

機械が研究するとはどういうことだろうか

人間の探究をめぐる議論に抛る概念的試論

Shiro Takagi*

要旨

機械が研究するとはどういうことだろうか。そして、機械が研究するとき、研究の何が変わりうるだろうか。機械が生み出した成果の報告は相次いでおり、それとともに、正確には何が解かれたのか、成果は誰にとって新しいのか、作業のどの部分を機械が行ったのか、をめぐる争いも生じている。本書は、これらの問いをめぐる概念的な試論である。出発点は、薄い作業上の特徴づけである。すなわち、社会 S に相対的に、研究とは S にとって新しい認識的達成をもたらそうとする試みである。これにより、研究を行うものが機械であってもよく、 S が人間の社会でも、混合の社会でも、人工の社会でもよいことになる。そのうえで本書は、人間の研究者を念頭に組み立てられてきた知識・理解・探究・研究の社会的組織についての議論に抛り、研究を行うものが機械であるときそれぞれが何を示唆するかを、説ごとにたどる。問うのは次のことである。機械の出力はいつ、誰にとって知識や理解になるのか。答えがまだないなかで、問いを研究可能にすることから、失敗の所在を突き止め、基準を改めることまで、探究はどう進むのか。人工の構成が研究を行ったこと、自律していることは、いつ認められ、成果は誰の知識になるのか。探究の配分、批判と認定、決定の権限は、人工の構成の内側に移りうるのか。そこから二つのことが見えてくる。一つは、機械が研究しているという主張が、それぞれ別々に成り立ったり成り立たなかったりする複数の主張に分かれることである。もう一つは、人間の研究では、研究者が人であり多くの人に分かれていたために決めずに決まっていた多くのことが、決めうるものになることである。それは作り手によって定められることも、システム自身によって形づくられることも、気づかれないまま失われることもある。誰が信念をもつか、批判者がどれほど独立しているか、誤りに誰が責任をもって応えるか、基準の改訂を何に照らして測るか、がその例である。本書はこれらの問いに決着をつけない。研究が何に拠ってきたかを名指し、これらの問いを問えるようにすることを試みる。

* 本書の文章は、主に、著者の指示のもとで AI エージェントの Claude (Anthropic) と Codex (OpenAI) が書いた。

本書は英語版の日本語全訳である。文献参照・URL・参考文献一覧は原文のまま残した。

© 2026 Shiro Takagi. クリエイティブ・コモンズ 表示 4.0 国際ライセンス (CC BY 4.0, <https://creativecommons.org/licenses/by/4.0/deed.ja>) のもとで提供する。他の著作からの引用部分はこのライセンスの対象外であり、それぞれの権利者に帰属する。

- 序論
 - 1. 操作の一覧は概念上の選択を覆い隠す
 - 2. 既存のラベルはひとつの目標を特定しない
 - 3. なぜ意見の不一致は特徴づけの必要をなくさないのか
 - 4. 作業的特徴づけは、恣意的でなくとも目的に沿ったものでありうる
 - 5. なぜ本書は科学だけでなく研究から出発するのか
 - 6. 議論の道筋
- 第I部 何を研究と捉えるのか
- 第1章 研究の作業的特徴づけ
 - 特徴づけに必要なもの、必要でないもの
 - 1. 認識的達成への薄い方向づけ
 - この文がまとめるものと、別の語彙が要るもの
 - この特徴づけはひとつの大域的な目的を確立するものではない
 - 2. 各語が担う働き
 - 3. 試みは研究活動をその成功から切り離す
 - 4. 達成をもたらすことは、産物をつくることとも、一度の閾値を越えることとも異なる
 - 5. 認識的達成では何を達成したかを明らかにする必要がある
 - 6. S にとって新しい (new to S) は、省略されていた参照先を明示する
 - 7. 作業的特徴づけから比較的直接に導かれること：未解決の点へ向かうこと
 - 8. 帰結しないこと
 - 方向づけとその区別を多様な研究実践と比べる
 - 9. 特徴づけが残す負債
- 第II部 何が認識的達成とみなされるのか
- 第2章 結果はいつ知識となるのか
 - 認識的变化、達成、知識
 - 1. 新しい結果から認識的達成へ
 - 2. 過程における状態——どのような種類の認識的態度が問題なのか
 - 2.1 信念の担い手は結果の産出者である必要はない
 - 3. 真理と可謬性
 - 4. 真なる成功を偶然的ではなく認識的なものにするのは何か
 - 4.1 認識的成功を位置づけるいくつかの仕方
 - 4.2 条件的な帰結
 - 4.3 基礎づけと作動しているコミットメント
 - 4.4 システムがもつ証拠と検索の蓄え
 - 4.5 生成すること、知ること、貢献すること、研究すること——偶然的生成器の事例
 - 5. 検証は何に寄与するのか
 - 5.1 検証は多様だが、それで機能を失うわけではない
 - 5.2 確証：結果は何を、どれだけ支持するのか
 - 5.3 限定的な例示としての統計的方法
 - 6. 知識について社会的なのは何か——新しさの参照先、産出、保有、保持
 - 7. 読まれていない、人工システムによる発見
 - 8. 知識の手前と周辺における認識的变化
 - 8.1 研究は知識に到達せずとも適切に遂行されうる
 - 8.2 知識の手前の変化：無知の明示化と、それが進歩かどうか
 - 8.3 知識の周辺の達成と進歩の説
 - 9. 認識論の諸説は人工システムによる研究に何を要求するか
 - 9.1 研究の実行者としてのシステムに掛かるもの
 - 9.2 システム自身が知っていると考えられる場合
 - 9.3 他の者がシステムを通じて知っていると考えられる場合
 - 9.4 結果が S の認識的状況に入ったと考えられる場合
 - 9.5 知識に届かない変化
- 第3章 理解はいつ、誰において達成されるのか
 - 知識、説明、そして人工システムによる不透明な探究
 - 1. 研究が求めうる達成としての知識と理解
 - 2. 予測と説明は関連しているが同一ではない達成である

- 3. 何が理解とみなされるのか
 - 3.1 説明に基づく説
 - 3.2 対象的理解説と依存関係説
 - 3.3 能力説と理解可能性説
 - 3.4 事実性、理想化、認識的に有用な歪曲
 - 4. 説明的理解なき知識
 - 5. 不透明性には所在がある
 - 5.1 候補生成と最終的な根拠づけ
 - 5.2 不透明性の多次元的な見取り図
 - 5.3 透明性は必要か、十分か
 - 5.4 事後の再構成は自動的に合理化であるわけではない
 - 6. 誰が理解するのか
 - 7. 理解は何のゆえに重んじられ、時間を通じて何を可能にしようか
 - 7.1 それ自体のために、あるいは目的として重んじられる理解
 - 7.2 事態がどう違っていたかを把握すること：転移、介入、診断
 - 7.3 領域を結びつけること：保持、再利用、さらなる問い
 - 7.4 理解を他者に利用可能にすること：コミュニケーション、批判、記憶
 - 8. 局所的な達成が継続的な探究能力の問題を決着させない理由
 - 9. 理解の諸説は人工システムによる研究に何を要求するか
- 第III部 未完の探究はどのように進むのか
- 第4章 機械が問うとはどういうことか
 - 1. 「機械が問う」という主張を分ける
 - 2. 何が問われ、何が答えになるのか
 - 3. 問いを理解することと、問いを扱えるものにする
 - 4. 問いは探究のなかでどう形づくられ、立て直されるのか
 - 5. 問いの前提と、問いの閉じ方
 - 6. 誰が問うのか
 - 7. A の未知と S の未知
 - 8. 「機械が問いを生んだ」という主張は何を示すのか
 - 9. 問いから研究を捉える見方と、本書の捉え方
- 第5章 可能性はいかにして利用可能になるか
 - 1. 手続きの記述と、それが示さない仕事
 - 1.1 第III部が用いる見取り図
 - 1.2 見取り図の拠りどころ
 - 2. 単一の方法はなく、神秘への回帰もない
 - 2.1 出力の証拠と過程の証拠
 - 3. 候補と表象を利用可能にすること
 - 4. 候補の形成、追求、評定はどう関係するか
 - 4.1 局面ではなく、区別できる機能
 - 4.2 何を追求するか判断と、何が真かの判断
 - 5. 異質な操作、比較可能な機能
 - 6. 物質的・身体的・時間的・分散的な探究
 - 6.1 実践、参加者、役割
 - 6.2 分散と実験室の研究
 - 6.3 記録
 - 6.4 身体性
 - 6.5 探究者が人間でないとき
 - 7. 探究の可能性が変わること
- 第6章 結果は何にかかわるのか——テスト、失敗、修正
 - 1. 結果から、それがかわるものへ
 - 2. 結果が期待に反するとき
 - 2.1 結果はひとつの配置から来る
 - 2.2 すべてが等しく疑わしいわけではない
 - 2.3 最初に疑われるもの
 - 2.4 何が局所化を導き、どこでそれが難しくなり、誰と誰のあいだで争われるのか

- 2.5 焦点となる主張に届いていない失敗
 - 3. チェックに合格したとき
 - 3.1 何にかかわり、どれほど強く決着させるか
 - 3.2 見かけの成功、成功が支持するもの、一致がもたらす確かさ
 - 3.3 チェックを定める者と、合格を読む者
 - 4. 一連の結果にわたる診断
 - 4.1 コミットメントを保護することはいつまで合理的であり続けるか
 - 4.2 研究を行うのが機械であるときの保護と修正
 - 4.3 診断としての試行錯誤
 - 5. 基準そのものが問われるとき
 - 5.1 結果はどのようにして基準にかかわりうるのか
 - 5.2 制約のもとでの、内部からの修正
 - 5.3 基準を改善することと、置き換えること
 - 6. この過程の説明が先へ引き渡すもの
- 第IV部 何が人工の構成を研究者たらしめうるのか、そしてそれはどの社会のためか
- 第7章 人工システムはいつ研究者とみなされるのか
 - 貢献、遂行、能力、地位の帰属と、自律性の比較
 - 1. 帰属の問い
 - 2. 遂行：探究の組織化はどこにあるのか
 - 3. 能力、地位、功績
 - 3.1 能力
 - 3.2 研究者としての地位
 - 3.3 功績、称賛、著者性
 - 4. 指標づけられた比較としての自律性
 - 4.1 自律性の意味
 - 4.2 時間にわたる介入と、三つの依存関係
 - 4.3 探究が継承する要素に対する自律性
 - 4.4 探究を固定せずに委ねることが要求するもの
 - 4.5 自律性が確立しないもの
 - 5. 帰属から構築と評価へ
- 第8章 誰の知識か？
 - 人間の、混合の、そして人工の認識的社会
 - 1. 行為者、知識の担い手、参照社会
 - 1.1 人間の社会のための研究
 - 1.2 混合社会のための研究
 - 1.3 人工社会による、人工社会のための研究
 - 2. 世界はいかにして探究者にとって利用可能になるのか
 - 2.1 検査は何を前提するのか
 - 2.2 接地はひとつの関係ではない
 - 2.3 世界への応答可能性
 - 3. 非人間的な研究と認識的アラインメント
 - 3.1 異質な科学を分解する
 - 3.2 ミスマッチはどこから生じるか
 - 3.3 AIアラインメント問題との類比
 - 3.4 認識横断的な協調
 - 4. 二つの地平と、第IV部が残すもの
- 第V部 研究を取り巻くどの関係が人工の構成の内側へ移りうるのか
- 第9章 Sのための研究を取り巻くのはどのような関係か
 - 役割、価値、評価、統制
 - 1. 研究の帰属が研究システムと出会うところ
 - 2. 三つの関係と、それを保持する当事者
 - 2.1 三つの関係の導出
 - 2.2 研究を取り巻く役割
 - 2.3 実現、分割、再結合
 - 2.4 関係の連動と、Sにとっての地位

- 3. S にとって価値がある
 - 4. 評価することに含まれるもの
 - 5. 統制と、それを制限する権能
 - 6. 第V部の進め方
- 第10章 探究はどのように配分されるのか
 - 分業、多様性、誰の未知か
 - 1. 配分はなぜ生じ、何がそれを難しくするのか
 - 2. 哲学者君主の二つの力
 - 3. 動機なしに舵を取る：信号
 - 4. どの路線を、どれだけ長く生かし続けるのか
 - 5. 人間の配分が決めずに供給していたもの
 - 6. 誰の未知か
 - 7. 第10章が見いだしたこと
- 第11章 S のための批判と認定
 - 無効化、独立性、信頼、訂正
 - 1. 本章はどの評価を検討するのか、そしてなぜか
 - 2. 誤り検出の見方における独立性
 - 3. 批判者を作れるとき、違いはどこから来るのか
 - 4. どの誤りが重要か：認定における価値
 - 5. 認定における依拠、証言、信頼
 - 5.1 人工の構成による認定とは何か
 - 5.2 点検していない結果への依拠を何が根拠づけるか
 - 5.3 その依拠は信頼か、信頼は必要か
 - 5.4 人工の構成の内部の一致は依拠する者に何を伝えるか
 - 5.5 何が残るか
 - 5.6 認定と理解
 - 6. S が素人になるとき
 - 7. 批判の何が変わるか：費用、範囲、判断、やり方
 - 8. 時間のなかの批判：代理指標と基準、新しいものの評価、訂正
 - 9. S のための批判は人工の構成の内側で組織できるか
 - 9.1 説が答える問いと、説ごとの判定
 - 9.2 説を通して見えてくること
 - 9.3 そこから示唆されること
 - 9.4 第12章へ渡る問いと、理解についての限定
- 第12章 誰が決め、誰が改めてよいのか
 - 当事者資格、修正可能性、外の当事者との関係
 - 1. 権威は人工の構成の内側に置けるか
 - 1.1 問いと、前二章から渡されたもの
 - 1.2 目的の順序づけは人工の構成の内側で定まるか
 - 1.3 誰に対して決定を正当化するのか
 - 1.4 有能さは当事者資格を与えるか
 - 2. 決定と決定権
 - 2.1 六種類の決定
 - 2.2 決定権
 - 3. 権威の根拠と、認識的なものと規範的なものが出会うところ
 - 3.1 結果に基づく根拠：結果の質と、確かめられた有能さ
 - 3.2 研究共同体の自己統治
 - 3.3 民主的な授権
 - 3.4 同意と支配からの自由
 - 3.5 認識的なものと規範的なものは、どこで分けておけないのか
 - 3.6 同じ成績のもとで根拠を比べる
 - 4. 名目以上の当事者資格
 - 5. A と S がともに人工であるとき
 - 6. 権威を修正可能に保つ
 - 6.1 是正可能性：エージェントから制度へ

- 6.2 異議申立て可能性
- 6.3 誰が答え、誰が負うのか
- 6.4 関係としてのインフラストラクチャー
- 7. 修正する能力を取り除く決定
- 8. 第12章が見いだしたこと
 - 8.1 根拠ごとに見た、誰が権威をもちうるか
 - 8.2 構築者の第三の主張
 - 8.3 A と S の組み合わせ、そして時間
 - 8.4 権威が内側に入ったとき研究はどういう姿になるか
 - 8.5 第11章から渡された問い
 - 8.6 説を通して見えてくること
 - 8.7 そこから示唆されること
- 9. 第V部が見いだしたこと
- 第13章 結論：機械が研究するとはどういうことか
 - 研究、人工の行為者、探究の組織
 - 1. 四つの問い
 - 2. 作業的特徴づけによる研究の捉え方
 - 3. 知識と理解は、いつ、誰において実現するのか
 - 4. 研究という、答えがまだない中で進む営みでは、何が行われているのか
 - 5. 研究しているのは誰で、その達成は誰のものになるのか
 - 6. 研究を取り巻く関係を組み替える
 - 7. 決めずに決まっていたものを名指す
- 参考文献

序論

2026年10月27日の時点で、機械が研究にもたらしたものについての報告が相次いでいる。そのなかには数学からのものも多い。2026年5月には、OpenAIの未公開のモデルが、1946年にエルデシュ（Erdős）が立てた単位距離予想の反例を出したと報じられ、9人の数学者が、この反例の構成を検証した要約を公表した（Alon et al. 2026）。7月には、ヤコビアン予想の3変数の場合について反例が発表された。テレンス・タオ（Terence Tao）の解説によれば、これはAnthropicのモデルの助けを借りて見いだされた（Tao 2026）。8月には、Anthropicが、未公開のモデルによって、リーマンのゼータ関数の非自明な零点のうち臨界線上にあるものの割合について、証明された下界が41.6パーセントから67.2パーセントに引き上げられたと発表した（Anthropic 2026）。同じ月にOpenAIは、社内のモデルが得た10件の結果を、それぞれLeanによる証明を添えて公開した（OpenAI 2026a; OpenAI 2026b）。9月には、OpenAIが、滑らかな外力を加えた3次元のナビエ-ストークス方程式について、有限時間で速度が非有界になる解の構成を公表し、これがミレニアム懸賞問題の公式の定式化にある選択肢の一つを確立するものとした（OpenAI 2026c）。クレイ数学研究所は、誰が解いたかには触れずに、問題は「どうやら解決された（apparently been settled）」と述べた（Clay Mathematics Institute 2026）。10月にはOpenAIが、社内のモデルが書いた700本を超える原稿を公開のリポジトリに置き、功績や先行性の判断を数学の共同体に委ねた（OpenAI 2026d）。

形式的な検査や、専門家が作った試験でも、同じような結果が報告されている。Google DeepMindのエージェント・システムは、共同体のリポジトリ Formal Conjectures で形式化されていたエルデシュ問題353題のうち9題を解いた。証明はいずれもLeanで検査され、形式化された命題がもとの問題と合っているかはチームの数学者が確かめた（Tsoukalas et al. 2026）。現役の物理学者が用意した、未発表で研究水準の理論物理の問題50題では、GPT-6 Astraがウェブを使える条件での1回の試行で49題に、Claude Opus 5.5が46題に正答した（Chandra et al. 2026）。形式化も研究の規模に達している。人間の監督のもとで、コーディング・エージェントが、MIP* = REの証明が依拠する定理について、12万6千行を超えるLeanの完全な証明を書き、その過程で公刊された命題の付帯条件を修正した（Lu, Deng et al. 2026）。

実験科学では、報告されるのはむしろ、設計や解析、実験室での最初の試験である。言語モデルのエージェントからなるチームが、新型コロナウイルスの最近の変異株に対するナノボディを設計し、人間が行った実験で、そのうち二つが最近の変異株への結合を改善した（Swanson et al. 2024）。システムには系譜もある。Googleのco-scientistは、2025年には研究者が試すための仮説を生成するものだったが、のちに実験を実行し論文を書くところまで拡張された。ある事例では、拡張されたシステムが2次元材料を成長させるための前駆体を提案し（著者たちはこれを危険性のないものとしている）、人間の専門家がその提案を手順へと仕上げた。得られた生成物は、目標の材料と矛盾しないものだった（Schmidgall, Zhu et al. 2026）。素粒子物理学では、ALEPH実験の保存データに取り組んだエージェントの枠組みが一つの測定結果を出し、著者たちはそれを、発見の主張ではなく、共同体の吟味にかける候補として示している（Moreno et al. 2026）。

機械は、機械自身を生み出す方法にも向けられている。2025年、Google DeepMindは、言語モデルを使ってプログラムを進化させるAlphaEvolveが、Googleの世界全体の計算資源の平均0.7パーセントを回収するスケジューリングの経験則と、Geminiモデル自体の訓練に使われるカーネルを速める方法を見つけたと報告した（Google DeepMind 2025）。2026年には、システムがエージェントのコードを書き換え、書き換えた版を評価によって選ぶループが、一つの研究の流れになった（Srikanth et al. 2026）。2026年10月半ばまでに再帰的な自己改善を掲げたarXivの記録173件のうち153件がこの年に投稿されており、ある調査によれば、研究の大半は、改善の過程ではなく、改善された後のシステムを評価している（Chen et al. 2026）。最前線の研究所

は、自社の研究のどれだけをモデルが担っているかを報告しはじめてもいる。複数のAI企業と大学の研究者が書いた政策論文はそうした数字を集めており、そのなかには、人間が大きな監督をするだけで自律的に完了した研究開発の作業の割合が、2026年3月から8月のあいだに1パーセントから26パーセントに増えたというAnthropicの報告もある（Chan et al. 2026）。問いから論文までを通して実行するシステムも、大きな規模で動いている。ある運用では417時間で166本の論文が書かれ、そのすべてが査読のために公開されたと報告されている（Tang et al. 2026）。機械は研究の認定にも入りこんでいる。AAAI-26という会議は、査読まで進んだ22,977本の投稿のそれぞれに、人間による査読と並べて、AIによることを明示した査読を一つずつ添えた（Biswas et al. 2026）。

これらの報告が何を意味するのかは、それ自体が争われている。しかも争いの多くは、こうしたシステムを作り、使う人々のあいだで起きている。一つは、何が解かれたのかという問いである。Leanによる検査が保証するのは形式化された命題の証明であり、形式化された命題が数学者の意図した問題であるかどうかは、別に判断されなければならない。Formal Conjecturesには、見つかって修正された誤った形式化が291件記録されている（Firsching et al. 2026）。ナヴィエ-ストークス方程式についての構成についても、そのLeanによる形式化の一部が文章で書かれた証明に対応していないという議論が、ほかの研究者から出されている（Bastounis, Circelli, and Hansen 2026）。もう一つは、結果が誰にとって新しいのかという問いである。2014年に未解決とされた問題の反例が、じつは1962年に出た本のなかで別の目的のために構成されていたこともわかった（Ju, Gao et al. 2026）。エルデシュの未解決問題については、新しい自律的な解の数がいったんは9件とされたが、文献を調べるにつれて4件に減った（Feng et al. 2026）。さらに、仕事のどの部分を機械が担ったのかという問いがある。AIの助けを借りた結果を、段階ごとに分けて分類しはじめた著者たちもいる。誰が問題を定式化し、どう解かれ、解がどう検証され、誰がそれを書き上げたのかを分けるのは、「AIが発見した数学」のような一つの言い回しでは、段階ごとの違いが覆い隠されてしまうからである（Tian and Narayanan 2026）。何が解かれたのか、誰にとって新しいのか、機械がどの部分を担ったのかという三つの問いは、以下の章で、より一般的な形で現れる。検査された命題が意図された問題であるかどうかは、形式的な検査そのものは下さない判断である（第3章第5.1節）。結果が誰にとって新しいのかは、新しさを測る基準となる社会をどう定めるかという問いになり（第1章第6節）、仕事のどの部分を機械が担ったのかは、機械に何を帰属させるかという問いになる（第7章）。

1. 操作の一覧は概念上の選択を覆い隠す

機械による研究の報告が相次ぎ、その意味が争われているなかで、本書は、機械が研究するとはどういうことか、そして機械が研究するとどうなりうるかを考える。この問いは、見かけほど単純ではない。研究者がすることなら、いくらかでも挙げることができる。研究者は文献を探索し、仮説を定式化し、実験を設計し、結果を証明し、史料を解釈し、人工物を構築し、データを分析し、論文を書き、互いの仕事を批判する。しかし、受け継がれた操作の一覧は、それらの操作を研究の一部たらしめているものについての説明ではない。目に見える一連の手順を再現しても、それだけでは、その根底にある営みが再現されたことを示しはしない。

操作とその機能との区別は単純だが、重大な帰結をもつ。文献探索は、未解決の問いの所在を突き止めることも、方法を習得することも、比較クラスを特定することも、予期せぬ結果を診断することも、のちの新規性の主張を検証することもできる。操作はそれと分かるほど似ていながら、探究のなかで占める位置は変わる。逆に、提案された答えが信頼に値するかどうかを評価するという機能は、実験、証明、シミュレーション、史料に基づく論証、フィールド観察、批判的解釈、人工物の構築と試験、あるいは研究者と装置にわたって分散した仕組みによって実現される。それと分かる一つの機能が、異種的な実現をもちうるのである。

それゆえ、おなじみの研究サイクル図を、そのまま研究の存在論、すなわち研究が何から成るかにについての説明へと移し替えることはできない。文献レビュー、アイデア生成、実験、分析、執筆、査読、出版は、それぞれに選択を含んでいる。分析は、説明されるべき現象を構成することもあれば、最初の実験が誤って構想されていたことを明らかにすることもある。執筆は、論証を単に報告するのではなく組み替えることがある。査読は、裏づけを検証し、主張を変え、信用を配分し、共同体がのちに何を追求に値するとみなすかを変えることがある。この図は作業を組み立てるうえでは有用であり続けうるが、それによってこの活動が何であるかが決着するわけではない。

一見すると認知的な段階についても、同じ問いが生じる。**アイデア**（idea）とは何であり、何がそれを実り豊かなものにするのか。好奇心は必要なのか、それとも探索を方向づけるための人間的メカニズムのひとつなのか。創造性とは再現の対象となる能力なのか、事後的に判断される出力の名前なのか、それともさらなる分析を要する一群のプロセスなのか。これらの名詞はいずれも、人工研究者に必要なだと考えられうるものを名指している。それぞれが何を名指し、どうすればエンジニアリングの目標になりうるのかは、それ自体が検討を要する問いである。第5章は、アイデアやほかの候補がどのように利用可能になるかを、**創造的**というラベルが何を説明しないまま残すか、創造性が何に帰属されるかも含めて扱う（第2節）。**好奇心**という語が問いの形成において何を名指しうるかは、第4章第7節が扱う。第7章第3.3節は、行為者への称賛の語としての創造性と発見に立ち返る。

本書を貫く問いは、研究には何が含まれ、その活動、参加者、あるいはそれを支える関係が人工のものになると何が変わるのか、というものである。構築はこれらの問いを具体的ににする。プロジェクトは何を実現するかを選ばなければならないが、問い、証拠、失敗、ほかの探究者との出会いは、その選択の不十分さを明らかにしうる。各章は二つの方向を行き来する。研究を明晰化することは構築の目標を精確にし、人工システムにおける実現を考えることは人間の研究から受け継いだ前提を試す。知識、理解、行為者性、権威の諸説は、研究に何が含まれるかについても、機械が研究すると何が変わるかについても、意見を異にする。本書は、これらの不一致をたどりながら、異なるプロジェクトが、人間の探究を拡張し、人工の研究行為者を成立させ、あるいは研究が営まれる社会を組み替えうる仕方を考えてみたい。そのようなプロジェクトについての主張を評価することは、この分析の一つの用途である。第6節で議論の道筋を示し、結論では検討を通じて何が見えてきたかに立ち返る。

本書を通じて三つの区別が重要になる。一つは、能力が存在することと、それがああるエピソードで発揮されることや、評価者がその帰属の証拠をもつこととの区別である（第1章第3節、第7章第3.1節）。二つめは、活動を研究にするものと、その活動の具体的な問いにうまく答えるものとの区別である（第1章第3節、第3章第7.1節）。三つめは、結果への依拠を根拠づけるものと、他者に代わって決める権限を与えるものとの区別である（第12章第1.4節）。以下の章は、対立する説と事例を通じてこれらの区別を展開する。どの区別も、すべての研究者があらゆる機能を単独で果たすべきだという要求ではない。

2. 既存のラベルはひとつの目標を特定しない

人工システムによる研究に関連する技術的な仕事は、自動発見、文献に基づく発見（literature-based discovery）、機械学習、実験室の自動化、定理証明、計算論的創造性、そして領域固有のシステムにまたがって分散している。そうした仕事の見取り図を描くのは経験的な課題である。それは、どの能力が実装されてきたのか、どのような制約のもとで、どれほどの人間とインフラストラクチャーの寄与を伴っていたのかを問う。第二の課題は概念的なものである。それは、結果として生じた仕組みが研究を遂行している、あるいは人工研究者を含んでいると合理的に言えるようになるまでに、それらの能力が個別に、あるいは合わせて何を成し遂げなければならないかを問う。

著名なシステムやプログラムは、すでに異なる目標を操作化している。初期のロボット・サイエンティスト（Robot Scientist）は、人間が構築した詳細な論理モデルの内部で、仮説形成、実験選択、物理的実行、解釈を反復した（King et al. 2004）。The AI Scientistは、アイデア生成、文献探索、計算実験、分析、論文産出、査読を通じて機械学習研究を組織するが、その一方で重要な人間による仕様策定と絞り込みを保持している（Lu et al. 2026）。分野の進む道筋を示そうとする人たちは、さらに別のところに線を引いている。ロボット・サイエンティストを作ったロス・キング（Ross King）は、いま、自律性を道具にあたる水準0から水準5までに等級づけている。水準5では、人間は科学の方向を定めるのではなく外から統治する。そのうえでキングは、システムの水準を、そのシステムがどう説明されているかからではなく、実際に観察された人間の介入から割り当てるよう求めている（King 2026）。First Proofという取り組みの主催者たちは、研究水準の問題を自分たちで動かすシステムに与えるが、試すのは証明だけである。予想を立てることや定義・枠組みを選ぶことは、研究にとって少なくとも証明と同じくらい重要だとしながらも、まだ測れないとして脇に置いている（Abouzaid et al. 2026）。ARIAのある公募は、問題を人間の応募者に委ねており、応募者は自らの既存システムが取り組む研究問題を特定しなければならない。そのうえでこの公募は、探索、仮説と実験の構築、解釈、報告のうちどれだけが人間の介入なしに生じるかを評価している（ARIA 2025, 4-5, 7）。これに対して北京市の2026-2028年の実施方案は、仮説の提出から革新的な発見までの全過程を、実験室の自律的な運用の範囲に含め、自律性の基準と等級別の評価はこれから定めるとしている（Beijing Municipal Committee 2026）。

これらの操作化は、互いに等価ではない。それぞれが、異なる活動、達成、行為者の周りに境界を引いている。いま挙げたシステムやプログラムは部分的で、強く条件づけられた事例であり、ここでは、同じラベルのもとに異なる構築上の主張がありうることを示すために用いる。

哲学的分析は、この食い違いをさらに広げる。AIによる発見についての諸説明は、独立した発見に何が必要かについて意見を異にしている。ベルガマス・キ・ガナピーニ（Bergamaschi Ganapini）は、発見は称賛に値する認識的達成であるため、関連分野での有能さだけでなく、その有能さの及ぶ範囲と限界についての反省的な洞察も必要だと論じ、現在のAIシステムにはまだそれが見られないとする（Bergamaschi Ganapini 2025; 第7章第3.3節）。これに対してデュエデ（Duede）とフリードマン（Friedman）は、この種の要件を退ける。二人はこうした立場を、発見を発見者の認知的能力に依存させるものとして、認知主義的（cognitivist）と呼ぶ。デュエデとフリードマンの説では、発見とは認識的ギャップ（epistemic gap）をうまく乗り越えることである。認識的ギャップとは、既存の知識、理論、予期が及ばないところをいう。発見は、それらでは収まりのつかない現象を明るみに出すことによってそうしたギャップを露わにするか、すでに露わになり明確に定まったギャップを、新たな知識、理論、方法によって埋めるかのいずれかである。二人は、有能さ、反省的な洞察、そして特定の課題に限られない領域全般にわたる知識は、発見にしばしば伴うものの、発見に不可欠ではないと論じる。たとえばジョスリン・ベル・バーネル（Jocelyn Bell Burnell）は、自分が気づいた信号の意義をまだ説明できない博士課程の学生として、最初のパルサーを発見した。そのうえでデュエデとフリードマンは、そうしたギャップを埋める人工システムには発見を帰属させようと結論する。二人が挙げる例は、大規模言語モデルを中心に組まれた探索の枠組みが、オンラインのピン詰め問題（到着する品物を順に、容量の決まった箱へ、できるだけ少ない箱数で詰める問題）について、新しい経験則的な解法（ヒューリスティック）を見いだした事例である。また二人は、システムに発見を帰属させることは、それを科学者として認めることではないとつけ加える。さらに、発見に科学共同体でのしかるべき立場も要るのであれば、現在のAIシステムはたいていそれを欠くと認める。ただし、その障壁は社会的・歴史的な事情に左右されるものだとみる（Duede and Friedman 2025, §§1-5; 第7章第3.3節）。集団を中心に据える研究は、発見が、ひとりの個別的なエージェントにではなく、人々、機械、制度、物質的存在者からなる異種的な組織に帰属されようと論じる（Clark and Khosrowi 2022）。ほかの研究は、次のような問いを立てている。基礎研究における人工システムの行為者性（agency）はどのようにモデル化されるべきか（van Lier 2023）。AI駆動の研究は社会的に分散した認知活動であり続けるのか、そして異質な科学（alien science）を生み出しうるのか（Kureha and Kukita 2019; Kureha 2023）。中核的な認識的目的を、人工の科学者のための非人間中心的な目標へと翻訳できるのか（Krenn and Champion 2026）。

これらの説明が意見を異にするのは、ひとつには、発見、貢献、参加、行為者性、認識的主体性、研究者としての地位が異なる可能性を指しているからである。人間による発見を可能にするプロジェクトと、人工の知る者を成立させるプロジェクトでは、出力が似ていても異なる達成が求められうる。したがって、目標を明示することは、構築したことの証拠とみなされるものとともに、何を構築するのかをも変える。

3. なぜ意見の不一致は特徴づけの必要をなくさないのか

研究実践は、学問分野、装置、専門職、制度、共同体を通じて形成されてきた。数学、文学研究、分子生物学、人類学、計算機科学、工学、芸術的研究は、対象、操作、証拠的關係、産物において異なる。それらの基準もまた変化してきた。受け入れられているあらゆる用法を論争なしに再現するような短い定式は存在しないだろう。

この観察は、正反対の二つの応じ方を誘う。ひとつは、概念分析を脇に置き、研究者がいましていることを自動化することである。本書の目的からすると、この応じ方では、研究の現在の組織のあり方が検討されないまま、所与として受け入れられることになる。もうひとつは、研究の哲学が決着するのを待ってから、有用なシステムをつくることである。検索、証明検査、計算論的再現、実験のスケジューリング、装置制御のための限定されたツールは、そのような決着を必要としない。

概念上の問題は、研究のどの側面を実現し、どの側面を周囲に委ねるかをプロジェクトが選ぶときに入り込む。それは、限定された性能からより広い帰属へ進む段階でとりわけ明瞭になる。分析の再現、伏せられた知見の回収、実験上の目的関数の最適化、慣例に沿った論文の産出は、実際の能力を示しうる。これらの能力が構想された研究に相当するかどうかは、課題が何を代表し、何が固定され、あるいは他者から供給されていたかに依存する。評価はその選択を見えるようにするが、その選択はすでに構築を方向づけていたのである。

これは語集の問題であると同時に、構成概念妥当性 (construct validity) の問題でもある。研究を遂行すること (conducting research) のような広い能力は、ひとつのスコアのなかに直接観察されるものではない。それは、選ばれた課題、環境、人工物、指標、判断を通じて操作化される。AIベンチマークに関する研究は、有限で文脈的な課題が有用であり続ける一方で、一般的な能力の主張が、課題とその目標とのあいだで擁護された関係を超えてしまうことがあることを示してきた (Raji et al. 2021; Bean et al. 2025)。ここに当てはめるなら、その教訓は研究のベンチマークが不可能だということではない。評価が支えるのは、目標との関係が擁護されているようなより広い推論だけだ、ということである。

現在の研究エージェント向けベンチマークは、この点を具体的に示している。CORE-Benchは計算論的再現を検証し、それを新規の自律的研究の一例としてではなく、そこへ向かう一歩として提示している (Siegel et al. 2025)。ScienceAgentBenchは、データ駆動型の研究課題のための実行可能なプログラムを評価する (Z. Chen et al. 2025)。FIRE-Benchは、高水準の問いと実験の範囲から出発する制約付きの再発見を評価する (Wang et al. 2026)。MLR-Benchは、より長い機械学習のパイプラインを評価し、首尾一貫した論文が、妥当でない、あるいは捏造された実験結果と共存しうることを示している (H. Chen et al. 2025)。各ベンチマークは、それが構成する目標についての証拠を提供する。その目標が研究一般とどう関係するかは、別の問いであり、第2章第4.2節、第5節、第5.2節がこれを認識論の観点から扱う。

限定された能力だけを主張するプロジェクトは、その能力についての擁護可能な説明を必要とする。研究を自動化する、発見を行う、あるいは人工研究者をつくると主張するプロジェクトは、より広い負担を引き受けることになる。そのようなプロジェクトの主張は、どのような構想のもとでその帰属が認められるのか、その証拠が研究のどの側面に取り組んでおり、どの側面が評価の外に残っているのかが明示されてはじめて、評価することができる。これは哲学的な合意を必要としない。必要なのは、主張の範囲が批判に開かれていることである。

4. 作業的特徴づけは、恣意的でなくとも目的に沿ったものでありうる

研究の定義は異なる目的に仕える。OECDの『フラスカティ・マニュアル』(Frascati Manual) は、国際統計のために、新規性、創造性、不確実性、体系的な遂行、移転可能性ないし再現可能性によって研究開発を分類している (OECD 2015)。これらの基準は、行政上の目的にとっては権威あるものだが、だからといって研究活動の哲学であるわけではない。この例は、定義が、研究とは何かを述べることなしに、ある仕事にとって権威をもちうることを示しており、それゆえ、説明を、それが果たすことを意図された仕事に相対的に捉えることを示唆する。第1章で提案する特徴づけは、この意味で目的に沿ったものであり、機械が研究するとはどういうことか、機械が研究するとどうなりうるかを考えるための足場として差し出される。そのような特徴づけに対する三つの懸念を、以下で順に取り上げる。それが恣意的だという懸念、目的のために概念を修正すると主題がすり替わるという懸念、そして説明は制度に用いられるやいなや、それが排除するものを通じて帰結をもつという懸念である。

カルナップ (Carnap) の解明 (explication) の方法は、ひとつのモデルを提供する。不明確な概念は、述べられた目的のために、より正確で実り豊かな概念によって置き換えることができ、しかもその際、両者が当てはまるものの範囲 (外延) が同一であると装う必要はない (Leitgeb and Carus 2020, suppl. D)。本書が用いる形でのカルナップのモデルでは、述べられた目的のために行われる置き換えは、その目的に照らして評価されるので、恣意的ではなく制約を受ける。これが第一の懸念に対する本書の応じ方である。本書がこのモデルを用いる理由もここにある。このように導入される概念は述べられた目的に応えるものであり、古い概念と外延を共有する必要がないので、本質を待つのではなく、明示的で批判可能な特徴づけを足場として、機械が研究することについて考えを進めることができる。本書の場合、述べられた目的はそうした考察なので、このモデルだけに従えば、特徴づけが応えるべき基準は、実質的にはその考察にとっての有用性になる。人工システムをつくろうとするなら、それはエンジニアリング上の有用性になるだろう。

しかし、目的相対的な修正は、主題を明確にするどころか主題をすり替えてしまうことがある。これは、カルナップに対するストローソン (Strawson) の批判に由来する反論である。ここに当てはめれば、懸念は次のようになる。目的に合わせてつくられた特徴づけが研究と呼ぶことをあるシステムが行っているとしても、それが日常的な意味で研究をしているのかという問いには答えていないかもしれない。この反論がどこまで及ぶかはそれ自体争われており、現代の諸説明は困難の所在を異なる仕方とらえる。ここで取り上げる三つの説明は、いずれも反論を退けるものではなく、それぞれ反論の及ぶ範囲を絞るか、その効力の所在を移すものと読める。マーク・ピンダー (Mark Pinder) は、この反論が成り立つのは和らげられた限定的な形においてだけだと論じる。すなわち、哲学的な問題を述べるのに必要な概念が、科学にとって実り豊かであるよう設計された概念によって、補われるのではなく置き換えられるときに反論は効力をもち、解明が主題をすり替えるかどうかは、それがどう行われ、何に使われ、問題がどう解釈されているかに依存する (Pinder 2020, §§3.2, 5.2, 6)。トリストラム・マクファーソン (Tristram McPherson) とデイヴィッド・ブランケット (David Plunkett) は、主題の連続性 (topic continuity) を文脈に依存するものとして扱うことを提案する。そこで問われるのは伝達・表象・推論の連続性、すなわち、改めた語を使っても話が通じ、以前の使い手の主張を報告できるか、語の表す内容が保たれるか、その語に結びついた推論の型が保たれるかである。その際、これらは提案を評価する人々の関心に応じて重みづけられ、連続性は、提案の全体的な評価ではなく、覆されうる一つの望ましい条件として扱われる (McPherson and Plunkett 2024, §§1–2)。マチュー・ケロー (Matthieu Queloz) は、概念の権威、すなわちその概念を使うよう求められる人々がそれを採用する理由は、彼らが反省のうえで是認するであろう関心にその概念が仕えるかどうかにかかると論じる。ケローの説では、主題の連続性が重要なのは、この権威の代理としてにすぎない。連続性が求められるのは、置き換えられる概念を使う理由を与えていた関心に、修正された概念も仕えることが望まれるからである。そのため、連続性は十分条件ではない。もとの問いに答えを与え続ける概念でも、よい答えを与えるとは限らないからである。また必要条件でもない。関心のなかには問いの形で表せないものがあり、そうした関心には、もとの概念の主題は保たずに機能を保つ概念が仕えうるからである。同じ理由から、精密さの向上が権威を与えるのは、論理学のように、精密さがもともと使い手の関心であるところに限られる (Queloz 2022, 1247, 1250, 1253,

1255–1256, 1262)。ここに当てはめれば、研究の特徴づけを精密にしたというだけでは、精密さそのものを求めているのではない使い手としての権威は生まれない。

したがって、述べられた目的にとっての有用性は、特徴づけが応えるべきもののひとつではあっても、唯一のものではない。ここまでの議論に照らせば、本書の作業的・特徴づけには、ほかに求められることがある。すなわち、それが異種的な研究実践と結びついたままであり、課題の一覧が覆い隠す事例を明確にし、重大な選択を明るみに出し、機械が研究するときに何を問えばよいかをはっきりさせること、そして、知識、理解、実践、行為者性、あるいは研究システムについてののちの説明が、特徴づけが多くを排除しすぎていること、あるいは誤ったものを含んでいることを示したときには、修正可能であり続けることである。ピンダーの区別をここに当てはめると、このように用いられる特徴づけは、ピンダーが完全な解明と呼ぶものよりも、部分的な解明、すなわち関連する理論的文脈で古い概念を置き換えることなくより正確な概念を導入するものに近い。そしてピンダーの再構成では、主題のすり替えという非難が効力をもつのは完全な解明に対してである (Pinder 2020, §§2, 3.2, 5.2)。つまり本書がカルナップのモデルから受け取るのは、置き換えが述べられた目的に照らして評価されるという点であり、日常の「研究」の概念をこの特徴づけで置き換えることではない。これが第二の懸念、すなわち目的のために概念を修正すると主題がすり替わるという懸念に対する本書の応じ方である。

第三の懸念は、説明が主題をすり替えるかどうかではなく、それがどう用いられるかにかかわり、第二の懸念に答えたあとにも残る。定義は制度的に無力なものではない。ギアリン (Gieryn) は、長く哲学者や社会学者にとって分析的な問題であった、科学をほかの知的活動から区別する境界画定を、科学者にとっての実践的な問題として検討する。その説では、科学とさまざまな非科学とのあいだに境界を築くことは、科学者が職業上の目標を追ううえで役に立つ。その目標とは、知的権威と職業上の機会を得ること、科学者が疑似科学者と呼ぶ者たちにはそうした資源を認めないこと、そして科学研究の自律を政治的な干渉から守ることである (Gieryn 1983, abstract)。チー＝トン・ホアン (Chih-Tung Huang) が紹介するギアリンの事例では、ジョン・ティンダル (John Tyndall) は、科学を有用で、経験的で、懐疑的で、客観的なものとして描くことで宗教とのあいだに境界を引き、自然について語り、教育課程を定め、公衆の支持を得る権威を科学者のために求めた。他方、産業革命を推し進めたのは科学者ではなく職人だと多くの英国人が考えていた時代に、機械工や技術者に対しては、科学を、より高貴な用途によって正当化される理論的で純粋な営みとして描いた。エディンバラでは、解剖学者たちが、骨相学を、理論が曖昧で、専門家の判断ではなく世間の意見に頼る疑似科学として描き、科学の共同体から押し出して、権威と資源に近づいた (Huang 2012, 2–4)。ギアリンは、こうした記述を緊張と利害の両方にたどる。科学に帰属させうる別々の特徴の組は、制度の内部の両義性を映しており、そのため科学は、経験的にも理論的にも、純粋にも応用的にも見せることができる。そして、どの記述が選ばれるかは、科学者の権威や資源への要求を正当化するかたちで境界を最もよく引くのはどれかによる (Gieryn 1983, abstract)。この分析を研究の特徴づけに当てはめると、研究の特徴づけもまた同じように用いられうることになる。本書に必要なのは、より限定された論点だけである。すなわち、ある説明がベンチマーク、研究資金の基準、出版規則、あるいは功績の基礎となるやいなや、その説明による排除は帰結をもつ。

たとえば、論文を単位とする説明があったとすれば、それは現在の出版形態を特権化するだろう。実験を中心に据える説明なら一部の領域を特権化するだろうし、個別エージェントを中心に据える説明なら分散した労働とインフラストラクチャーを隠してしまいうるだろう。ケローの説からは、こうした排除にかかわる帰結が一つ引き出せる。概念の権威は、その使い手が認めるであろう関心に相対的なもので、同じ概念が、ある集団にとっては権威をもち、別の集団にとってはもたないことがありうる (Queloz 2022, 1265–1266)。そうした見方に立てば、機械が研究することについて考える人々の関心に仕える特徴づけは、それだけでは、その特徴づけによって実践を分類される研究共同体にとっての権威をもたない。これは、本節の第一段落がOECDの基準に付したのと同じ種類の限定である。ある目的にとっての権威、あるいはある集団にとっての権威は、それだけでは別の目的や別の集団に及ばない。この限定は一方方向にはたらく。限られるのは、特徴づけがそれらの共同体に対してもつ権威であって、それらの共同体が特徴づけを批判する立場ではない。特徴づけは、それらの共同体の実践に応えるものであり続け、その実践に照らして改められうる (第1章第1節はこの点を述べ、第8節では特徴づけをそうした実践と比べる)。

科学におけるモデリングと比べると、本書の進め方の全体がつかみやすくなるかもしれない。ワイスバーグ (Weisberg) は、モデリングを、モデルを通じて現象を間接的に研究することとし、三つの段階で描く。モデルを作り、それを分析し、そのうえで、そうすることが適切な場合には、モデルと対象との関係の評価する。モデルと対象との関係は、部分的にはモデルの解釈 (construal) によって定まる。解釈は、モデルの部分を対象の部分に割り当て、意図された射程を述べ、モデルが満たすべき忠実性の基準を定める (Weisberg 2007b, 4, 14–16)。作業的・特徴づけは、この意味での研究のモデルとして読むことができる。それは第1章で作られ、そこでその帰結が分析される。そして後の章は、それを研究の実践に照らしながら、このモデルから機械の研究について何が引き出されるかを考える。それは、他のモデルを不要にするような完全な表象として示されているのではない。

本書が依拠するものの多くも、同じように研究の部分的なモデルとして読める。知識、理解、探究の過程、行為者性、研究の社会的な組織といった、研究の一つの側面のために作られたモデルである。これまで研究をしてきたのは人間なので、これらのモデルが扱ってきたのも、当然ながら人間の研究である。本書は、これらの部分的なモデルをつなぎ合わせながら、研究を行うのが機械であるときに何が変わり、何が変わらないのか、それはどのような意味でなのかを考えてみたい。「機械が研究するとはどういうことか」という問いで本書が目指しているのは、このことである。そのために、それぞれの議論のどこまでが人間の研究に固有で、どこからが機械の研究にも成り立ちうるのかを見分けようと試みる。また、人間の研究についてはある含意をもっていた議論が、機械の研究では別の含意をもつこともある。さらに、知識がそのために求められる者たち (第1章で *S* と呼ぶ社会) が人間でない場合には、また別の含意をもつこともある。たとえば、人間の研究は人間の身体に形づくられてきた。これを「探究者はその探究者自身の身体に形づくられる」という形にすれば、機械にも成り立ちうる。そのうえで、身体性の議論のどの部分が人間に固有なのかを問うことができる (第5章第6.5節)。第9章第6節は、同じ問いを別の形で立てる。研究を取り巻く関係 (探究の配分、批判と認定、権威など) がこれまで別々の当事者に分けて担われてきたことによって供給されていたかもしれない財が、何に依っていたかを問い、その答えを四つに仕分ける。研究を担ってきたのが人間であることについての前提、現在の科学のシステムがどう組織されているかについての前提、どんな探究者にも残る探究の条件、そして研究を担う単位の外にいる当事者との関係である。ワイスバーグの言葉でいえば、これは、どの結果が表象の枠組みの変更に対して頑健であるかを問うことに近い (Weisberg and Reisman 2008, 15–17, 25)。

これに近い動きは、計算科学の哲学ですでになされてきた。ハンフリーズ (Humphreys) は、科学的実在論者であるとは、何が存在するかについての人間中心的な立場を退けることであり、したがって人間の感覚を、それぞれ適用範囲の限られた五つの検出器にすぎないと認めることだと書いている

(Humphreys 2010, 112)。またハンフリーズは、ウィンズバーグ (Winsberg) とアルバラード (Alvarado) が伝えるところでは、過程が認識的に不透明であることを認知主体と時点で相対化し、ある主体がある時点でその過程の認識的に重要な要素をすべては知らないこととして定義した。そして、主体が自らの本性のゆえにそれらの要素を知りえないとき、その過程はその主体にとって本質的に不透明であるとした (Winsberg and Alvarado 2026, §2.5; 第3章第5節)。科学的実在論をめぐる論争では、科学が目指すのは観察可能なものについて正しい理論だとする構成的経験論 (constructive empiricism) をとるファン・フラッセン (van Fraassen) が、**観察可能** (observable) を私たちにとって観察可能という意味に読み、その私たち、すなわち認識的共同体の成員は人類と一致するとは限らないとした (van Fraassen 2005, 3-4; 第1章第6節)。ハンフリーズの感覚についての議論は、人間の感覚を特別扱いすることを退ける。それに対して、主体に相対的な不透明さの説明と、ファン・フラッセンの観察可能性の説明は、概念をほかの探究者にも当てはめられるように述べ直す、より直接的な例である。本書が議論ごとに試みるのは、この種の述べ直しである。

5. なぜ本書は科学だけでなく研究から出発するのか

本書の対象には、自然科学、数学、工学、社会的探究、人文学、および関連する領域が含まれる。本書は、実験的な自然科学を標準的な事例として出発し、ほかの実践を類似性によって受け入れるという方法をとらない。証明、史料解釈、フィールド研究、人工物、実験室での実験は、互いに大きく異なる。共通の特徴づけの目的は、ひとつの実現をほかのすべての実現の存在論にすることなしに、比較を可能にすることである。

研究 (research) が類であり科学 (science) がその一つの種であることを確立するような、時間を超えた分類法は存在しない。歴史研究は、近代における科学と人文学の区分の多くを長い19世紀に位置づけ、修飾語のない英語の *science* が、いかにして一般に物理科学や自然科学を指すようになったかを記述している。researchは、注意深い探索や調査を意味する語から専門的な学術活動を指す語へと発展し、さらに大学、政府、産業、研究開発にまたがる組織化されたカテゴリーとなった (Krämer 2018; Godin and Schauz 2016)。これらの歴史は、現代の語彙がなぜ重なり合っているのかを説明するが、正しい概念上の境界を決着させるものではない。

制度上の分類は、*研究* という語を、数学、自然科学と社会科学、工学、医学、人文学、芸術、パフォーマンス、人工物、デザイン、試験的開発にわたって広く用いている (OECD 2015; Research England et al. 2019)。これらは哲学的定義ではなく実務上の分類であり、日常的な研究 (research) という語は、商業的、ジャーナリズム的、個人的な調査をも含む。この語は、それだけでは対象を画定しない。ここでは、この語は、広い比較クラスを指すための活動語として用いられ、その範囲は明示的に述べられる。

隣接する探究の哲学は、有用な区別を提供してくれるが、確立した上位クラスを提供してはくれない。探究は、説明によっては、真なる信念、正当化された信念、知識、理解、あるいはより広い認識的改善へと向けられうる。探究はまた、より広い共同体がすでに知っている何かについての、ひとりの人の理解を拡大することもありうる。本書は、研究の特徴づけを提案するまえに探究の理論を決着させる必要はない。しかし、特徴づけが依拠しているのは、未知のものへと向けられていること、調査上の規範に従うこと、問いに決着をつけること、知識やその他の認識的達成をもたらすことのあいだの違いが保持されることである (Haziza 2023; Kelp 2021b, S1583-S1584; El Shazly 2025, 2-5, 12-13)。

したがって、研究という語の選択は、語彙上の立法ではなく目的相対的なものである。その価値は、知識、理解、その他の認識的達成に向けた異種的な形の探究を最初の比較クラスの内部にとどめておくことにある。抽象化が意図しているのは、学問分野の違いを消し去ることではなく、操作、基準、産物、そしておそらくはいくつかの目的が変化することを許すような目標を可能にすることである。このように変化を許す理由の一つは、機械が、いま営まれているとおりの研究を行うだけでなく、研究の新しいやり方を作り出すようになる場合を比較のなかに残しておくことにある。そうすれば、その活動、参加者、あるいはそれを支える関係が人工のものになると何が変わるのかを問うことは、研究のやり方そのものの変化を問うことも含むことになる。そうしたことが起こるのか、どのように起こるのかを本書は決着させない。それについて検討を通じて何が見えてきたかには、第13章第7節で立ち戻る。

6. 議論の道筋

次章では、あえて多くを決めない、ひとつの薄い方向づけを提案する。それは研究の唯一の本質を発見したと主張するものではない。本書の残りの部分では、その方向づけを足場に、互いに結びついた五つの問いを考える。

1. 研究についてのどのような構想が問われているのか。
2. 何が認識的達成とみなされるのか。
3. 問い、方法、基準が変化しうるとき、探究はどのように組織されるのか。
4. どこまでを一つの行為者とみなすかの境界を、人工システムや、人と機械の組み合わせに引いたとき、そこに研究能力、研究者としての地位、あるいは自律性を帰属させることを何が根拠づけるのか。そして、その研究はどの社会のために組織され、意図する達成はどこで実現するのか。
5. 研究を、その意図する達成の新しさの基準となる社会 (第1章の *S*) のために方向づけ、認定し、統治する関係のうち、どれを人工の構成の内側に置くことができ、置くと何が変わるのか。ここで人工の構成とは、研究が帰属される単位のうち人工のものを含むもの、すなわち一つのモデル、エージェント、その集まり、人間とAIからなる組織などである (第7章第1節)。いまは研究の外にいる者や仕組みが担っている関係を、そうした単位そのものが担うようになることを、「内側に置く」と言う。

第1章は、特徴づけを提案することで、第一の問いについて暫定的な見通しを示す。残りの四つは、第1章第9節が挙げる四つの負債、すなわち特徴づけが後の章に答えを負わせる問いに対応する。達成、過程、帰属、研究体制の負債である。研究体制とは、研究の分業、探究の配分、批判と認定、記憶と訂正、功績、権威からなる、研究を取り巻く社会的な秩序を指す。この分け方がどのような整理で、何のためのものかは、第1章第9節で述べる。各問いは本書の五つの部のひとつを組織する。第I部 (第1章) は第一の問いを、第II部 (第2章と第3章) は第二の問いを、第III部 (第4章から第6章) は第三の問いを、第IV部 (第7章と第8章) は第四の問いを、第V部 (第9章から第12章) は第五の問いを扱う。第13章は結論である。

冒頭の方向づけは、認識的達成を開かれた成功語、すなわち成功したことを表すが、どの種類の成功かは先に限らない語として用いる。主に分析するのは知識と理解であり、それ以外の達成は名指すだけで確立されるわけではない。本書の議論は、多くの箇所では知識の場合を中心に進む。何をもって達成が成り立つとみなすか、証拠をどう評価するか、結果をどう認定するかを論じる章（とくに第1、2、6、8、11章）がそうである。理解は第3章で取り上げ、知識とどう関係するかを軸に、何をもって理解が成り立つか、誰が理解するのか、理解にどのような価値があるのかを検討する。理解が認定と帰属にもつ帰結は、第9章第2.1節や第11章第5.6節などでも扱う。一方、探究の過程、研究の帰属、探究の配分、決定の権限を扱う章の多くは、知識にも理解にも同じように当てはまる書き方になっている。認識的達成という語によって、知識や理解など複数の達成を構築の目標として扱えるが、それぞれの成立条件は擁護を要するまま残る。この語がそれだけではひとつの成功条件を定めないという代償は、第1章第1節で明示的に評価する。

議論の順序は一方向ではない。後続の章は特徴づけに依拠するが、そこで分かったことが特徴づけの修正を求めることもある（第1章第1節）。後続の章が主に考えるのは、研究をこのように捉えたとき、機械が研究することについて何が含意されるかである。本書は作業的特徴づけを試すことを目的としないが、第1章第8節では、特徴づけを多様な研究実践と比べ、どこまでを覆い、どこで別の捉え方と分かれるかを見ておく。第13章では、特徴づけが、研究するのが機械である場合に何を区別し、何を問えるようにしたのか、何を脇に置いたのか、どの結論が特徴づけ以外の説や事例に依拠するのかをたどる。また、変化する問いと方法から研究行為者の成立、それらのあいだの関係の再編まで、各章が人工システムによる研究について見ようとしてきたことを結び合わせる。哲学上のあらゆる不一致を先に解決しなくても、この探究を通じて目標をより明確にできる。

各章を通じて見えてくることのうち、第13章がまとめるのは二つである。一つは、機械が研究しているという主張が、別々に成り立ったり成り立たなかったりする複数の主張に分かれることである。もう一つは、人間が研究するときには、研究する者が一人ひとりの人であり、多くの人に分かれていたために決めずに決まっていたものの多くが、機械が研究するときには決めるものになることである。

第I部 何を研究と捉えるのか

第1章 研究の作業的特徴づけ

特徴づけに必要なもの、必要でないもの

1. 認知的達成への薄い方向づけ

ここで必要な特徴づけが担う仕事は、限られているが、求められることは多い。ひとつの分野の仕事の進め方を研究そのものの形とみなすことなく諸分野にまたがること、そして、人工物が何をしているのか、人々と制度がなお何をしているのか、評価が何を確立するのかを問えるところまで、人工研究者という目標をはっきりさせることである。研究の時代を超えた本質を特定する必要はなく、範囲を限ったエンジニアリングを始める前に、争われている境界のすべてに決着をつける必要もない。

この限定が重要なのは、候補となる定義が異なる問いに答えるからである。研究活動であるための条件は、自動的に、良い研究や信頼できる研究であるための条件になるわけではない。たとえば、知識を産出することは、理解を産出すること、進歩をなすこと、研究システムのなかで有用な役割を占めること、制度的な承認を受けること、あるいは重要な何かを産出すること、必ずしも同じではない。知識の産出が理解の産出とどう関係するかは、理解を知識の一種とみるかどうかによって変わり（第3章第3節）、進歩をなすことと同じかどうかは、進歩についてどの説をとるかによる（第2章第8.3節）。以下で提案する作業的特徴づけのもとでは、失敗したプロジェクトも研究であったことはありうる。制度的に承認されたプロジェクトが認知的に欠陥を抱えていることもありうる。データのキュレーションは、それが支える研究上の問いをそれ自体としては解決しないままに、大きな進歩を可能にすることがある。

研究を特徴づける候補には、操作によるもの、達成によるものもある。統制された実験、仮説検証、論文産出のような操作で研究を特徴づけると、証明や解釈のような実践は、大きく記述し直さなければ取まらない。説明、理解、問題解決のような達成で特徴づけると、別の問いが生じる。これらの実践を、そうした達成に向けられた探究として捉える方が、それらがしていることをよく表すのか、という問いである。第5章の領域横断的な事例は、研究を一つの操作と同一視しない理由を与えるが、どの達成に向けた方向づけが適切かまでは決めない。知識や理解のような達成も同じ比較にさらされる。その適切さも、操作による候補と同じく、何を明晰にし、何を排除するかで判断される。

人工研究者をつくるうえでは、この比較は特に重要になる。人工システムは新しい出力を大量につくれるが、それによって認知的に何かを成し遂げたことになるかは別の問いだからである。そのため、新しい出力をつくることと、認知的達成をもたらすことを区別しておく必要がある。そこで本書は、研究を、特定の操作ではなく達成への方向によって特徴づけ、目指す達成の種類は先に決めずにおく。この方向づけを、試み、認知的達成、社会にとっての新しさという三つの語で表す。

本書の作業的特徴づけは次のとおりである。

ある社会 S に相対的に、研究とは、 S にとって新しい認知的達成をもたらそうとする試みである。

この文は、研究の定義を確定させて守るためのものではない。人工物が研究するとはどういうことかを考えるための足場として置く作業的な特徴づけであり、それ自体も検討にさらされる。各語をこの形で置いた理由は、以下と第2節以降で述べる。

ここで認知的達成とは、認識上の成功を指す開かれた総称であり、本書では主として知識と理解を通じて検討する。この語は網羅的な分類を与えず、達成が同じ構造をもち、同じ価値をもち、あるいは互いに排他的であるとも仮定しない。その実現が何を含むか、主体による保有を必要とするかも、採る説明による。また、この語自体は、重大な障害の克服や相当の認知能力の発揮を要件にしない。そのような強い意味での達成は、第3章第7.1節で検討する。

研究を知識の産出として捉えるか、理解をもたらす営みとして捉えるかで、研究のどの側面が見えてくるかは変わる。本書が認知的達成という総称を用いて両方を扱うのはそのためである。ただし本章では、作業文の各語が何を求め、何を求めないかを、おもに知識を例にとって説明する。知識については、その成功が提起する問い（真理、誰に帰属するか、偶然でないこと）や社会的な帰属が認識論と社会認識論で詳しく論じられており、各語の働きを確かめやすいからである。第2章が命題的知識から分析を始めるのも、同じ理由による。理解については、違いが出る箇所でのつど触れ、まとまった検討は第3章で行う。理解の理論は、理解が知識の一種なのか、それとも知識とは別の、おそらくはより価値のある認知的達成なのか、そして理解の説明的な形態（ある事柄がなぜ成り立つかの理解）と対象的な形態（ある現象や主題領域の全体の理解）とがどう関係するのかを問う（Grimm 2021）。この問いも第3章第3節で扱う。

ここで試み（attempt）は、組織された方向性を示すのであって、必ずしもすべての参加者が抱く意識的な意図を示すのではない。特徴づけは、活動の何らかの特徴——その計画、機能的な組織、役割の配分、証拠への応答性、あるいは研究実践のなかでの位置——が、なされることを、見込まれる認知的達成へと結びつけていることを求める。それ自体は何にも方向づけられていない出来事から、のちに偶然に知識が引き出されうる、というだけでは足りない。どのような種類の組織が必要とされるのかは未決のままである。この要求の当面の目的は、分散した事例や非人間的な事例を保持しつつ、研究の試みを幸運な出力から区別することにある。

S には二つの役割がある。一つは、新しさを測る参照先としての役割であり、もう一つは、その達成がどの社会のために求められるか、つまり試みが向けられる先を定める役割である。どちらの役割も、達成を S に帰属させるものではない。 S は、人類、専門家の共同体、組織、人間と人工物からなる集団、ありうる人工の共同体のいずれでもありうる。ただし、ある出力を新しく見せるために事後的に寄せ集めた集合ではない。研究を行う者が S と同じである必要はなく、 S を名指すことは、達成の担い手を決めることでもない。これらは第2節と第6節で述べる。

先行する短い論考は、知識の産出を通じて研究を特徴づけ、そのうえで、その特徴づけの構成要素として問いを導いていた (Takagi 2023)。本書は、試みと社会に相対的な新しさを保持しつつ、目指す達成を広げ、「産出する」を「もたらす」に改める (第4節)。理解が知識の一種かどうかを先に決めたり、理解を知識への手段としてだけ扱ったりせずに、第II部で検討する選択肢を全体の方向づけに接続するためである。

この文は、本書で検討する目標の構想を特定する。必要十分な境界界定の基準でも、完全な研究の哲学でもない。広く読めば、教育、報道、調査、発明、日常の問題解決にも当てはまりうる。逆に、認識的な方向づけは、制作、参加、関係、変容を中心とする実践を誤って記述しうる。この方向づけが、研究と呼ぶことが適切なあらゆる実践に必要なか、未決のままである。これらの圧力は、第4節と第8節で検討する。第4節は産物、役割、達成を区別し、第8節は、方向づけが捉えるものと、それが実践に課す区別の両方を、多様な研究実践と比べる。

ホイニンゲン＝ヒューネ (Hoyningen-Huene) の科学論は、本章と似た問題、すなわちきわめて多様な実践について一般的なことをどう言うかという問題に向き合っている。ホイニンゲン＝ヒューネは、科学的知識は他の種類の知識、とりわけ日常的な知識から、主としてより体系的であることによって区別されると主張し、この比較的なテーゼを九つの次元に沿って展開する (Hoyningen-Huene 2018, 3, 7 n. 12)。このテーゼは知識の集まりを比べるもので、研究という活動を比べるものではない。ホイニンゲン＝ヒューネは、**体系性**という抽象的な語が曖昧であることを認め、それと対比される語 (でたまため、偶然的、恣意的な、無計画な、無秩序な) によって得られる精確さも高くないとする。抽象の水準ではそれ以上のことはできないようなので、この語は文脈のなかで具体化されなければならない、というのである。体系性の具体的な諸概念は、次元、分野、時代によって異なり、必要十分条件によってではなく家族的類似、すなわち全体に共通する一つの特徴はなくとも部分的な類似が重なり合うことによってだけ関係づけられている。そのため、このテーゼが主張する諸科学の統一は、ホイニンゲン＝ヒューネの説では極めて弱いものであり、すべての次元がすべての分野に当てはまるわけでもない (たとえば予測は当てはまらなくてよい) (Hoyningen-Huene 2020, 89–91)。この説では、科学は非科学からどこでも鋭く区切られているわけではない。科学的な研究と、実用の目的に合わせた関連する活動とのあいだには移行領域があり、そこでは必然的に粗く曖昧な区別しか引けない (96)。ホイニンゲン＝ヒューネは2018年の応答論文で、自らの比較が示すのはせいぜい体系性が科学にとって必要であることであって、十分であることではないと明言した。そして、この説の境界設定の基準は、疑似科学とみなされるものの全体的な体系性が一定の期間にどれだけ増したかを、それと競合する確立した科学での増し方と比べるものだとする。体系性の量的な尺度は疑わしいが、この比較には量的でない比較の語で足りるという。また、歴史家による事例研究を、この説を具体的に試す場として歓迎している (Hoyningen-Huene 2018, 1, 3–5)。作業的特徴づけが抽象的なのも、同じように、多様な研究実践について空疎でない一般的なことを言うためである。

ホイニンゲン＝ヒューネはまた、上で挙げた第一の圧力、すなわち広く読めば隣接する活動にも当てはまるという圧力に、活動の側からではなく知識の側から向き合っている。その議論では、信頼性への要求が高い、専門家による知識の産出の領域、たとえば凶悪犯罪連関分析システム (VICLAS)、市場調査、製品開発、チェス理論、航空機事故調査などは、きわめて体系的に進められうるが、その産物は明らかに科学ではない。違いをホイニンゲン＝ヒューネは、自ら認識的な連結性 (epistemic connectedness) と呼ぶもの、すなわち知識が他の知識の断片と結びついていることに置き、こうした知識はそれを十分な程度にはもたないとする (Hoyningen-Huene 2018, 2–3)。同じ出来事を扱う歴史学の雑誌の論文と質の高い新聞記事とは、伝える要点が同じでありうるが、論文は多くの注によって、知識どうしの結びつきを明示する (Hoyningen-Huene 2020, 96–97)。本章の作業文と比べると、先に挙げた専門家の事例は、どれも知識の産出に向けられている。したがってそれらの事例は、ホイニンゲン＝ヒューネがこれらを科学から除くように研究からも除くとすれば、*S* にとって新しい知識への方向づけだけではそれができないことを示唆する。ホイニンゲン＝ヒューネ自身の基準は、上で見たように、新しい知識を他の知識に結びつけることにかかわる。*S* にとって新しい知識という主張が前提する、*S* がすでに知っていることとの関係が、同様の働きをなしうるかどうかは、本書では追わない。第8節で述べるように、本書は、認識的な方向づけが研究をあらゆる隣接活動から区別するとは主張しないからである。

この文がまとめるものと、別の語彙が要るもの

この特徴づけは、活動の大まかな方向性の水準で働く。研究の産物の存在論、プロセス・モデル、科学的進歩の理論、社会的価値の尺度を一度に兼ねるためのものではない。これらの仕事は、ここで提案する方向づけを手放さずに、切り離して扱うことができる。

1. **活動の水準**では、研究は、*S* にとって新しい認識的達成をもたらそうとする組織された試みとして、改訂の余地を残しつつ捉えられる。
2. **過程の水準**では、探究は目標に達する前に、問い、証拠、信念度 (確信の度合い)、概念、方法、能力、認識された無知、社会的な利用可能性を変えうる。
3. **達成の水準**では、本書は主に知識と理解、およびその実現について競合する説明を検討する。説明、概念の再編、能力の向上は、主張に応じて産物、手段、達成でありうる。それらを名指すだけでは、新たな達成の種類や成功条件は確立されない。
4. **価値とガバナンスの水準**では、重要性、有用性、美、解放、正統性、リスク、制度的認定をそれぞれ評価する必要がある。それによって、これらがあらゆる認識的達成や研究としての地位の隠れた条件になるわけではない。

これらは分析上の水準であって、四つの独立したプロセスでも普遍的な時系列でもない。ひとつの証明や実験がそのうちのいくつかを同時に変えることもある。同じように、試みをどれだけ長く持続させうるか、そのエピソードがどこで区切られるかを定める取り決め、たとえば助成や学位論文 (第3節) は、四つ目の水準で評価されながら、活動と過程の水準にかかわる。そうした取り決めも、研究としての地位の条件にはされない。四つの水準を分けるのは、正反対の二つの誤りを避けるためである。成功を表す語をあまりに広くして、価値あるあらゆる変化の別名にしてしまう誤りと、共通の方向づけを、分野ごとの善を構造なしに並べた一覧で置き換えてしまう誤りである。ただし、第8節の説や第11章第4節で取り上げる説のように、価値が、根拠づけ (warrant)、つまり主張を受け入れることの認識的な支えや、受容そのものに入るとする説では、四つ目の水準はこのように他の水準から切り離せず、価値の判断が、結果が受け入れられるかどうかにかかわりうる。

認識的な方向づけは、望ましい効果、新しい人工物、受容された言明を、認識上の成功から区別する。それにより、構築の目標は、何が成し遂げられ、どこで実現し、成功が誤り、偶然、有用だが不十分な寄与とどう区別されるのかについての説明に応答すべきものとなる。これらは問いであって、一つの理論の答えをすべての達成に押しつけるものではない。

作業文は、意図される達成を、認識的達成という開かれた総称で名指している。そう名指すことには代償がある。認識的達成という総称は、それだけでは成功について多くを定めない。認識的達成があったという主張だけからは、知識が得られたのか、どの命題が（もしあるとすれば）知られたとされるのかは分からない。それが分かるのは、どの種類の達成を主張しているかを特定したときである。

ただし、認識的達成という総称を用いることは、成功の条件を緩めることではない。ある命題の知識を主張するなら、認識論者の大半がとる見方では、その命題の真理が求められる（第5節）。理解も、第3章で比較する説のもとで、その対象に応答することが求められる。理解の説のなかには、理想化されたモデル、つまり意図的に偽を含むモデルが理解に寄与すると認めるものもあるので（第3章第3.4節）、この語を使えば真理の条件が緩むように見えるかもしれない。しかし、理想化された表象は知識に向けた探究にも寄与しうるので、理想化をどこまで使えるかは、この語ではなく、それぞれの達成についての説が決める。この語が定めるのは、達成の種類を特定する前に言えることまでであり、個々の知識や理解の主張には、それぞれの成立条件が求められる。

また、認識的達成を開かれた総称にしておくことは、何でも達成に数えることでもない。概念、方法、装置、能力を名指しだけで、それらが達成の一種として確立されるわけではない。あるものが達成の一種であるかどうかと、本書がそれをどこまで分析しているかは、別の問いである。実践的、美的、関係的な善も、それぞれの意味で評価される。価値があると呼ぶだけでは、それらは認識的なものにならない。

この特徴づけはひとつの大域的な目的を確立するものではない

この定式は、研究とは何かを、何をもちたそうとする試みかによって述べているので、容易に*研究の目的*についてのテーゼとして読まれうる。そう読むと、この定式は本書が意図するよりはるかに強い主張になってしまう。ある実践が目的をもつという主張は、重要な点で異なる諸関係を指しうる。すなわち、実践者の意識的な意図、集団的あるいは団体的な目標、制度の歴史的または現在の機能、それなしにはある活動が研究でなくなるような構成的特徴、成功の基準、規範的な理想、ある実践を理解可能にする解釈などである。ここに挙げた意味は、以下で論じるラスク（Lusk）とエリオット（Elliott）が検討する根拠と、ヒュッテマン（Hüttemann）が展開する解釈の道筋に、実践者の意図を加えたものである。論証なしにこれらの意味のあいだを行き来すれば、**試み**という語に、それが担いうる以上の哲学的な重みを負わせることになるだろう。

研究や科学の目的についてこうした強いテーゼを実際に擁護する立場もある。バード（Bird）は、科学を認知システムの社会的な機能的類似物として扱い、知識が科学の構成的な目的であると論じる。この説では、適切に機能する認知システムが生み出す信念が知識であるのと同じように、科学が適切に機能するときに生み出すものも知識である（Bird 2019, 171–172, 176, 178–179）。バードはこの論文で、科学が真理ではなく知識を目指すことを、何が進歩とみなされるかから論じる。信頼できない方法によって真である理論を信じるに至った共同体は、それによって進歩したことにはならず、その信念を捨てても後退したことにはならない。ある活動が進歩するのはその目的に近づく限りにおいてである。そうであれば、真理や真理近似性の単なる蓄積が進歩にならない以上、科学の目的は真理ではなく知識だ、というのである。理解と問題解決については、別の場での議論に委ねている。また、科学が目指すのはその理論の経験的十全性だとするファン・フラッセン（van Fraassen）の提案は、目的がどの範囲の主張に及ぶかについての問いであり、自分の問いとは独立だとして脇に置く。目的を観察できるものに限っても、科学が目指すのは観察上の帰結が真である理論なのか、それが真だと知られている理論なのか、というバードの問いは残るからである。バードの問いは、目指される成功の種類にかかわり、その範囲にはかわらない（Bird 2019, 173, 176; Mohler and Monton 2025, §1.1）。第2章と第3章は、これらの代替案に、競合する達成として立ち戻る。バードののちに、同じ方向の主張を、目的ではなく機能の語で述べている。知識を産出する機能が、客観的な科学の規範を直接に基礎づけうるし、何が科学とみなされるかを構成するのに役立つという主張である（Bird 2025a）。

これに対して、研究や探究にひとつの大域的な目的を立てる必要はないとする立場がある。ジェーン・フリードマン（Jane Friedman）は、探究のひとつの目的を指定することなしに、探究の構造と規範、そして探究者の目的について直接に理論化できると論じる。フリードマンの説では、探究は、ゲームのように人間が取り決めて作ったものではないので、目的が取り決めによって定まるゲームの類比は探究には当てはまらない。また、探究者がみな、知識のような特定の心的状態を得ると同じ一つの目的を共有している必要もない。知識ではなく高い信念度を得ようとする者も、探究していることに変わりはないからである（Friedman 2024）。科学の水準では、ロウボトム（Rowbottom）が、科学の目的（単数であれ複数であれ）についての語りをやめ、より特定の記述的、評価的、規範的、定義的な主張に置き換えることを勧めている。ただし、科学者や研究グループが目的をもつことは認めている。ロウボトムによれば、そうした語りを科学者が何を目指しているかについての主張として文字どおりに読めば、哲学上の論争は社会学的な調査で決まる問題になる。他方、科学そのものにおいて何が成功とみなされるかという主張に言い換えれば、どの探究に価値があるかについての価値判断の違いが、科学についての事実をめぐる不一致のように見えてしまう（Rowbottom 2014, 2–5, 12–16）。もっとも、**目的**という語を避けるだけでは、研究にひとつの大域的な目的があるか、何がそれを根拠づけうるかという問いは決着しない。バードの機能の議論が示すように、同じ強い主張が別の語で戻ってきうるからである。

ラスクとエリオットは、この問いに正面から取り組む。二人の問いは、科学における価値をめぐる論争から生じている。非認識的価値、つまり証拠や知識にかかわる価値以外の価値、たとえば社会的、政治的、個人的な価値が科学的推論に影響することを、証拠が残す隙間においてだけ認める見解は、科学の大域的な目的に依拠しているように見える。ここで大域的な目的とは、あらゆる分野にわたって成り立ち、局所的な目的、すなわち個々のプロジェクトや学科や分野の目的のうちどれが受け入れられるかを限定する目的である（Lusk and Elliott 2024, 2, 4–6）。そこでラスクとエリオットは、そうした目的を何が正当化しうるかを問う。まず検討されるのは、誰の意図にも訴えない五つの根拠であり、二人によれば、そのいずれも説得力をもたない（7–11）。目的が真理や説明のような規範的理想だとすれば、どの理想がそれなのかを決めるものは、食い違う哲学者の直観のほかにはない。また、理想を実践から読み取ろうとすれば、理想に届かない実践と、はじめからその理想を目指していない実践とを原理的に区別する手続きが必要になる。目的が機能だとすれば、科学が存在する理由も社会への寄与も多様であり、信頼できる情報、予測、健康、戦争での成功、経済発展などにわたるので、ひとつの機能だけが際立つことはない。さらに、科学を認知システムになぞらえるバードの類比に対して、ラスクとエリオットは、この類比がひとつの認識的な機能を支持するのは、すでにひとつの認識的な目的を受け入れている者にとってだけだと反論する。新しい技術を生み出すことのような目的には、個々の生物の認知のうちに明白な対応物がないからである。目的が本質的特徴だとすれば、境界設定の基準、つまり科学を他の活動からはっきり区別する基準がないことが、そう

した特徴から目的を特定する試みを妨げる。そして、科学によく見られる特徴の緩やかな集まりに訴えれば、得られるのは、それぞれ局所的にしか成り立たない多くの目的である。目的が成功の基準だとすれば、何が勝ちかが規則で定まるチェスとの類比は成り立たない。科学にはそれほど明確な規則がなく、何が科学の成功とみなされるかこそが争われている点だからである。そのうえで二人は、先に見たロウボトムの論点に依拠して、この争いは価値判断にかかっており、科学者が何を意図しているかに訴えずに異論の余地なく決着させることはできそうにないと述べる。目的が団体的目標、つまり成員が何を望むにかかわらず組織がその構造を通じて追求する目標だとすれば、科学には一様な組織構造が欠けている。公的・私的な資金提供機関が、それぞれ異なる方針と誘因をもっているからである。ラスクとエリオットの説では、人々の意図に訴える根拠もうまくいかない。科学者たちは同じ目標を共有しておらず、目標の優先順位も同じではない。また、科学全体を設計した者は誰もいない。さらに、誰もが受け入れられるほど一般的な基準は、局所的な目的を限定するには内容が乏しすぎる。複数の大域的な目的を立てても事情は変わらない。それらのあいだには原理的な優先順位が必要になるが、それは誰も与えておらず、そうでなければ、それぞれの文脈で実際に働く目的は事実上局所的なものになるからである（12-15）。二人はこの分析を、決定的な反駁としてではなく、大域的な目的を立てる側に立証責任を移す概略的な検討として提示している（3, 6-7）。

大域的な目的に代えて、ラスクとエリオットは局所的目的アプローチ（local aims approach）を擁護する。これは、科学には局所的な目的しかないとする見解である。二人の説では、科学全体については失敗した根拠も、範囲を狭めれば目的を与えうる。たとえば、あるプロジェクトが何を達成するよう設計されたかや、ある分野の創設者がその分野に与えた使命から、局所的な目的を特定できる。そして、範囲が限られていることが、局所的な目的は先の反論を免れると二人が考える理由である（15-16）。また、ひとつのプロジェクトはふつう複数の局所的な目的を同時に追求するので、ラスクとエリオットはそれらを、パーカー（Parker）がモデル評価のために展開し、二人自身が方法や仮説にまで広げた目的適合性（adequacy-for-purpose）の見方によって結びつける。この見方では、そうした対象は、真であるかどうかによってではなく、特定された目的に適合しているかどうかによって評価される。そのうえで、ある文脈で探究が何のために行われるか（purpose）が、正確さや精度のような認知的な目的と、使いやすさや完了までの時間のような非認知的な目的とを、それぞれどこまで達成しなければならないかを定める（16-17, 20）。この見方では、非認知的な目的は認知的な目的に従属する必要がない。二人はこの立場を対等目的アプローチ（equal aims approach）と呼ぶ。それでも、二種類の目的は、その探究の目的（purpose）を果たすことの一部としてあわせて満たされるものとされるので、認知的な制約がなくなるわけではない。これに対して、大域的な認知的目的がなければ認知的な腐敗を防ぐものがない、という懸念がありうる。ラスクとエリオットはこれに、プロジェクトが何のために行われるかは、それが依拠する分野の局所的な目的からも、法的、規制上、倫理的な制約からも批判できると答える。とくに誠実さは、研究者に、自分の分野の認知的な基準に従うか、それから外れていることを明らかにするかのどちらかを求める（18-22）。

大域的な目的に対するラスクとエリオットの理由は、異なる前提に依っている。そのいくつかは、科学全体がこれまでどう組織されてきたかにかかわる。科学者が目標を共有せず優先順位も同じでないこと、科学全体を設計した者がいないこと、資金提供者の方針の違いから一様な組織構造がないことなどである。そうでない理由もある。理想についての哲学者の直観が食い違うこと、何が成功とみなされるかこそが争われていること、誰もが受け入れられるほど一般的な基準は局所的な目的を限定するには内容が乏しすぎることなどである。二人は人工の構成（研究を担う人工システム、または人工システムを含む組み合わせ。序論第6節）を論じていないが、組織が一つにまとまった人工の構成に当てはめれば、前者の理由は成り立つとは限らない。そして、プロジェクトが何を達成するよう設計されたかから局所的な目的を特定するのと同じように、その組織から目的を特定できるかもしれない。そうした目的が二人の意味で大域的か局所的かは、その人工の構成を研究の全体とみなすかどうかによる。残りの理由は、その組織のあり方に左右されない。バードの類比に対する二人の反論も、そうした構成については形を変える。その反論は、新しい技術を生み出すことのような科学の目的には個々の生物の認知のうちに明白な対応物がないのに、科学全体を認知システムになぞらえることにかかっていた。組織が一つにまとまった人工の構成は、それ自体が、バードの類比が科学をなぞらえる先の種類のシステムでありうる。そのとき、そこでひとつの認知的な機能が際立つかどうかは、その構成が何をするように作られ、選ばれてきたかによる。ただし、実際の選択が何に報いてきたかは、掲げられた目的と食い違いうる（第2章第4.1節）。

目的の語り方を保つ見解のあいだでも、その目的が知識かどうかには争いがある。ベダー（Beddor）は、探究は知識を目指すという見方が、合理的な者は絶対的に確実でなくても何かを知りうるというテーゼと両立しないと論じる。このテーゼは可謬主義と密接に結びついている。可謬主義とは、証拠が p の偽である可能性を排除していなくても、 p を知りうるという見方である。可謬主義に、証拠が排除していない可能性があるなら完全に確実であってはならないという原則を加えると、このテーゼが導かれる。ただし、このテーゼは可謬主義より弱く、可謬主義を退ける者も受け入れうる。ベダーの議論は、ある事例から始まる。ある科学者が、実験にもとづいてある仮説を知っている。それでもその科学者は、結果がすべて交絡要因によるものだという可能性に、わずかな信念度を合理的に割り当てている。そこへ、別の研究者から、より包括的な研究のデータが届く。この科学者にはそのデータを見る理由があるように思われ、しかも見ることは同じ問いについての探究の続きである。ところが、探究が知識を目指すのなら、その科学者はすでにその目的を達成しており、探究を続ける意味はないことになる。ベダーはこの事例を、認知的決定理論にもとづく議論で補強する。認知的決定理論は、合理的な選択の理論である決定理論が行為を評価するのと同じように、信念度を認知的な観点から評価する。真な命題への信念度は、高いほど正確、つまり真理に近い。そして、最大限に正確でない信念度は、最大限の認知的価値をもたない。そのため、探究が信念度をできるかぎり認知的価値の高いものにすることを目指すのなら、確実さなしに知っている者は、まだその目的を達成していない。二つを両立させる試みはいずれもうまくいかないとみて、ベダーはどちらを手放すかを問い、知識に確実さを求めるほうが代償が大きいと判断する。人々は、ある人が完全には確実でないと認めながら、その人が何かを知っているとためらわずに言い、また、知っていることと確実に知っていることを区別しているからである。そこでベダーは、知識を探究の目的とする見方を退け、代わりに、ある問いについての探究の目的は、その答えへの信念度をできるかぎり認知的価値の高いものにすることだと提案する。そのためには、真の答えに完全な確信、すなわち信念度1を割り当てることが必要である。この説では、知識を目的とする見方が人を引きつけてきた理由の一部は、知識が認知的価値をもつと考えられてきたことにあり、この提案はその価値を直接に目的とする。さらにこの提案は、確実さに届かない知識を認め、認知的決定理論ともよく合う。ただしベダーは、確実さに達するまで探究を続けなければならないとは結論しない。探究を続けることが合理的かどうかは、ほかの探究のほうが大きな認知的価値の期待値を見込めるかどうかにもよるからである（Beddor 2024, §§1-3, 5-6）。探究の目的をめぐるこの争いには、第3節で立ち戻る。

もう一つの道は、目的を帰属させるとはどういうことかについて、別の答えを与える。目的の帰属を、典型的な科学的実践を合理化する、つまり筋の通ったものとして理解できるようにする解釈として扱う道である。ここでは、目的は実践の構成的な特徴として立てられるのでも、語りから外されるのでもな

く、実践を理解可能にする読み方として帰属される。ヒュッテマンはこの道筋を展開すると同時に、目的をこの仕方では帰属させるなら、科学は知識を目指すというバードのテーゼに、科学がどの種類の知識を目指すのかをより詳しく特定するという限定を付す必要が生じうることも論じている。ただし、テーゼそのものには異を唱えていない。限定が必要になりうるのは、実践が追い求めるのがどんな知識でもよいわけではないからである。科学者は、たやすく得られる知識を集める代わりに、莫大な費用をかけて加速器を建設する。また理想化についてのある説では、理想化は系の振る舞いを誤って表しながら、その振る舞いの根底にある構造を特定するのに役立つ。そうであれば、実践を合理化するのは、ある種類の真理は目指すが別の種類は目指さない、という目的である (Hüttemann 2025, 1-3, 5-6)。この道筋をとれば、目的の帰属は実践に照らして試されることになり、帰属される目的が強いテーゼより限定されたものになることもありうる。

こうして、これらの立場は、科学や探究がどの目的をもつのか (バードとベダー) だけでなく、目的を帰属させることが何に帰着するのか (ヒュッテマン)、そして何がその帰属を根拠づけるのか (バードとラスクとエリオット) についても意見を異にしている。

したがって、作業的特徴づけが引き受けるのは、取り消し可能な解釈上の主張である。方向づけを、研究実践に照らして改められうる解釈として扱う点で、これはヒュッテマンが展開する道筋に近い。ある S にとって新しい認識的達成への方向を提案するが、それを、全参加者の意識的な意図、制度の歴史的機能、例外のない本質とは同一視せず、ラスクとエリオットの議論を踏まえて、プロジェクトが追求しうる局所的な目的を限定する大域的な目的とも同一視しない。ラスクとエリオット自身も、科学全体について述べられた目的であっても、局所的な目的の追求にそうした限定を課さないなら、自分たちが退ける種類の目的ではないと認めている。二人はこの点を、自らの目的適合性の見方について述べている (Lusk and Elliott 2024, 17 n. 5)。この点で、ここで提案する方向づけはそうした弱い種類の主張である。認識的達成を、プロジェクトの他の目的より上に置くことを求めないからである。プロジェクトは、予測、構築、ケア、解放、商業的有用性を追求しながら、認識上の成功にも向けて組織されうる。本書はこの解釈を試すことを目的としなが、第8節では、特徴づけを多様な研究実践と比べ、どこまでを覆い、どこで別の捉え方と分かれるかを見ておく。

この限定は、人工研究者をつくることに直接に重要である。この定式は、研究を「知識を最大化せよ」という一つの量を最大化させる指示に置き換えてよい根拠にはならない。それは、どの未知のものが注意に値するのか、誰の目的が探究を統べるべきなのか、認識的価値と非認識的価値がどのように相互作用すべきなのか、どのようなリスクが許容されるのか、あるいはいつ結果が決定を変えるべきなのかを決定しない。これらは、認識的な目的から自動的に導かれるものでも、あってもなくてもよい付け足しでもない。この定式が決めずにおく構築とガバナンスの問題であり、第V部は、それぞれの説が何を含意するかを問うことで、それらを取り上げる (第9章第1節)。

2. 各語が担う働き

第1節で示した作業文を、もう一度掲げる。

ある社会 S に相対的に、研究とは、 S にとって新しい認識的達成をもたらそうとする試みである。

この文の各表現は、それぞれ限られた働きだけを担い、それ以上のことは決めない。たとえば、どの語も成功を保証せず、達成の担い手を特定しない。各語の詳しい扱いは第3〜6節で行う。

- **試み**は、組織された活動を、成功した達成から区別する (第3節)。
- **もたらす**は、認識的達成の実現に注意を向ける。それは産物をつくることと同じではなく、一度の閾値を越えることである必要も、知られたり理解されたりする対象を創造する必要もない (第4節)。
- **認識的達成**は成功を表す。それを主張するには、どの達成を主張し、それがどの条件のもとで成り立つかを明らかにする必要がある。本書では主として第II部の知識と理解の分析を通じてこれを行う (第5節)。
- **S にとって新しい**は、意図する達成を、研究者だけの以前の状態でなく、社会の以前の認識的状況と比較する (第6節)。
- **社会 S** は、参照先を、専門共同体、人類、組織、混合集団、ありうる人工の共同体へ開く。その記号だけでは達成の担い手は特定されない (第6節)。

第1節で見た S の二つの役割のうち、その達成がどの社会のために求められるかを定める役割は、試みの方向に関わるものであり、社会が利益を得ることや達成が実現することの保証ではない。どちらの役割も、それだけで達成を集団の主体としての S 、その成員、あるいは公共的な資源の配置に帰属させるものではない。知識と理解は、これらの関係についてそれぞれの説明を要する (第2〜3章)。参照先を事後的に自由に選ぶことはできない。活動が先行成果をどう扱い、誰に成果を向け、どの基準に応答するかが選択を制約する (第6節)。複数の参照先が擁護できる場合、主張を評価するには、どれを用いているかが明らかである必要がある (第2章第6節)。この記号からは、ひとつの社会的な心があることも、保存、アクセス可能性、受容、知識、理解が互いに同じものであることも導かれない。

研究についての主張には、二つの当事者がかかわる。試みを行う行為者あるいは組織されたシステム A と、達成がそれにとって新しく、そのために求められる社会 S である。これまで研究をしてきたのは人間なので、研究を行う者もその成果を保有する者も、当然ながら人間であった。ハンフリーズ (Humphreys) は、科学は人間の科学であるという想定が、明示的にであれ暗黙のうちにであれ、これまでのすべての科学哲学を動かしてきたと指摘している (Humphreys 2009b, §2)。ハンフリーズは後の論考で、自動化された科学をめぐる議論でしばしば暗黙のうちに働いている、と自ら言う論証を名指している。科学は人間によって、人間の利益のために行われる活動なのだから、科学における究極の認識的権威はつねに人間であるべきだ、という論証である。ハンフリーズは、人間中心主義の魅力にはもっともな理由があると認め、科学の追求とその産物の使用に対する道徳的責任を人間が保持することは不可欠だとする。それでも、知識を拡張して捉えるなら、「人間の現在の理解を超えた」科学的知識を認めるべきだと論じる (Humphreys 2020, 12, 14)。したがってこの説では、科学が人間のために行われることは、それだけでは、人間を科学の究極の認識的権威とすることも、科学が知るようになる

ことが人間の現在の理解の内にあることを求めることもない。ハンフリーズはすでに、シュラーベ（Schrape）が引用する人類の絶滅後も続く科学の事例で、人間によって行われる科学と人間のために行われる科学とを分けていた。ある活動が科学的とみなされるのに、私たちによって行われることも、私たちのために行われることも要らず、科学を行う者も、科学が役立つ目的の持ち主も、人間でなくてよいと結論したのである（Schrape 2015, 162–163; 事例は第8章第1.3節で述べる）。ウェスターブラッド（Westerblad）も、ハンフリーズが科学を人間の行為者と受益者に限ることを退けたと報告している（Westerblad 2026, §2）。ここで区別した二つの当事者は、この区別を引き継いでいる。「によって」の側に *A* が、「のために」の側に *S* が当たる。ただし *S* はハンフリーズの言う受益者ではない。その区別での「のために」は、活動の成果が誰の目的に役立つかわかる。これに対して *S* は、達成がそれにとって新しいとされ、試みがそれへと向けられる社会であり、上で述べたとおり、その向きは *S* が利益を得ることを保証しない。逆に、研究から利益を得る者は、その達成が *S* の認識的状況に入るのを待つとは限らない。ダリオ・アモデイ（Dario Amodè）は、AI開発企業の最高経営責任者として、うまくいった場合にAIが世界に何をもたらしかを論じたエッセイで、AIによって加速された生物学の目標を、自然に生じる感染症のほぼすべての予防と治療や、人間の寿命の倍増などの利益として述べる。そして後者について、アモデイの考える期間のうちに実現したとしても、「まだそれを知らないかもしれない」と注記する（Amodè 2024, §1 and n. 18）。

また、どちらの当事者も、誰が達成の担い手であるかを決めない。ハンフリーズの、人類の絶滅後も続く科学の事例では、人間の担い手はいない。研究を行うのが人工システムでありうるなら、研究を行う者と成果を保有する者は一致するとは限らない。そのため、達成を実際に担う者、知識でいえば知っている者は、この二つとは別に、研究が成功したかを問うところで問われる。担い手は *A* 自身でも、*S* の成員や組織でも、第三者でもありうるし、知る主体なき知識を認める説、つまり結果が知識であるかどうかを誰かの態度ではなく公的な認定と受容に求める説のもとでは、担い手がないこともありうる。そのため「*S* にとって新しい」も「*S* のための」も、*S* が知っていることを意味しない。ただし、どちらも、*S* の認識的状況が、その成員、組織、あるいは記録を通じて、求められる種類の達成を含みうることを前提にしている。そうでなければ、達成を *S* と比べることに、試みを *S* に向けることに意味がないからである（第6節）。前提されているのは、*S* がその種の達成を持ちうることであって、達成に至った経路を *S* がたどれることではない。説によっては、どの成員も導出をたどれない結果であっても、機械が生成した証明を、何をどう確かめるのかが分かっている検証器（透明な検証器）が確かめる場合のように、*S* において知られるに至りうる（第3章第5.1節）。*S* がその種の達成を持ちうるのが前提される以上、人工の *S* については、*S* が参照先であるだけの場合にも、人工システムの信念と保有の問いが生じる。

したがって、ある活動とその結果について問えることは三つに分かれる。活動が *S* のための研究であるか、*S* にとって新しい達成がどこかで実現したか、そしてそれが *S* の認識的状況に入ったか、である。最後のものには、第2章第6節の保有の説が異なる答えを与える。人工システムは、それ自体がその出力を知ることなしに、人間の知識のために研究を遂行することがありうる。信念についてのある説のもとでは、*S* のどの成員もまだ知らない結果を、システム自身が知っていることもありうる。また、*S* のために求められた結果を、まだ誰も知るに至っていないこともありうる。人間または人工の行為者が生み出した結果を、のちに、人間と人工システムの両方を構成要素とする組織（本書ではハイブリッドな組織と呼ぶ）が取り込むこともありうる。人工の共同体は、関連する条件が満たされるならば、原理的には行為者であり、参照社会であり、知識の主体でもありうる。これらの分離は分析上のものであり、どの個別的、集团的、あるいは分散的な存在者が文字どおりに知ることができるのかを予測するものではない。誰が担い手であるかは、信念の説と社会的な保有の説によって決まり（第2章第2節・第6節）、第8章第1節は場合を *S* によって分ける。

人工の行為者が人間の成員の理解を助ける場合にも、同じ区別が当てはまる。行為者、理解の担い手、参照社会は、ときに一致するとしても別のものである。集合的理解を主張するには、その集団の達成が何によって成り立つかについての説が必要になる。成員に新しい説明が利用可能になるだけでそれが成り立つかどうかは、説によって異なる（第3章第6節）。

後の章で人工研究者に何が求められるかを述べるときには、つねにこれらの区別が前提になる。たとえば、結論が証拠に基づいていることや、理由が覆されれば結論を改めることといった、システムについて述べる要求は、システム自身が知っている、あるいは理解しているとされる場合にだけ、そのままシステムにかかる。ほかの者が知る者であるなら、それが人間であれほかの人工システムであれ、システムの出力は、その者にとっての証拠や、その者が信じるに至る過程の一部であり、要求はその者の信念やコミットメントにかかる（第2章第2.1節・第4.3節）。そのため同じ出力も、誰の達成が主張されているかによって、評価のされ方が変わりうる。

この特徴づけは、それだけで研究を説明しようとするものではなく、他の理論と仕事を分け合う。探究の理論は、**試み**のうちに隠されたプロセスを照らし出すことができる。認識論は、ありうる成功と、それが知識、理解、その他の認識的達成とみなされる条件を検討する。科学的実践と技術の哲学は、物質的および表象的な実現を見えるようにする。社会認識論は、産出と保有がどのように分散しうるかを問う。誤りの制御、客観性、ガバナンスの理論は、質、批判、訂正を検討する。科学的進歩の理論は、どの達成が増えることを進歩とみなすかを問う（第2章第8.3節）。行為者性と自律性についての説明は、何がある活動を行為者自身のものとし、自己統治されたものにするのかを検討する。本書はそれらに依拠して、いつあるシステムが安定した研究能力をもつか、あるいは研究者としての地位を根拠づけるのかを問う。本書は、これらの問いをひとつの定義にまとめない。まとめれば、意図、集団の知識、自律性などをめぐって各分野で争われている問いに、研究を定義する段階で先に答えを出すことになるからである。人工システムによる研究についての主張も、どの問いに答えているかを分けて評価することになる。

3. 試みは研究活動をその成功から切り離す

実験は決定的な結論を出せないことがあり、証明の戦略は崩れることがあり、アーカイブは歴史上の問いを解決できないことがあり、プロトタイプは目標とする制約を満たせないことがある。もし研究が認識的達成の実際の実現を意味するのであれば、これらの活動は、その目標を外すたびに分類し直されることになるだろう。治療と研究の区別についての機能的分析は、関連する論点を示している。すなわち、成功しなかった活動も、意図された効果を達成できなかったときでさえ、知識の生成をその機能として保持しうる（Lewens 2006）。この説明は機能を計画と意図のうちに位置づけている。それは活動と効果の区別を支持するが、ここで用いる、より一般的な組織された方向づけの考えを与えるものではない。

成功を研究活動の条件にしないなら、本書の特徴づけのもとで活動を研究にしているのは何か。それは、活動が認識的達成をもたらす方向に向けられていることであり、本書はこれを**試み**と呼ぶ。では、その方向づけはどこにあるのか。ここでの**試み**は、すべての参加者が「私は知識を産出することを意図している」といった文を抱くことを要求しない。成功を表す語ではなくこの語を用いるのは、成功を研究活動の条件にしないためである（本節の主題）。そして、この語をこの意味で用いるのは、分散した事例や非人間的な事例を保持するためである（第1節）。活動と認識的な目的との関係は、個人の意図、集団の計画、人工物の組織された機能、構成要素の役割、あるいは規範に統べられた実践への参加によって基礎づけられうる。本書が方向づけの基礎に挙げる機能的な組織は、システムのなかで構成要素が果たす役割という意味での機能である。探究に失敗した人工システムでも、仮説を出し、実験し、証拠に依拠して仮説を改めるなら、その意味での組織をもつ。他方、人工物の機能を、その人工物が何のためのものかという意味で捉え、新しい人工物についてはそれを設計者の意図に由来するとみる説もある。その説をとると、新しいシステムの探究の方向づけは設計者に由来することになり、その活動を誰に帰属させるかが問題になる。これは活動が研究であるかどうかではなく帰属の問題であり、第7章第2節で扱う。

第1節で触れたように、ホイニンゲン＝ヒューネは、科学的な研究と、実用の目的に合わせた関連する活動とのあいだに移行領域を置く。そうした移行領域には、技術の研究と開発のあいだや、市場についての経済学の研究と企業の行う市場調査のあいだがあり、そこでは、科学の側と非科学の側とで、ほかの点での体系的の程度が変わらないことがある。たとえば、燃費を改善しようとする自動車メーカーの仕事は、ふつう、すでに市場に出ているか間もなく出るはずの特定のエンジンに合わせられている。これに対して、研究大学の工学部での同様の仕事は、ふつう、意図される応用の範囲がより広いか、特定の応用をまったく念頭に置かず、新しい設計原理を目指している。この説では、その違いは、少なくとも潜在的には、より一般的であることによる認識的な連結性の度合いの高さにある（Hoyningen-Huene 2020, 96–97）。近代科学がなぜ技術的に利用可能なかを論じた以前の論考で、ホイニンゲン＝ヒューネは、個々の行為について関連する点を述べている。実験の独立変数の特定の値を実現する点ごとの検査手続きには、すべての行為と同じく、物理的なパラメータという外的な側面と、目標を含む内的な側面がある。肯定的な結果を得た手続きは、従属変数の特定の値を得るという別の目標のもとで繰り返すことができる。そのため、検査と技術的な応用とは、外的な側面では一致し、目標においてだけ異なりうる。新しい薬の臨床試験の成功は、彼が挙げる例の一つである（Hoyningen-Huene 1989, 10–11）。ホイニンゲン＝ヒューネはまた、仮説の実験的な検査にさえ、彼が強いられた偶然と呼ぶものの要素が含まれると述べる。予測からずれた結果は、仮説のせいかもしれないが、実験の設定のなかで暗黙に当然とされていた補助仮説に誤りがあることを示すかもしれない。そのため背景の想定は、意図されてではないにせよ、実験という活動によって絶えず問いに付される（Hoyningen-Huene 2020, 98–99）。意図される応用の違いは、各参加者の心の状態の違いとしてではなく、行為が活動のなかで占める位置の違いとして読むことができる。組織された方向性という意味での**試み**に当てはめれば、この比較からは、その方向性を見定めるには、行為が活動の目指すものにどう関係するかを知る必要があると考えられる。検査と応用が外的な側面では一致することが示すように、物理的に同一の行為が、目標において異なりうるからである。さらに、強い偶然という指摘からは、行為が何に関わるかは、それが着手されたときの計画によって一度に固定されるわけではないと考えられる。ある仮説を検査するために設定された実験が、誰も意図しないまま、設定のなかで当然とされていた想定に関わるようになることがあるからである。組織された試みについては、ここから、何がある行為を未解決の事柄に関連するものににしたのか、どのようなありうる成果であればその役割が変わったのか、そして活動の方向を転換する能力がどこにあったのかという問いが生じる。第7章第2節は、活動が人工の構成やハイブリッドな構成に帰属されるところで、これらを問う。

方向づけを参加者の心の状態と同一視しないなら、三つのものを分けておく必要がある。活動の方向づけ、その活動を評価する規範、そして参加者の心の状態である。探究の目的をめぐる認識論の論争では、後の二つが分かれる様子を具体的に見てとれる。まず、第1節で見たように、探究が何を目指すのかについて見解が分かれている。ケルプ（Kelp）は、特定の問いについての探究の成功を、問いに適切に決着をつけることとする広く受け入れられた特徴づけから出発し、それを具体化する実質的な説として、そうした探究の目標を知識とする。そして、一般的な現象についての探究の目標を、ケルプが体系的な知識によって分析する理解とみなす。これに対してベダーは、探究は信念度をできるかぎり認識的価値の高いものにすることを目指すと論じ、ジェーン・フリードマンは、探究が関連する意味での構成的な目的をもつのかどうかを問い、探究の構造、規範、探究者の目的を直接に論じることが勧めている（Kelp 2021b, S1583–S1585, S1588; Beddor 2024, §§3.1, 6.1; Friedman 2024）。リー（Lee）は、この種の主張には二通りの読み方があると指摘する。探究が成功する条件についての規範的な主張としての読み方と、探究者なら誰もがもつ個人的な目的についての記述的な主張としての読み方である。そのうえでリーは、探究することは自分のもつ疑いを解こうと試みることに部分的に存する、という自らの見解が、どちらの読み方に照らしても成り立つと論じる。第一に、この見解は成功条件についてのどの説とも両立する。知識が探究を成功させるものだとは仮定すると、昨夜雨が降ったかどうかを互いに独立に探究する二人のうち、証拠を拙く評価した一人は知るに至らず、そのため成功するのはもう一人の探究だけである。それでも、どちらの探究も疑いに支えられていた。第二に、フリードマンが論じるように探究者ごとに目的が異なり、全員に共通する目的がないとしても、この見解は揺るがない。知ることを目指す人も、確実さを目指す人も、証拠によって正当化された信念を目指す人も、それぞれの目的を、疑いを解こうとすることによって追うからである（W. Lee 2026, §3.1）。つまりリーの見解は、探究の目的についての主張ではなく、個々の探究者が探究しているときにどのような心の状態にあるかについての主張である。実際この見解は、信念が探究と両立するかどうかをめぐる論争で標準的だとリーが呼ぶものに従って、探究には不思議に思うことや好奇心のような疑問の態度（interrogative attitude）が伴うとし、そうした態度なしに行われる活動を、真正の探究ではなく代用品の探究（ersatz inquiry）に数える。ただし、より緩やかな文脈ではそれが探究に数えられることも認める（W. Lee 2026, §1.1）。探究に疑問の態度が伴うという見方と、代用品の探究という区別は、フリードマンも論じている（Friedman 2013; Friedman 2019）。

本書がリーから受け取るのは、試みることと成功することが別だという点である。雨の例の二人は成否で分かれるが、どちらも探究していたことは変わらない。これは本節の活動と成功の区別を支持する。他方、リーのいう「試みる」は、疑いに支えられた個人の心的活動である。リーは人工システムを論じていないが、それに当てはめれば、その活動がリーの基準で真正の探究に数えられるかどうかは、それが疑問の態度をもつかどうかによることになる（第4章第9節）。これは、第2章第2節が認識的態度について立てるのと同じ種類の問いである。本書の**試み**は、必ずしもそうした心の状態を必要としない。集団の計画、人工物の組織された機能、構成要素の役割によっても基礎づけられうるので、すべての参加者に疑問の態度を求めない。そのため、リーの基準では真正の探究に数えられない活動も、本書の意味では試みでありうる。両者は別の問いに答えている。リーが問うのは、個人が探究しているとはどういうことかであり、本書が問うのは、活動がどのような達成に向けて組織されているかである。ケルプ、ベダー、フリードマンのどの見解が正しいかについても、本書は決着をつけない。本書の**試み**は、探究一般に構成的な目的があるかどうかについての主張ではなく、個々の活動が何に向けて組織されているか

についての記述だからである。この論争から本書が受け取るのは、活動の方向づけ、その活動を評価する規範、参加者の現に生じている心状態を分けておく理由である。

それでも、どの見方をとるかによって、ある活動をどのような試みとして記述するかは変わる。その活動が何をもちたそうとしたのか、何を達成したのかの記述が、見方ごとに異なるからである。ひとつの例で、何が変わり、何が変わらないのかを見ておく。この意見の不一致のうち二つの見方は、探究が何を指すかを定めている。ケルプが擁護する、探究は知識を目指すという見方と、探究は信念度をできるかぎり認知的価値の高いものにすることを指すというベダーの見方である。そのため、どちらの見方からも、ひとつのエピソードが何を指し、何を達成したかを、探究の指すものに照らして記述できる。そこで以下では、同じエピソードをケルプとベダーの見方からそれぞれ記述する。

取り上げるのは、数学の証明を証明支援系で検査したエピソードである。2020年、ペーター・ショルツェ（Peter Scholze）は、リキッド・テンソル実験（Liquid Tensor Experiment）を始めたブログ記事で、ダスティン・クラウゼン（Dustin Clausen）とともに証明した定理の形式的な検証を求めた。ショルツェは、基礎としてこれほど重要で、他の人々がブラックボックスとして使うであろう定理については、99.9%確かであるだけでは足りないこと、そして、ほかに誰もその証明の細部をあえて見ようとしなかったと思うので、自分にはまだ小さな疑いが残っていることを書いた。半年後、ショルツェが確信をもてずにいた論証の部分が証明支援系Leanで検査されたのち、主証明の正しさについてもう疑いは残っていないと書いた（Scholze 2020, §6; Scholze 2021）。探究は知識を目指すというケルプの見方では、その検査は、主証明が正しいことを知ろうとした探究として記述され、その成否は、それによって知識が得られたかどうかで判定される。その知識が誰にとって新しかったかは、第6節で取り上げる。探究は答えへの信念度をできるかぎり認知的価値の高いものにすることを指すというベダーの見方では、ショルツェの書いた99.9%を本人の信念度と読むなら、同じ検査は、主証明が正しいことへの信念度がすでに高いものの完全な確信には届いていない段階で続けられた探究として記述できる。このエピソードは、すでに知っているかもしれない者に決定的な証拠を得る機会が与えられるという、ベダー自身の事例と同じ形をしている（Beddor 2024, §2）。ベダーの説では、ショルツェが検査の前にすでに主証明が正しいことを知っていたとしても（第6節が定理について残している可能性と同じ種類の可能性である）、主証明が正しいかどうかについての探究の目的は、まだ達成されていなかった。真な命題への0.999という信念度は、それより高い信念度ほど正確ではないからである。そのため、検査はその目的に照らして意味をもっていた。そうであれば、検査が何を達成したかは、その信念度をどれだけ真理に近づけたかによって記述される。主証明が正しいなら、信念度を完全な確信に向けて引き上げたことで、その信念度はより正確になった。この引き上げの大きさは、確証理論のいう確率的関連性としての確証、すなわち証拠によって確率が高まることとしての確証にあたる。他方、確固さとしての確証、すなわち証拠のもとで確率が十分に高いこととしての確証の意味では、主証明が正しいという主張は検査の前にすでに確証されていた（第2章第5.2節）。もう疑いは残っていないというショルツェの言葉が、完全な確信、したがって目的の達成を表すのか、それとも目的に近づいた信念度を表すにとどまるのかは、さらに解釈を要する問題である。二つの記述のあいだでも変わらないのは、主証明のうちショルツェが確信をもてずにいた部分を検査するように、活動が組織されていたことである。本書がその検査を試みとみなす根拠は、その検査がどのような認知的達成をもたらしたかではなく、活動がこのように、証明のうち確かめるべき部分に向けて組織されていたことにある。そのため、どちらの見方のもとでもその検査は試みであるが、それぞれの見方のもとでは異なる試みとして解釈される。ケルプの見方では新しい知識をもたらそうとする試みとして、ベダーの見方では、すでに高い信念度をできるかぎり認知的価値の高いものにしようとする試みとして解釈され、何を達成したかの記述もそれに応じて変わる。

活動を成功から切り離すと、逆向きの推論もできなくなる。失敗した活動が研究でありうるのと同じように、成功に見える出力があっても、そこに研究があったとは限らない。幸運な出力が帰属を決着させることはないのである。たとえば、ランダムな生成器が新規で真なる数学的言明を出力するなら、誰もそれを認めたり裏づけたりしないうちに、真なる言明がすでに書き出されていることがありうる。別の行為者がその結果を証明し、検査し、取り込むなら、その生成器は、のちの知識の産出にとって因果的に不可欠であるかもしれない。しかし、そこから、その生成器がその結果を知っていた、研究のエピソードを遂行した、あるいは研究能力を保有していた、ということは帰結しない。偶然は、気づき、評価、保持、検証、修正がそれに組織された役割を与えるとき、確率的探索、予期せぬ観察、セレンディピティによる方向転換を通じて、研究の内部で働くことができる（Copeland 2019）。

したがって、ここでは研究のエピソードを、研究を遂行する安定した能力から区別する。確率の低い、あるいは反復不可能なひとつのエピソードも研究でありうるが、ひとつの幸運な出力は、その源を研究者と呼ぶ理由をほとんど与えない。関連する条件が変わっても信頼できること、証拠の変化に応じること、研究の諸機能が統合されていること、失敗のあとに修正することといった条件が、そのより強い帰属を支える。これらの条件は、試みが果たす最小限の仕事ではなく、第7章第3.1節の能力の帰属の分析に属する。

エピソードと能力を区別すると、ひとつのエピソードとは何か、それはどこで区切られるのかが問題になる。本書の意味での試みも、研究のエピソードという言い方も、研究が区切られた単位で進むことを前提しない。探究がどこで始まりどこで終わるか、何を一つの試みとし何を複数の試みとするかは、探究そのものによってではなく、研究を組織する慣行——いまの研究でいえば論文、助成、学位論文、公開、実験ノートの記入の番号づけ（第5章第6.3節）など——や、それを記述する者の目的によって定まりうる。たとえば、ある公開の創業プロジェクトでは、試験用の幼虫を使える季節と、ある研究者が去ってスクリーニングを実施できる者が一人になったことが、新しい化合物を発送すべき期日を定めた。同じ研究者の学位論文の締切は、共有の原稿を未完成のままその学位論文に入れるという提案を促したが、その一方で、ある貢献者はそれでもさらに化合物を作ることにした（Open Source Mycetoma 2022）。本書が研究のエピソードについて語るとき、その境界は当面の分析の目的のために引かれている。エピソードの数え方に依存する帰属は、エピソードをどう区切ったかを述べるときにはじめて評価できる。

成功のほうも、ある時点で一度に確定するとは限らない。エピソードの区切りが慣行や記述の目的によって定まりうるのと同じように、試みが成功したかどうかの判定も、後から改められうる。経験的な裏づけはとりわけそうであり、達成の側から見たこの事情は、達成が一度の閾値ではないこととして第4節で扱う。研究であつたかどうかを成功に依存させれば、その判定も、裏づけが改められるたびに改められることになる。研究を試みによって特徴づけることで、研究であつたかどうかの判定は、成功の判定のこうした改訂から切り離される。

失敗の側も、ひとつの判定に尽きるとは限らない。プロジェクトは、当初の目標を外しながらも、境界条件を確立したり、ある可能性を排除したり、実験上のアーティファクト（測定や解析上の人為的な産物）を明るみに出したり、有用な痕跡を保存したりすることがある。あるいは、回収可能な認知的な利

得を何も残さないこともある。その活動が研究であったかどうか、それが有能に遂行されたかどうか、それがその問いに答えたかどうか、それが進歩をなしたかどうか、それが副次的な知識を産出したかどうかは、互いに別個の判断である（目標が知識である場合について、第2章第8.1節がこうした判断を挙げ、同章第9.1節が、研究を遂行する人工システムに有能な遂行が何を求めるかを扱う）。本節で試みを成功から切り離れたのは、これらの判断を、研究であったかどうかというひとつの判断にまとめないためである。

4. 達成をもたらすことは、産物をつくることとも、一度の閾値を越えることとも異なる

知識や理解をもたらすことは、知られたり理解されたりする対象を創造することを要しない。天文学的な関係や過去の出来事は、探究が成功したときに存在し始めるわけではない。知識の主張では、産出という通常の語りは、第2章で検討する認識的關係の実現に関わり、その獲得や社会的保有の説明は争われている。理解については、新しい表象が、現象が何にどう依存しているかという関係（依存関係）を創造せずに、その把握を可能に示う。どちらの場合も、表象を産出することと達成を実現することは別である。証拠、能力、アクセス可能性の変化は、それ自体が成功を構成せずに、成功を可能に示う。

研究はまた、以前にはその形では存在していなかったものを構築する。たとえば、証明、形式的対象、概念、解釈、人工物、データセット、モデル、装置、分類、試料、パフォーマンス、実験的に安定化された現象などである。したがって、ここでは直接の産物と認識的達成とを区別する。

この区別をしなければ、少なくとも四つの関係が混同されうる。

1. 活動が、文章、証明、モデル、データセット、試作品、装置、解釈、パフォーマンスなどの**直接の産物**を産出する。
2. 産物が**内容**を担い、あるいは主張、パターン、可能性、区別を利用可能にする。
3. それが、証拠、根拠、説明、方法、表象、記憶、足場などの**認識的役割**を果たす。
4. それを通じて、知識や理解などの**認識的達成**が実現しうる。その成立条件と所在は説明を要し、産物や可能化する役割だけでは確立されない。

この四つを分けるのは、産物があることと達成が実現したことを区別するためである。たとえば、人工システムが大量の出力を出したことは、それだけでは、何かが達成されたことを示さない。出力の大半が偽であれば、チェックに合格した出力でさえ、第2章第5.2節で当てはめるバイズ的な概念のもとでは、支持が弱いままでありうる。この区分を使うと、人工システムによる研究についての主張に、その証拠が四つのうちのどの関係を確立しているかを問える。たとえば、生成された証明の文章は、産物があることと、ある主張が利用可能になったことを確立するが、それだけでは達成を確立しない。四つは特定の論者から借りたものではなく、産物から達成までのあいだで両者が分かれうるところを本書が分けたものである。そのため、産物と達成にかかわる関係をすべて挙げたものではない。たとえば、結果が記録され、受け止められ、保存されるといった社会との関係は、第6節で別に扱う。また、四つは関係であって、必須の時系列上の段階ではない。ひとつの証明が、定理を確立し、なぜそれが成り立つのかを説明し、ある依存関係を明らかにし、ある戦略を伝達することがある。プロトタイプは、可能性を実証し、失敗モードを明らかにし出し、設計空間を再編成することがある。また、そのメカニズムが誤解されたままで、実現可能性を確立することもある。データセットは、ある主張を支持し、モデルを較正し、予期されていなかった問いのために観察を保存することもあれば、最初の解釈を支持できないこともある。

四つの関係を分けておくことがなぜ重要かは、領域ごとの議論のなかで具体的に見てとれる。データやデザインについては、データについてのプロセス志向の説明やデザインを通じた研究（research-through-design）が、人工物がその存在だけで知識の産出を決着させることなく探究の足場となりうるのはなぜかを論じている（Leonelli 2023; Zimmerman, Forlizzi, and Evenson 2007; Gaver 2012）。たとえばレオネリ（Leonelli）は、データ、モデル、出版物を、状況に置かれた足場として扱い、それらが再利用されるたびにその妥当性と証拠としての価値を評価し直さなければならないとする。そして、それらが仕える達成を、ハソク・チャン（Hasok Chang）が能動的な知識と呼ぶもの、すなわち活動を目的に沿って遂行し調整する能力であって、情報を蓄え取り出すことには還元できないものとみなす（Leonelli 2023, 46, 52–53）。四つの関係でいえば、データという産物があること（第一の関係）は、それが証拠としてどう働くか（第三の関係）を決めず、まして達成が実現したこと（第四の関係）を決めない。その働きは再利用のたびに問い直されるからである。数学では、ラヴ（Rav）が、数学の実践における概念的な証明を形式的な導出から区別し、証明が概念のあいだにつける道筋は確信させるだけでなく、数学的な観念を照らし出す相互のつながりを打ち立てると考える（Rav 1989, 67, 73 n. 30）。ラヴの用語を採るペルク（Pelc）の読みでは、ラヴは証明を、定理が正しいことを確信させる手段としてだけでなく、数学の知識の貯蔵庫としても扱う。証明を学ぶことで、自分の研究に取り入れうる新しい観念、概念、戦略を知るのである（Pelc 2014, 2–3）。四つの関係でいえば、ここで争われているのは、数学の知識（第四の関係）が、定理の言明が利用可能になること（第二の関係）にあるのか、それとも証明にもあるのか、である。ラヴの見方では、証明は方法や概念を担い、定理の根拠として働くこと（第三の関係）以上のことをしている。サーストン（Thurston）は、概念と考え方を数学的進歩の一部として強調する（Thurston 1994）。この見方では、新しい言明が利用可能になることだけで進歩が尽くされるわけではない。工学では、方法の記述、熟練した実践者によって実行される方法、状況に埋め込まれた社会技術的な使用中の方法は、互いに交換可能ではない（Stacey et al. 2025）。方法の記述という産物があり、その内容が利用可能になっても（第一・第二の関係）、それを実行する能力が実現しているとは限らない。芸術では、スクリプナー（Scrivener）が、芸術作品はいくらかの知識を伝えうるとしても、それ自体は知識の一形式ではないと論じる。スクリプナーは、知識を伝える人工物を、知識を伝える意図によって定義し、知識がそこに蓄えられているとは含意しない（Scrivener 2002, 1, 3, 11）。四つの関係でいえば、作品が内容を担い、知識を伝える役割を果たすこと（第二・第三の関係）と、作品それ自体が知識であることは別である。説明書は知識を伝えるために作られるが、知識が紙の上に蓄えられているわけではないのと同じである。また、ある経験についての真実な記述は、その経験がそれを通じて意味をもっていた能力や感受性を保存しそこなうことがある。芸術について把握（apprehension）を中心に据えるスクリプナー自身の提案は、第8節で検討する。同節では、方向づけとその区別を、理論家自身が研究として提示している実践と比べる。

これらの議論から分かるのは、正しい記述、根拠づけ、保有、実行、伝達が互いに乖離しうるということである。そのため、人工システムが正しい定理の言明や方法の記述を出したという証拠は、それだけでは、証明による根拠づけも、システムやその言明を読む人がその知識を保有していることも、方法を

実行する能力も確立しない。四つの関係を分けておけば、そうした主張がどの関係を確立したのかを、これらの論争に決着をつけずに問うことができる。これらの文献は、証明や方法がそれ自体知識であるかどうかを決着させない。ただしラヴの説明は、ペルクの読みでは、証明を数学の知識の貯蔵庫として扱い、レオネッリの説明は、産物が仕える達成を、命題を知るのではなく、活動を遂行し調整する能力に見いだす。本書の枠組みでは、そうした能力は、命題的知識とは別の形の認識的達成の候補である。ただし、第1節の達成の水準で述べたとおり、能力は主張に応じて産物、手段、達成のいずれでもありうるので、ある能力を達成に数えるかどうかは主張ごとに特定することになる。また、知識が能力を含むかどうかは第5節が挙げる問いの一つであり、第2章の冒頭で述べるように、命題的知識を出発点とすることはその問いを決着させない。

最後に、この意味で達成をもたらすことは、論文に知識としての刻印が押されるような、ひとつの不可逆的な閾値を含意しない。第3節では、試みの成否の判定が後から改められうることを見た。達成そのものも、ひとつの時点で確定するとは限らない。部分的な証明、較正、追試、批判、理論的統合、そしてのちの訂正は、裏づけ、確信、使用可能性、そして生きている選択肢の空間を絶えず変えていく。知識についていえば、多くの説では、知識が成り立つかどうかは裏づけのあり方や、裏づけを失わせる反対の情報（無効化要因）の有無に依存するので（第2章）、裏づけが変われば、知識が成り立つかどうかとも変わりうる。デューイの説明では、反省的な探究のどの段階も、当然視してよいものを、それまでの思考の成果として、またその先の探究を方向づけるものとして示す。そうして得られた知的な内容は暫定的なものであり、それに基づいて行為するなかで正されていく（Dewey 1916, 164, 241）。理解については、さらに、程度をもつものとして扱われることがある。たとえばケルプの説では、最大の理解は、その現象についての最大限に包括的で、最大限によくつながった知識であり、それに満たない理解は程度の問題である（Kelp 2021b, S1588。第3章第7節も参照）。そうした説のもとでは、理解をもたらすことは、ひとつの閾値を越えるのではなく、理解が深まっていくことでありうる。したがって、達成をもたらすことは、あらゆるプロジェクトがひとつの離散的な閾値を越えることを含意しない。

5. 認識的達成では何を達成したかを明らかにする必要が生じる

第1節で述べたように、作業文の**認識的達成**は、知識や理解を含む開かれた総称である。どの種類の達成を目指すかを先に決めないので、特徴づけは、知識に向かう研究にも理解に向かう研究にも当てはまる。しかし、種類を決めずにおくことは、個々の主張の負担を軽くすることではない。ある活動が何かを達成した、あるいは達成しようとしていると主張するなら、それがどの種類の達成であり、その達成が成り立つ条件は何かを明らかにする必要がある。本節は、この必要が何を求めるかを、知識の場合で具体的に見る。知識については成功の条件が認識論で詳しく論じられており、多くの説がその条件を厳しいものとみなしているからである（第1節も参照）。理解についての対応する検討は第3章で行う。ここで挙げるのは知識の条件であって、あらゆる認識的達成の条件ではない。

ある活動が新しい知識をもたらそうとする試みだと主張する場合、つまり目指す達成の種類を知識と特定する場合、**知識** (knowledge) という語は、結果をほめるための敬称にとどまることはできない。その語を使えば、次のような哲学上の問いに答える責任を引き受けることになる。真なる結果と知識とを分けるものは何か。主体はそれを信じていなければならないのか。どのような種類の証拠的・信頼性的・様相的（状況が少し違っていても正しかったかにかかわる）・行為者的・社会的な関係が求められるのか。知識は能力や理解を含むのか。集団は、現時点でどの成員も抱いていないことを知りうるのか。

本章は、これらの問いに前もって答えるものではない。それは第II部で、競合する説を比べながら検討する。本章では、知識を主張するなら、どの説をとるにせよ向き合うことになる三つの論点を挙げておく。第一に、知識は世界に向けられた事実的 (factive) なものとして扱われる。イチカワとシュトイブ (Ichikawa and Steup) によれば、偽であることは知られえないという点で認識論者の大半は一致している (Ichikawa and Steup 2026, §1.1)。この見方に立てば、偽なる主張は、社会に広く受け入れられ、その受容が安定していても、知識にはならない。第二に、知識が誰のものが問われる。認識論の標準的な分析は、「X は p を知っている」という形で、知識を所有者のない命題にではなく主体に帰属させる。その主体は個人のこともあれば、社会認識論の多くの説が認めるように集団のこともある。これに対して、ポパー (Popper) の知る主体なき認識論に由来する説（第2章第6節、表2.2）では、結果が知識であるかどうかは、誰かの態度ではなく、認定と公的な受容にかかっている。第三に、運よく真であった信念や出力には、関連する非偶然的な認識的地位が欠けているように見える。たとえば、正しい答えを偶然出したシステムは、それだけではその答えを知っていたことにならない。本書は、これらの論点について、真理とは何か、誰が知る主体になりうるか、どのような関係が運を排除するかに決着をつけることを目指すものではない。

また、科学についてのすべての説明が、**知識**を、真であることを要求する意味で用いるわけではない。たとえばホイニンゲン＝ヒューネは、**科学的知識**を科学者が用いるように、関連する共同体に広く抱かれ、暫定的とも反駁済みとも扱われていない確立した信念について用い、それが真であるかどうかは問わない (Hoyningen-Huene 2018, 13)。第1節で紹介したホイニンゲン＝ヒューネのテーゼ、すなわち科学的知識は主として体系性によって日常的な知識から区別されるというテーゼは、この意味での知識についてのものである。したがって、そのテーゼは、本節で検討している種類の成功条件を与えるものではない。

上に挙げた三つの論点にどう答えるかについて、現代認識論は互いに競合する説明を与えている。説明の手がかりは、信念、証拠、アクセス可能な理由、信頼できる過程、安全性、有能さ (competence)、能力 (ability)、無効化要因 (defeater) からの自由、社会的批判、知識第一の説明など、説ごとに異なる。この競合のきっかけの一つはゲティア (Gettier) 事例である。ゲティアは、正当化された真なる信念でありながら知識とは言いがたい事例を示し、知識を正当化された真なる信念とする分析に問題を突きつけた。しかし、それに代わるただ一つの分析は確立していない。知識第一のアプローチは、そもそも知識が独立に特定された構成要素へと還元されねばならないという前提を退ける (Ichikawa and Steup 2026)。たとえばウィリアムソンは、イチカワとシュトイブの紹介によれば、知ることをそれ自体ひとつの心的状態、それも最も一般的な事実的な心的状態として扱い、知識をより基本的な認識的概念によって分析するのは誤りだとする (Ichikawa and Steup 2026, §11)。

どの説をとるかによって、人工の研究システムを構築する側にも、それを評価する側にも、異なる負担がかかる。たとえば、信頼できる過程を重視する説のもとでは、問題の信念を形成した過程が信頼できることを示すことが求められる。人間がシステムを通じてある結果を信じるようになる場合には、システムが出力を生む過程もその過程の一部である。証拠を重視する説のもとでは、知っているとされる者が、その結果を支える証拠をもっていることを示すことが求められる。第II部は、競合する説明のどれが正しいかを決めず、ひとつのチェックリストにまとめることもせずに、それぞれの説をとると人工シ

システムによる研究について何が見えてくるかを検討する。本書が立てた研究の作業的特徴づけのほうは、そのまま並べておく競合する説とは違い、それ自体が改訂の対象になりうる。中心的な研究実践を認識的達成への試みとして捉えるために、その実践自身の区別を歪めたり、望ましい変化をすべて認識的と呼んだりしなければならないなら、作業的特徴づけの内容が射程を改める必要がある。

6. S にとって新しい (new to S) は、省略されていた参照先を明示する

新しさは、参照社会の以前の認識的状况に関わる。知識については、ある人が知ることを別の人は知らず、ある専門共同体が知ることを別の共同体は知らないことがある。理解についても、共同体ですでに理解されていることを、ある個人が新たに把握しうる。どちらの変化も、一人の研究者に初めて生じたというだけでは社会的に新しくない。新しさを主張するには、どの達成が、どの点で、どの社会にとって新しいのかを特定することになる。

出発点となる観察はささやかなものである。通常の学術的実践において、確立された定理を学ぶことは、それだけでは新たな研究上の貢献とはならない。分野がすでに知っていることを独立に再発見することは、能力を示したり独立のチェックを提供したりするかもしれないが、新しさは、探究者個人がそれまで何を信じていたかだけからは判断できず、分野がすでに何を知っていたかと比べる必要がある。たとえば、リキッド・テンソル実験では、ヨハン・コメリン (Johan Commelin) がある証明を形式化するなかで構成した複体が、数か月後にマクレーン (MacLane) の構成であったと判明した一方で、コメリンがそれについて証明した補題は、本人の知らずかり新しいものだった (Commelin 2022)。逆に、関連する共同体にとって未知であった最初の結果は、当初はごく小さな下位集団しかそれを理解していないとしても、貢献となりうる。したがって、新しい知識 (new knowledge) という言い方は、誰にとって新しいのかという社会的な参照先を、口にしないまま前提している。ただし、その参照先にとって新しいことは、成員の全員が同時に信念を改めることを求めない。

作業文の変数 S は、この参照先を明示し、一般化する。それは人類、専門家共同体、組織、混合的な集団、あるいは候補となる人工社会を指しうる。少なくとも、それは、関係する認識的状况を比べることに意味がある集まりである。その比較に意味があるのは、 S の認識的状况が、その成員、組織、あるいは記録を通じて、比べられる種類の達成を含みうる場合に限られる。決して含みえないなら、その種の達成は S にとって新しいのではなく、 S の手の届かないところにあるにすぎず、試みを S に向けることもできないからである (第2節)。この記法それ自体は、法人的な心、制度的な受容、普遍的なアクセス可能性、あるいは集団的知識についての単一の理論を要求しない。表2.2が系統に分けて示す、互いに競合する説明は、異なる判定を下しうる (第2章第6節・第7節)。

省略されていた参照先をこのように明示することには、科学的実在論をめぐる論争に先例がある。ファン・フラーセンにとって、観察可能なものとは、補助のない人間の感覚で接近できるものである。それでも、観察可能という語は持ち運び可能 (portable) と同じように指標的 (indexical、「ここ」や「私たち」のように、発話の文脈によって指すものが変わる) であるとする。つまり、それは私たち、すなわち認識的共同体にとって観察可能という意味であり、私たちに数えられる存在の範囲は「必然的にも、歴史的にさえも、固定されていない」。そのため、ファン・フラーセンによれば、認識論は、たとえばイルカや地球外存在が共同体に迎え入れられる場合のように、私たちと人間とが一致しない未来を考えに入れなければならない (van Fraassen 2005, 1–4, 7)。この「私たち」が定めるのは、何が観察可能かであって、何が新しいかではない。また、ファン・フラーセンは研究を論じていない。その扱いが S と共有しているのは、日常の用法では省略されている参照先が明示され、人間を越えて広がらうものとされ、その結果、誰がそこに属するかは独立の問いになるという点である。成員を迎え入れることが混合社会である S に何をもたらすかは、第8章第1.2節で取り上げる。

この開かれ方には限界がある。事後的に寄せ集められた任意の集合が、ある出力を新しくするためだけに自動的に S となることはできない。成員、専門家の下位集団、記録、あるいは実践は、検討中の社会的な比較が意味をなす程度に結びついていることが求められる。どのような種類と程度の結びつきが求められるかは、社会認識論と、当該の達成の主張が果たすことを意図された実践的役割とに属する問題である。

また、この記法は、知識が社会的にある一時点で現れることを含意しない。結果が産出され、記録され、評定され、保存され、検索可能にされ、受け止められ、使用され、訂正され、撤回されることは、それぞれ別の出来事として起こりうる (第2章第6節で区別する諸関係を参照)。読まれていない結果は、説明のしかたによっては、知識を可能にしながらも、まだ S によって保有された知識ではない、ということがありうる。より十全な社会認識論的分析は、達成を扱う第II部で行う。

S にとって新しいかどうかは、 S がすでに何を知っていたかと比べて決まる。演繹によって得られる結果も、この比較によって新しくなりうる。数学の定理は公理から論理的に従うので、この段落の終わりで触れる理想化のもとでは、公理を知っている共同体は、その帰結である定理もすでに知っていることになる。しかし、結果が S の内部ですでに知られている前提から従っても、そこで誰かがその結果をすでに知っているとは限らない。前提から結論が論理的に従うことは命題どうしの関係であり、誰かが実際に導出を行って、知らない状態から知っている状態へ移ることとは異なる。認識論でよく論じられる原理に、知っていることから正しく演繹すれば、その結論も知るようになるというもの (閉包原理と呼ばれる) がある。一般的な形では認めない認識論者もいるが、認めたとしても、この原理が述べるのは、演繹を行えば知識が広がるということであり、まだ演繹していない帰結まで前もって主体の知識に数えることではない。また、導出が存在するというだけでは、それがどれほど長く、たどり着きたいものであっても、 S がその結論を知る位置に置かれるわけではない。知る位置にあるとは、まだ知ってはいないが、手の届く手段によって知ることができる状態のことである。第2章 (第4.4節・第7節) は、実際に知っていることと、この意味で知る位置にあることを区別する。本書はこの区別を、論理的にどう導けるかからではなく、誰が何を行い、何が手に入るかという実際の認識的状况に即して当てはめる。そのため、結論が受け入れられた公理から従う場合にも、証明は数学の共同体にとって新しい知識を産出しうる。個々の証明がそうするかどうかは、なお、その根拠と、共同体がすでに何を知っていたかに依存する。この区別に従う評価は、そうした新しい導出を、システムにとってだけの再発見から区別する。後者は、 S の内部ですでに知られていたが評価対象のシステムにだけ伏せられていた結果を、システムが導き直すことである (本節の末尾で触れる)。主体は知っていることのあらゆる論理的帰結をすでに知っているとみなす理想化、すなわち論理的全知の理想化のような想定をとれば、この区別は消える。

第3節で取り上げたショルツェの検査に戻ると、証明の検査が何を新しくしたかは、少なくとも二つの点に依存する。一つは、関連する共同体がその定理をすでに知っていたかどうかであり、第2章第6節の社会的知識の諸説はこれに異なる答えを与える。新しかったのは、その定理の知識ではなく、ある特定

の論証が正しいという知識であったのかもしれない。もう一つは、疑いが残っていても知っていると言えるかについて、どの説をとるかである。ウィリアムソン (Williamson) の知識第一の説では、あることを知っていながら、それに多くを賭ける用意がないことがありうる。また数学では、おおまかに言えば、ある命題の証明をもっていれば、その命題を知っていることになる (Williamson 2000, 86, 209, 251–252, 263–264)。ウィリアムソン自身はこの事例を論じていないが、この説を当てはめると、ショルツェの書いた99.9%は本人の信念度を述べたものと読め、ショルツェが検査の前からその定理を知っていた可能性が残る。そうであれば、その検査は、ショルツェに定理の新しい知識を与えたとは限らない。どちらの点も、その検査が試みであったかどうかではなく、何が誰にとって新しかったかにかかわる。

理解についても、 S にとって新しいかを同じように問うことができる。ただし、何と何を比べるかは、理解をどう捉えるかによって変わる。理解を知識の一種とみる説 (第3章第3節で見るカリファ (Khalifa) の説や、第4節で見たケルプの説など) では、理解の新しさも知識の新しさとして問われる。たとえば、ある現象がなぜ成り立つのか、その説明は何か、その現象についての事実どうしがどうつながっているか、といったことについての知識が、 S の中にすでにあったかどうかである。ただしケルプの説でも、現象を理解するには包括的でよくつながった知識が要るので、そうした知識の断片が S の中にあったかどうかだけでは、理解の新しさは決まらない (Kelp 2015, 16)。理解を知識とは別の達成とみる説では、それに当たる把握、たとえば依存関係や説明の把握が、 S の中にすでにあったかどうかを問うことになる。どちらの説のもとでも、定理のような結果がすでに知られていることだけでは、理解の新しさは決まらない。既知の定理に新しい証明が与えられ、それまで理解されていなかった依存関係が明らかになる場合がそうである。ただし、証明の書き方が違うだけでは、新しい理解があったことにはならない。また、理解を可能にする説明や資源が新たに利用可能になったことが、それだけで集団や科学に帰される理解にあたるのか、それとも成員に理解が実際に実現したことは分けるべきなのかも、説によって異なる (第3章第6節)。分ける説のもとでは、上で導出が存在することとその結論を実際に知っていることを分けたのと同じく、社会が理解しうることをすべてすでに理解しているとはみなさないことになる。

こうした比較は、人工システムがかかわる場合に具体的な違いを生む。人間の数学共同体ですでに理解されている証明を、人工システムが理解するようになったとしよう (人工システムが理解しうるかどうかは第3章第6節で扱う)。理解する担い手が新しいというだけでは、その数学的理解が共同体にとって新しくなるわけではない。しかし、そうしたシステムを構築することは、人工システムがもつ理解について、たとえば機械がどのような仕組みでその証明を理解するようになるのかについて、新しい知識をもたらさう。その場合に新しいのは、数学的な理解ではなく、人工システムがもつ理解についての知識である。その研究は、数学の研究とは別の達成を目指していることになる。逆に、システムが、人間の共同体の知らなかった依存関係を把握する場合には、共同体にとっての新しさがありうる。それは、その依存関係と、共同体ですでに何が知られ、理解されていたかによって決まる。それでも、共同体自体がその理解を獲得したことにはまだならない。システムの内部の状態に人間が近づけないなら、共同体の把握は変わっていないかもしれないからである。説明が共同体に利用可能になったか、成員がそれを受け取ったか、集団として理解したかについては、採る説に応じて、それぞれ異なる条件が加わる。したがって、何が把握されたか、それが S にとって新しいか、 S 自身がそれを獲得したかは、別々の主張である。 S がそのシステムを成員とする人工社会であるなら、人間の数学者がすでに理解していた証明についてのシステムの理解も、ほかのどの成員もその証明を理解していなければ、それ自体 S にとって新しくなりうる。そのとき S がその理解を獲得したかどうかは、ほかの成員がそれに近づけるか、それを共有できるかにかかり、その条件もまた説によって異なる (第3章第6節)。新しさについての主張が誤っていれば、崩れるのは貢献についての成功の主張であり、研究が試みられたことまでは否定されない。これは第3節で見た、試みと成功の区別の一例である。

こうした比較と似た構造は、創造性の哲学にも現れる。ポール (Paul) とストークス (Stokes) は、創造的なものは新しくなければならないという要件がどう精緻化されてきたかを述べるなかで、ニュートンの微積分は、ライブニッツが先に到達していたとしても創造的であり、自由は決定論と両立するという初学者の考えも、哲学者たちが何千年もそうした説を擁護してきたにもかかわらず創造的でありうることを指摘する。マーガレット・ボーデン (Margaret Boden) の用語でいえば、そうした考えは、それを生み出した者にとって新しいという意味で心理的に創造的であるが、歴史全体のなかで新しいという意味で歴史的に創造的ではない。そして心理的な新しさのほうがより基本的である。歴史全体のなかで新しいものはどれも、それを生み出した者にとって新しいが、その逆は成り立たないからである (Paul and Stokes 2023, §2)。ボーデンの区別がかかわるのは、考えがそれを抱く個人にとって新しいことであって、参照社会にとっての新しさではない。 S にとっての新しさに当てはめたとき引き継がれるのは、限られた領域に相対的に新しさを評価することがそれ自体で整合的であり、歴史全体のなかでの新しさはその源ではなくその特殊な場合だ、という点である。その領域を個人ではなく S とする理由は、 S が、認識的状況の比較が意味をなす集合体であることにある。そう読めば、ボーデンの区別からは、何が新しいかを、どこかで知られていることすべてにではなく、選ばれた S に相対的に評価することは、世界にとっての新しさから派生したものではなく、範囲の整合的な選択であると考えられる理由が得られる。そうであれば、上の事例では、ライブニッツが先に到達していたとしてもニュートンの微積分がニュートンにとって新しかったであろうのと同じように、 S の外の人工システムがすでにもっている達成も、人間の S にとって新しくありうる。これに対して、機械そのものが S の成員である場合、つまり混合の S では、機械がすでにある達成をもっているかどうか、その達成が S にとって新しいかどうかにかかわる。上で見た人工の S の場合と同じである。

S にとって新しいかどうかの判断は、 S がかつて知っていたことを失った場合にも、単純には決まらなくなる。失われた知識を取り戻すことは、 S にとって新しいのか、そうでないのか、という問いが生じるからである。ハソク・チャンは、多くのことを当然視しなければならず、あらゆる問いを追うことはできない専門科学は知識を失いうるのであり、捨てられた知識や忘れられた知識を復権させることは知識の創造の一形態であると論じる (Chang 2021, 108, 110)。チャンの例のひとつは、ピクテ (Pictet) の実験である。この実験では、熱と同じように冷気も放射され、反射されるかのように見える。チャンは、過去に報告された現象が再現できると確かめることを、忘れられた科学的知識の一片を取り戻すことと捉え、ピクテの実験については、1985年の追試そのものほとんど顧みられなかったようだと言及する (Chang 2017a, section "Recovery")。この例では、ピクテの実験の結果は、かつて科学的知識であったが、その後忘れられた。では、いまこの実験を掘り起こして結果を確かめ直すことは、それを忘れた共同体にとって新しい知識をもたらすのか。人工システムが忘れられた文献や実験を掘り起こす場合にも、同じ問いが生じる。答えは、社会がどこまで知識を保持しているとみるか、つまり保持についてどの説をとるかによる (第2章第7節)。バードの見方では、通常の経路を通じてなお利用可能な記録は共同体の知識であり続ける。そのため、記録がどのように利用可能である限り、掘り起こしても共同体の知識に新しいものは加わらない。ラッキー (Lackey) の見方を、カレストルブがそこから展開する

形でとれば、そうした記録は共同体を知る位置に置くにすぎない。そのため、掘り起こしは共同体に知識を取り戻させるものであり、その意味で新しいとみる余地がある。 S が人工あるいは混合であれば、チャンの述べる専門化が残る一方で、記録は成員の手の届くところに残り続けうる。そのとき、脇に置かれた知識は忘れられたとは限らない。バードの見方では、それに立ち戻っても何も加わらない。これに対してラッキーの見方では、どの成員も保有していなかった知識を、なお取り戻させうる。どちらの見方をとっても、この回復は、評価の対象となるシステムにだけ伏せられていた結果をシステムが再発見することとは異なる。後者では、結果は S の中でずっと知られており、新しいのはシステムにとってだけだからである。

テッド・チャン (Ted Chiang) の短編小説は、さらに別の事例を示す。記録は S の経路に入っているのに、その内容は入っていない、という事例である。この作品は、未来の学術誌の論説という形で書かれている。その未来では、「メタヒューマン」(metahuman)、すなわち生まれる前の遺伝子治療によって脳を変えられた人間が研究を支配しており、彼らは、変えられていない人間には使えないデジタル神経転送によってしか成果を公開しない。人間の読者向けの学術誌は人間の言語に訳された又聞きの説明を載せるが、その訳は「最も聡明な人間でさえ」当惑させる。人間の科学者たちは、もう独創的な寄与はできないだろうと考え、メタヒューマンの出版物の解説、メタヒューマンの人工物のリパースエンジニアリング、メタヒューマンの研究施設の遠隔観測へと向かう。そして論説は、この解釈学 (hermeneutics) を正当な方法とみなし、それが「独創的な研究と同じように」人間の知識を増やすとする (Chiang 2000, 517)。この作品はフィクションであり、そのメタヒューマンは機械ではない。それでも思考実験として読めば、この作品は三つのことを切り離している。結果は人間の S の外の担い手において実現しており、その記録は S の学術誌に載っており、それでいてどの成員もまだその記録をたどれない、ということである。この事例に当てはめると、上の保持の見方は、解説の成功がもたらす新しさを、それぞれ異なるところに見いだす。バードの見方を、一度読まれ受け入れられた結果の保持からこうした記録へ広げ (第2章第7節)、どの成員もたどれない記録も通常の経路を通じて利用可能な記録に数えるなら、その内容はすでに共同体の知識である。これに対してラッキーの見方では、その記録は共同体をせいぜい知る位置に置くにすぎない。さらに、本節で用いた「知る位置にある」の意味、すなわち手の届く手段で知るに至りうるという意味に照らせば、どの成員もそうした手段でたどれない記録は、 S をその位置にさえ置かない。後の二つの読み方では、解説は、その結果をもつ者にとっては新しくなくても、 S にとっては新しくありうる。

7. 作業的特徴づけから比較的直接的に導かれること：未解決の点へ向かうこと

作業的特徴づけからは、研究が従うべき普遍的な方法や手順は導かれない。それに比べると、比較的直接的に導かれる含意がひとつある。 S にとって新しい認識的達成への試みは、何が事実か、なぜそうなのか、対象をどう把握すべきかなど、関係する点で認識的にまだ決着していないことに向かう、ということである。作業文の「 S にとって新しい」は、目指す認識的達成を、 S の認識的状況がまだ含んでいないものとして規定する。その状況をどう捉えるかは問わない (第6節)。「試み」は、活動がその達成に向けられていることを求める。この二つを合わせると、試みは S にとってまだ決着していない点に向かうことになる。知識でいえば S においてまだ知られていないことに、理解でいえば S においてまだ理解されていないことに向かう、ということである。そして、 S の認識的状況はその種の達成を含むようになりうるのだから (第2節)、 S にとってまだ決着していない点とは、そこで決着しうる点である。達成が実現したとき、はじめはそれを A やほかの担い手だけが保有しているとしても、このことは変わらない。何かがまだ知られていない、あるいは分かっているかないということは、それについて問いが立ちうるということでもある。そこで、試みがこのように、まだ決着していない点に向かっていることを、本書は最小限の問いの構造と呼ぶ。最初から言葉で問いが定式化されている必要も、対象についてのすべての事実が未知である必要もない。対象についていくつかの事実が知られていても、理解は未達でありうる。また、試みが向かう点が、実際に S においてまだ決着していない必要もない。ここで導かれるのは、試みがどこに向かっているかであって、 S の実際の状態ではないからである。目指した達成がすでに S において成り立っていた場合、第6節で述べたように、誤っていたのは新しさの主張であり、研究が試みられたことではない。どの活動が試みに属するかは、目指す達成との組織されたつながりによる。

これは、あらゆるプロジェクトが一つの明示的な疑問文から始まるということも、一人の行為者がそれを抱いているということも、最終的な問いが当初から変わらずに存在していたということも意味しない。探究のなかで問いがどのように形をとり、立て直されていくかは、第4章第4節が、研究の記録と探究の説に即して見る。ここでは、最小限の問いの構造がどの意味での問いなのかを確かめておく。

最小限の問いの構造という主張がかかわるのは、試みが向かう先にある認識上の未決着であって、一つの心のなかに表象された固定した問いではない。問いという語そのものが、疑問文にも、その内容にも、それを問う行為にも用いられる (Ciardelli 2022, 16 n. 3; Habgood-Coote 2026, 2)。問う行為の説明は、問う者に何を求めるかで異なる。ワトソン (Watson) は三つめの用法をとり、問いを行為として、情報を引き出すというその特徴的な機能によって同定する。そのため、問いは疑問文の形をとる必要がなく、相手を困らせるといった別の明示的な目的で問われた問いも、真正の問いでありうる (Watson 2015, 86–102)。ハブグッド＝クート (Habgood-Coote) は、問うとは疑問の態度を表し、答えを得ることを目指し、聞き手に答えを与えるよう指示することだ、という見方を標準的な見方と呼び、それが三種類の問いかけに合わないと論じる。答えを知っている問いを教師が生徒に問う場合、フレデリック・ダグラス (Frederick Douglass) が1852年の演説で行ったように自ら答えるために問う場合、そしてダフィット・ヒルベルト (David Hilbert) が1900年の国際数学者会議で、数学の共同体に解かれることを望む問題を示したような共有された問いかけ (shared asking) である (Habgood-Coote 2026, 2, 4–6)。二人の説で問う行為の同定に何が要り、何が要らないか、ハブグッド＝クートがそこからどのような見方を提案するかは、第4章第1節と第6節で見る。

本節のいう問いの構造は、まだ疑問文、その内容、問う行為のどの意味での問いでもないことがありうる。ワトソンの説は、問いを疑問文の形や問う者の意識的な動機から切り離す点で、この構造に近い。ただし、ワトソンが同定するのは、情報を引き出す機能をもつ行為である。本節のいう問いの構造は、行為ではなく、試みが、 S においてまだ決着していないものとして向かっている点である。疑問文はのちにそれを表現しうる。その内容は探究のなかではじめて確定しうる (第4章第3節・第4節)。一人または複数の参加者による問う行為が、それを追求しうる。したがって、この主張は、そのような行為を、 A のなかに、あるいは S のいずれか一人の成員のなかに位置づけることに依存しない。

また、最小限の問いの構造という主張は、問いがどこから来るのかも定めない。問いは、 A の探究のなかで生じることもあれば、研究を指導したり資金を出したりする者、共同体、装置、分野の歴史などから受け継がれることもある (第4章第1節)。行為者がその問いを生み出したかどうかは、研究が試みられているかどうかではなく、自律性にかかわる (第7章第1節・第4.3節)。そのため、この構造だけでは、機械にどの意味で問うことを帰属できるのか、

問いの中身が当初から定まっているのか、誰がそれを担うのか、誰にとっての未知なのかは決まらない。第III部の三つの章は、三つの側から探究の過程を見る。探究とそれを組織している未決着の点との関係、答えの候補・表象・手段のあいだの関係、そして結果とそれがかわるものとの関係である。そのうち第4章が、機械が問うとはどういうことかを、問いの意味、中身が当初から定まっているか、担い手、誰にとっての未知かという四つの点に沿って問う。

8. 帰結しないこと

作業的特徴づけは、あらゆる研究活動が次のようであることを立証するものではない。

- 完全に定式化された仮説から始まる。
- 確定した解答空間を表象する。
- 発見を正当化から切り離す（仮説を思いつく過程と、それを受け入れる根拠を示す過程とを、別々の段階として扱う）。
- 検証と呼ばれる単一の手続きを実行する。
- 説明的理解を産出する。
- 論文を公刊する、あるいはその結果を直ちに利用可能にする。
- 一人の行為者によって統御される。
- 価値がある、重要である、あるいは進歩的である。
- あるいは、最大限の自律性をもつ研究者を含む。

また、この特徴づけは、知識が、あるいは認識的達成一般が、研究の唯一の目的であることも、認識的な目的が実践的・道徳的・政治的・美的・変革的な目的につねに優先することも、認識的な方向づけが研究をあらゆる隣接活動から一意に区別することも証明しない。この定式は目標の構想を一つ特定するものであり、本書はその構想が構築にとつても帰結を検討する。

この特徴づけはまた、研究を何のために行うかという目的の選択と、どの問いをどの順で取り上げるかというアジェンダの選択を、実質的には決めずにおく。社会は、無際限に多くの事実、説明、構成物、可能性について無知である。新しさは、そのうちどれが注意に値するかを決定しない。問いの選択は、たとえば、説明への関心、期待される認識的利得、実践的必要、道徳的懸念、実現可能性、制度的優先順位、機会費用、リスクと便益の分配などを反映しうる。キッチャー（Kitcher）の説では、ある問いが重要かどうかは、人々の特徴、人々の目的と価値、そして社会の組織のあり方に依存し、人間の好奇心を表す問いに答えることは、重要性の捉え方の一つである（Kitcher 2013, 216-217）。これに対して、第9章第3節で示すように、知識の価値が S から独立だとする立場もある。そこでは、その価値は S の目的と関心に相対的ではないか、それだけに相対的なのではない。この立場には二つの形があり、 S の未知のあいだの選択への帰結が異なる。一つは、グリム（Grimm）がクヴァンヴィグ（Kvanvig）とリンチ（Lynch）に帰する原理を採る形である。そこでは、あらゆる真なる信念が一応の（prima facie）価値をもつ。しかしそれだけでは、それらの価値を等しくすることも、 S の未知に順序をつけることもないので、選択は決まらないままである。もう一つは、キッチャーが認識的重要性への伝統的なアプローチと呼ぶ形である。そこでは、世界の構造がどの真理が重要かを定めており、 S の未知のあいだで選ぶことは、何が重要かを見いだすことに帰着するだろう。どちらの形でも、選択を決めるのは S にとっての新しさではない。前者は順序を与えず、後者は新しさとは別の根拠、すなわち世界の構造によって順序を与える。本書は重要性を、知識や研究としての地位の隠れた条件として扱うのではなく、第1節で区別した四つ目の水準、すなわち価値とガバナンスの水準に置く。

方向づけとその区別を多様な研究実践と比べる

本節で述べたように、作業文は、構築上の帰結を検討する目標構想を定めるものである。第5節は、それがどのような場合に改訂を要するかも述べた。中心的研究実践を認識的達成への試みとして扱うことが、その実践自身の区別を歪めたり、望ましい変化をすべて認識的と呼んだりすることを要する場合である。本書は作業的特徴づけを試すことを目的としない。それでも、研究として提示されながらこの構想に取まりにくい実践と特徴づけを比べ、どこまでを覆い、どこで別の捉え方と分かれるかを見ておけば、改訂の条件がどこで問題になりうるかを具体的に知ることができる。それはまた、この構想を採って人工システムを構築し評価する場合に、目標に何が入り、何が入らないかを、定義の上だけでなく事例に即して示す。

ここでは、そうした実践を三つ取り上げる。芸術的研究、参加型アクション・リサーチ、そしてショーン・ウィルソン（Shawn Wilson）が描く先住民の研究パラダイムにおける研究である。いずれも、その理論家自身が研究として提示し、しかもその記述が本書のコミットメントに圧力かける実践である。この二つを条件に選んだのは、特徴づけの語であらかじめふい落とされていない実践と比べるためである。特徴づけに合うことを理由に選べば、どこで別の捉え方と分かれるかは見えにくくなる。

三つの実践は、それぞれ特徴づけの異なるコミットメントにかかわる。比べるときに立てるのは、次の三つの問いである。第一に、その実践を論じる理論家自身の説明において、実践は、関係する社会にとって新しい認識的達成への組織された方向づけを含むか。作業文の「試み」と「 S にとって新しい」が求めるのはこれである。第二に、何がその達成を構成し、何が単なる産物や可能化条件にとどまるのか（第4節）。第三に、本書が提案するように、達成を価値や関係から分析上区別できるか。これらは本書の問いであり、検討する実践を拘束する中立的な語ではない。

芸術的研究の事例は、主に第一と第二の問いにかかわる。作業文の認識的達成は知識だけでなく理解も含むので、芸術的研究の制度上の定義が知識と並べる理解は、はじめから範囲の外に置かれることはない。この事例では、そのことと、それでも残る問いとを区別できる。残る問いとは、ボルドルフ（Borgdorff）のいう、まだ知らないことへの開かれを、達成への方向づけが捉えられるか（第一の問い）、そしてスクリブナー（Scrivener）のいう把握（apprehension）の対象が、第4節の意味での産物にとどまるのか（第二の問い）である。アクション・リサーチは第三の問いにかかわり、認識的達成を価値ある行為から分けられるかを問う。ウィルソンの関係的な説明は、第三の問いのうち達成を関係から分けられるかと、達成が誰に、あるいは何に帰属するかの両方にかかわる。これらの事例は圧力を網羅せず、合意をなすものでもないが、合わせると、上の三つの問いのすべてにかかわり、さらに、関係

的な説明が本書とは異なるところに置く帰属の所在にもかかわる。理解の説明は第3章で、産物の役割は第4節と第5章で詳しく検討する。ここでの問題は実践の特徴づけである。

芸術的研究をめぐる議論が応答してきた制度上の定義は、知識と理解を対にしている (Borgdorff 2011, 54–55, 63 n. 20; Scrivener 2002, 2)。ボルフドルフによれば、芸術的研究において創造の過程は、新しい洞察、理解、産物が生まれる道筋、あるいはその一部である (Borgdorff 2011, 46)。そのうえでボルフドルフは、芸術的研究が学術研究として認められるための基準を示す。そのなかには、研究は私たちが知り理解していることに寄与し、既存の体系に新しい知識や理解を加えようとする、という基準が含まれる (Borgdorff 2011, 54–55)。論考の冒頭で予告されているとおり、ボルフドルフはまた芸術的研究を、美的経験のうちに閉じ込められ、創造的実践のうちで演じられ、芸術作品のうちに具現化された内容、すなわち反省にも概念にも先立つ内容を、形にすること (前反省的で非概念的な内容の分節化、articulation) として特徴づける。そして、それはそれらの語の通常の意味での理論構築や知識産出を果のところで含まないのであり、「むしろ、知らないこと、あるいはまだ知らないことへ向けられている」と結論する (Borgdorff 2011, 44, 47, 59, 61)。スクリプナーは、芸術作品はいくらかの知識を伝えるが、それ自体は知識の一形式ではなく、芸術が与えるのは結論ではなく可能性と見方であると論じる (Scrivener 2002, 1, 7, 11–12)。そこでスクリプナーは、芸術における研究を「新しい把握 (apprehension) を生み出すために企てられる独創的な創造」と定義することを提案する。ここでの新しさは文化的な新しさであり、単に作り手や個々の見る者にとっての新しさではない。またスクリプナーは、芸術の共同体が何を研究と呼ぶかを、その共同体自身の決定として扱う (Scrivener 2002, 2, 12–13)。

ボルフドルフは、この論争の多くの参加者が、知識を、正当化された真なる信念や根拠づけられた主張可能性というより、世界の開示や世界の構成とみなすと報告する (Borgdorff 2011, 56, 61頁も参照)。作業文の認識的達成が理解を含むからといって、ボルフドルフの説との一致が確立するわけではない。ボルフドルフは、この論争のなかに、理解を、理論的知識・実践的知識・面識による知 (対象と直接に知り合っていることによる知) が交わりうる知識の一形式とする見方があると報告する。芸術的研究については洞察と了解 (insight, comprehension) を加え、説明的な把握 (explanatory grip) を得ることよりも、知覚的・受容的な関与や、意味を内側からくみ取ろうとする了解的 (verstehende) な関与のほうが研究にとって重要であることが多いと強調する (Borgdorff 2011, 55)。第3章で比較する説がその理解を捉えるかは、実質的な問いとして残る。また、理解が含まれるからといって、ボルフドルフが述べる洞察、世界の開示、表現のすべてが、認識上の成功についての説を満たすと確立されるわけでもない。ボルフドルフの説明からはなお、達成への方向づけが、まだ知らないことへの開かれを十分に捉えるかという問いが残る。第7節は作業文から、Sにとってまだ決着していない点へ向かうことを導いたが、それはそこで達成をもたらそうとする試みの一部としてであった。これに対してボルフドルフは、知らないこと、あるいはまだ知らないことへ向けられていることを、通常の意味での知識産出と対比している。そのため、第7節と照らし合わせると、問われるのは、その開かれが、まだ知らないことに決着をつける方向をもつかどうかである。ボルフドルフが学術研究の条件に含めるのは、学術的な場に対して研究の重要性を正当化し、その意義が当の芸術家の作品群を超えることである (Borgdorff 2011, 54, 57)。本書はこれらを、研究としての地位ではなく意義の問題に置く。

スクリプナーは、芸術制作を、知識と理解を得るための調査と明示的に対比している。スクリプナーのいう把握は、感覚と知性によって把握される対象であり、物事がどうであったか・どうであるかの知識ではなく、見る仕方とあり方を提供する (Scrivener 2002, 11–12)。したがって、作業文が理解を含むことだけでは、スクリプナーの提案に応えたことにならない。そのような対象は作品として作り出されるので、第4節の区別では、活動が産出する直接の産物 (第一の関係) にあたる。それを把握することが知識や理解などの認識的達成をもたらすか、どの条件のもとでそうなるかは、さらに論証が要る。文化に相対的な新しさはスクリプナーの提案に関わるが、この問題を決着させない。芸術共同体が何を研究と呼んでよいかを決める権限は、本書の語にはない (序論第4節)。スクリプナーの提案と本書の特徴づけが適合するかは、ここでは決めない。スクリプナーの議論には、知ることは確信を伴って経験される、新しい知識の主張は論証によって正当化されなければならない、といった、本書が採らない前提も含まれる (Scrivener 2002, 4–5, 10, 12)。そのため、芸術作品はそれ自体として知識の一形式ではないという結論は、スクリプナーのいう知識についての主張であって、本書の意味での知識についての主張ではない。本書がこの事例から引き出すのは、第二の問い、すなわち何がその達成を構成し、何が単なる産物にとどまるのかという問いへの答えではなく、そうした対象を把握することが知識や理解などの認識的達成をもたらすと示すにはそれ自体の論証が要る、という点である。

リーズン (Reason) とブラッドベリー (Bradbury) は、アクション・リサーチの主な目的のひとつとして、人々の日々の生活の営みに役立つ実践的な知識を産出することを挙げ、それが「新しい実践的な知識」と、知識を創り出す新しい能力とに通じると書く (Reason and Bradbury 2001, 2)。したがって、作業文に照らせば、少なくとも実践的な知識のうち命題として述べる部分については、その方向づけは二人自身の説明のうちに見いだされる。それと並んで、解放や癒やしのように、第1節がそれと両立しようとした目的もある (Reason and Bradbury 2001, 2, 10)。二人は、ヘロン (Heron) とリーズンに従って、相互に依存する四つの知のあり方を挙げる。人や場所や物との直接の出会いによる経験的な知、その出会いを物語や絵や動きによって初めて表現する提示的な知、概念や観念に拠る命題的な知、そしてやり方を知ること (knowing how) としての実践的な知である。このうち第一とされるのが実践的な知である。それは、ヘロンに従えば、他の知のあり方を行為において完成させ、経験的な知を「価値ある行為において」完成させる。アクション・リサーチにおいて知識は「名詞ではなく動詞」である、と二人は書く (Reason and Bradbury 2001, 1–2, 9, 11)。リーズンとブラッドベリーの意味での実践的な知は、三つの層に分けて捉えることができる。命題として述べる主張は、第5節の意味での知識として評価できる。技能については、本書はそれが知識の一種であるかどうかを決めず、能力として評価する (第2章第1節)。行為の価値については、本書は価値の水準で評価する。リーズンとブラッドベリーは参加について認識上の理由も挙げる。外部の研究者の先入観を通さずに、当事者の視点と関心に根ざした成果は「よりよい」研究になる、という理由である。本書は根拠づけの説を一つに決めていないので、批判に持ち込まれる視点を重視する説 (第2章第6節) のもとでは、この理由を扱うことができる。ただし二人は、アクション・リサーチの特徴は単なる方法論の問題ではなく、「それをはるかに超えて」、知る者と知られるものを参加的な関係のなかで結びつけることによって、経験主義的・実証主義的な世界観の基礎を掘り崩すと付け加える (Reason and Bradbury 2001, 4)。二人は妥当性と質の問いのなかに、探究の意義と、探究が営まれた関係を含める (Reason and Bradbury 2001, 10, 12)。本書はこれらを、自らの分析上の配置として価値とガバナンスの水準に置く。リーズンとブラッドベリーはこの分離を受け入れない。

リーズンとブラッドベリーは、アクション・リサーチを、新しい形の理解を創り出すことでもあると述べる (Reason and Bradbury 2001, 2)。それが第3章で検討する意味での理解であるかは、問いとして残る。作業文の認識的達成は理解も含むので、認識的な方向づけの評価は、上で特定した命題的な成分

だけに依存する必要はない。しかし、それによって本書が、技能を知識や理解と同一視したり、その成功を価値ある行為から切り離したりする実践的な知の説明を与えたわけではない。認識的達成という総称は、この実質的な問題に答えを出さず、参加と価値についての不一致も解決しない。

ショーン・ウィルソンの描く先住民の研究パラダイムでは、関係が実在であり、関係への説明責任 (relational accountability) がその価値論と方法論の基礎をなす (Wilson 2004, vi, 147–148)。ウィルソンは、そうした研究とそこから生じるものを導かなければならない原則を挙げる。そのなかには、研究の問いの基盤が先住民の経験の現実のうちにあること、方法が過程を重視するものであること、研究者が、研究が共同体にもたらす変容と成果に責任を負うことが含まれる (Wilson 2007, 194–195)。スタン・ウィルソン (Stan Wilson) に拠りつつ、ショーン・ウィルソンは、知識は関係的であり、個々の研究者に所有されるのではなく、すべての被造物と分かちもたれるとし、自分が示す知識は、自分のものでも、研究に参加した仲間たちの集まりのものでもない、と書く (Wilson 2003, 172; Wilson 2004, 223)。関係から得られる知識について語るが、知識は所有されることも発見されることもないとする。研究がしているのは新しい関係の集まりをつくることであり、新しい考えを存在させることには、それを取り巻く関係への責任が伴う、と書く (Wilson 2004, 152, 209, 232)。関係への説明責任のもとでは、「正しいか誤りか、妥当性、統計的有意性、価値があるかないか、といった価値判断は意味を失う」とウィルソンは書き、妥当性と信頼性の基準を満たし、あるいは「事実」を正確に記述している研究でも、関係から切り離されていれば先住民のパラダイムのなかでは受け入れられないとする (Wilson 2004, 147–148, 188)。また、そうした研究には、外から課された検査なしに、言うことが「真」であることを確かめる独自の仕方があること、継続的なフィードバックによって参加者が分析の正確さを確かめられること、そして正確さは、現象を記述することよりも、現象をなす関係の全体を記述することにおいて重要であることも書く (Wilson 2004, 188, 222–223)。ウィルソンの説明では、研究は儀式であり、そのようなものとして尊重されなければならない。それは、すべてのものが関係しつながつていることを敬意をもって確かめ直す「人生を変える儀式」である (Wilson 2003, 175)。

ウィルソンは、このパラダイムを他の研究パラダイムと比べることによって述べようとはしない。比較は二項対立をつくり出し、どうしても評価と従属に行き着くからである (Wilson 2007, 194)。このパラダイムのもとでの研究が本書の意味での知識を生むかどうかは、ここでは判定しない。本書の語で下した判定は、いずれにせよ、その実践を分類される共同体にとって権威をもたない (序論第4節)。記録するのは、本書の語とウィルソンの語が離れる三つの箇所だけである。第一に、関係が知識を担いうるか。本書は知識を主体に帰属させる標準的な分析から出発するが、知る主体なき知識の説にも触れており、どの存在者が知識の主体でありうるかは決めていない (第2節、第5節)。また、知識は所有されえないというウィルソンの主張は、それだけでは、知識が誰かに帰属されることの否定ではない。第二に、関係についての考慮をどこに置くか。本書はそれを研究としての地位の条件にしないが、ウィルソンはそれを先住民の研究として受け入れられることの条件にする (Wilson 2004, 147–148, 188)。第三に、真理、正確さ、新しい考えについてのウィルソンの語りが、第5節の事実的な知識とどう関係するか。

三つの問いに照らすと、結果は次のとおりである。第一の問い、すなわち実践が関係する社会にとって新しい認識的達成への組織された方向づけを含むかについては、リーズンとブラッドベリーの説明は、少なくとも命題として述べる実践的な知について、作業文の方向づけを含む。ボルフドルフの基準も、芸術的研究が私たちの知り理解していることに寄与しようとすることを求める点で、それを含む。ただし、ボルフドルフが同時に述べる、まだ知らないことへの開かれをその方向づけが捉えるかは残る。スクリプナーの提案と合致するかは未決のまま残り、ウィルソンのパラダイムのもとでの研究について本書は判定しない。第二の問い、すなわち何がその達成を構成するかについては、スクリプナーの把握の対象は第4節の意味では産物であり、それを把握することが知識や理解などの認識的達成をもたらすかには、さらに論証が要る。第三の問い、すなわち達成を価値や関係から分析上区別できるかについては、本書は、達成を価値、ガバナンス、関係から分けることを自らの分析上の提案として保つ。リーズンとブラッドベリーはそれを争い、ウィルソンの説明では達成は関係から切り離されない。本書はこの違いに決着をつけない。つまり、作業的特徴づけは、三つの実践を、その意味での研究として一律に包摂するのでも、そこから一律に締め出すのでもない。論者自身の説明のなかに、作業文の方向づけが見いだせるところと、見いだせるかが決まらないところがある。

その境目がそこに引かれるのは、作業文が達成を認識的達成という語で名指しているからである。この語は理解を含むので、ボルフドルフの基準が知識と並べる理解は、はじめから作業文の範囲の外に置かれることはない。一方でこの語は、第1節で述べたように成功語であり、知識や理解を主張するにはそれぞれの成立条件を、ほかの達成を主張するにはそれぞれの論証を求める。そのため、論者が知識や理解への寄与を目指すところでは方向づけが見いだせるが、まだ知らないことへの開かれ、スクリプナーの把握、実践的な知のうち命題にならない部分、関係的な洞察のように、実践が目指すものが、実現したときにその条件を満たすのか、あるいはそれを支える論証があるのかを示されていないところでは、決まらないまま残る。本書の語に言い換えたからといって、これらを取り込めたことにはならない。言い換えによって取り込んだと主張すれば、実践自身の区別を歪めることになり、第5節の改訂条件に当たるだろう。決まらないものを決まらないまま残すかぎり、この改訂条件には当たらない。したがって、三つの実践との比較は次のようにまとめられる。この方向づけが、研究と呼ぶのが適切なあらゆる実践に必要な第1節で決めずにおいたので、特徴づけを改訂する必要はいまのところ生じていない。ただし、特徴づけが覆うのはそこまでであり、その先では別の捉え方と分かれる。

人工研究者を構築し評価するうえで、この検討の帰結は二つある。第一に、この構想に沿って人工システムを構築し評価するなら、上で決まらないまま残ったものは、それだけではその目標に入らない。そうしたシステムが研究をすると主張するなら、どの意味での研究かを述べる必要がある。決まらないまま残ったものは、異なる前提にも依っている。美的経験のうちに閉じ込められた内容というボルフドルフの考えや、リーズンとブラッドベリーがヘロンに従って挙げる経験的な知は、経験をもつ知り手を前提している。これらの論者はいずれも人工システムを論じていないが、人工システムに当てはめれば、それらに対応するものがあるかどうかは、そのシステムが何にどう出会うかによる。参加についてのリーズンとブラッドベリーの理由や、ウィルソンの関係への説明責任は、誰がどのような関係のなかで探究に加わるかにかかわる。システムに当てはめれば、それらは、誰の視点とどの関係がその研究に入るのか、そして *S* が人間であれば、その人間の成員がどう加わるのかを問うことになる。評価の制度がこの構想を用いれば、構想が目標に入れないものは、制度のなかで見落とされうる。序論第4節で述べた制度的利用の帰結は、ここにも当てはまる。第二に、この構想の内側でも、研究を認識的達成をもたらそうとする試みとして記述するだけでは、人工システムの目標は定まらない。目標とその条件はなお特定する必要がある。それらを特定しなければ、認識的達成を語る評価でも、理解が抜け落ちたり、可能化する人工物が意図した達成として数えられたりしうる。

9. 特徴づけが残す負債

本書は作業文を、主に、それがはっきりさせる問いのために用いる。作業文は、自らは答えない問いをいくつも生む。たとえば第5節で述べたように、知識を主張するなら、真なる結果と知識を分けるものは何かを説明しなければならない。本書はこうした問いを負債と呼ぶ。作業文を採った以上、後の章が答える責任を負う問いだからである。

本書は負債を四つにまとめる。達成の負債、過程の負債、帰属の負債、研究体制の負債であり、それぞれ第II部から第V部までの一つの部が扱う（序論第6節）。四つを分けておくのは、一つに答えても別の一つは決まらないからである。たとえば、ある人工システムが正しい結果を出したと示しても、それだけでは帰属の負債は返されない。その研究の遂行や研究能力をそのシステムに帰属できるかは、別に示さなければならないからである。この整理を使えば、人工システムによる研究についての主張が、どの負債をどこまで返しているかを問うことができる。

四つの負債は、本書の各部が扱う問いをまとめるものであり、作業文の語から一つずつ導いたものではない。また、作業文の語とも、第1節の四つの水準とも、一対一には対応しない。各負債は作業文の複数の語にかかわるので、第2節が各語を第3～6節に対応させたような対応はない。水準との関係でいえば、達成と過程の負債は同名の水準にかかわり、帰属の負債は、活動・能力・地位をどこに置くかを問うので、水準を横断する。研究体制の負債がかかわる関係は、下で見ると、主に価値とガバナンスの水準で評価される。第1節がその水準に置いた問いは独自の負債をもたず、それらの関係にかかわるところで第V部が扱う（第9章第1節）。

第一は**達成の負債**である。意図する認識上の成功は、新しい出力、根拠ある受容、有用な人工物、可能化する資源とどう異なるのか。第2章と第3章は、知識と理解の成立条件、ありうる担い手、説明や能力との関係を検討する。産出、獲得、保有、利用可能性、使用、保存は一致する必要がない。一般的な語ではこれらの相違は解決されず、別の達成を主張すれば別の説明の負担を負う。

第二は**過程の負債**である。答えが関連する社会にとって利用可能でなく、プロジェクトが失敗しうろうえ、その問い、概念、証拠、方法、基準そのものが変化しうるとき、探究はどのように進むのか。候補形成、実験、証明、モデリング、解釈、構築、批判、追試、変則（anomaly）の診断、表象は、可能な役割であって、固定した順序ではない。

第三は**帰属の負債**（attribution debt）である。研究は、何に、そしてどの明示された境界において帰属されうするのか。一つの正しい結果を産出することは、安定した研究能力を立証しない。この問いに答えるには、人工の行為者に帰属されているのが、貢献なのか、活動の遂行なのか、能力なのか、研究者としての地位なのかを問う必要がある。また、探究のどの部分をその行為者が選択し、実行し、監視し、修正し、統御するのか、どの部分が外部から供給されたままであるのか、そしてどの点で、どこまで自律的であると主張されているのかを問う必要がある。研究の遂行を帰属させても、意図する達成がどこで実現するのかは決まらない。行為者は、求められる達成の担い手である必要はなく、その達成がそれに相対的に新しいとされる社会に属する必要もない。

第四は**研究体制の負債**である。ここでいう研究体制とは、研究を取り巻く社会的な秩序であり、研究の分業、探究の配分、批判と認定、記憶と訂正、功績、権威からなる。研究はすでに、認知的な作業を、共同研究者、道具、アーカイブ、査読の実践、組織、公衆にわたって分散させ、責任を、そこにかかわる人々と制度のあいだで分けている。人工研究者は、アジェンダ設定、注意、配分、妥当性確認、記憶、訂正、功績、権威を変えるかもしれない。このメタ科学的問題は作業的特徴づけとは独立の起源をもつが、人工の行為者が研究世界に入るとき、両者は出会う。特徴づけの三つの語、*S* にとって新しい、*認識的達成*、そして *S* そのものは、活動が *S* に結びつくところを示し、特徴づけの外の条件と合わさって、そうした行為者が変えうる関係を生むからである（第9章第2.1節）。*S* の未知は無数にあるのに探究の手段は限られ、多くの研究者がそれを追うので、*S* の未知のあいだでの探究の配分が生じる。*S* の成員はそれぞれ多くの結果について自ら証拠を集められず、他者を通して結果に依拠するので、結果の批判と認定が生じる。そして *S* は多くの成員からなつて一つの声をもたず、探究の帰結は *S* の外にも及ぶるので、配分と認定を誰が決め、誰が改めてよいかという権威が生じる。特徴づけは、これらの関係のいずれも研究としての地位の条件にしない。ただし、社会的な組織が根拠づけや保有に入るとする説では、認定が、何が *S* の知識と数えられるかにかかわる。集団の知識を、該当する領域で意思決定の権限をもつ成員に依存させる説では、権威の一形態もそこにかかわる（第2章第6節）。公共的・集合的な理解についての説でも、社会的な組織が主張される達成に関わりうる（第3章第6節）。関係が達成の条件に入らない場合、その評価は価値とガバナンスの水準にとどまる。

これらの負債は互いを制約する。達成の説明は実践を誤記述しうるし、実践の記述は根拠づけを確立せず、帰属の境界は分散した仕事を隠しうる。第2章と第3章は、知識と理解が主張されるところでその主張を評価し、第4～6章は問いの形成と立て直し、候補と表象が利用可能になること、失敗の局所化、第7～8章は遂行と能力の所在と、達成が誰のものになるかを、第V部は探究の配分、結果の批判と認定、両者に対する権威を検討する。目標構想を変えれば、それらの問いは消えないまま、答えが変わりうる。冒頭の方角づけを退ける読者も、どの結論がそれに依存するかを特定し、残る論証を評価できる。

第II部 何が認識的達成とみなされるのか

第2章 結果はいつ知識となるのか

認識的变化、達成、知識

1. 新しい結果から認識的達成へ

作業的特徴づけは、研究を、社会 S にとって新しい認識的達成をもたらそうとする試みとする。本章は、要求度の高い成功の一つである知識を検討する。根拠のない予想、未読の記載、偶然の真理、意義が未確定な人工物にも新しさはありうる。しかし、これらはどれも、新しい、あるいは有用だというだけで知識になるわけではない。一般的な方向づけは、この個別の達成の成立条件を緩めない。

知識という語は、少なくとも三つの問いを視野にとどめておく。結果はどの点においてその対象を正しく捉えているのか。結果として生じた認識的状态は、どの主体に帰属されるのか。対象への認識的に成功した結びつきを、幸運な偶然から区別するものは何か。これらは知識という語によって提起される負担であって、完全な分析の三つの独立に必要な構成要素ではない。認識論の説によって、これらの問いの整理の仕方も、それへの答えも異なる。

本章は、次の順で進む。第2節は、どのような種類の認識的態度が問題なのか、主として信念について、またその担い手が結果の産出者である必要があるかを問う。第3節は真理と可謬性を、第4節は真なる成功を偶然的ではなく認識的なものにするのは何かを扱う。第5節は検証が何に寄与するかを問う。第6節は、知識について社会的でありうるもの、すなわち新しさの参照先、産出、保有、保持を分ける。第7節は、結果の産出と S による保有とが離れる事例で、それまでの節が述べた信念、検証、社会的な保有についての説をまとめて試す。第8節は知識の手前と周辺における認識的变化に目を向け、第9節は、それぞれの説が人工システムによる研究に何を求め、各要求が誰に掛かるかを述べる。

本章は命題的知識から始める。命題的知識の分析は詳しく発達しており、分析の入口として使いやすいからである。ただし、この入口は作業的特徴づけが求めるものではない。作業的特徴づけは知識が命題的であることを要求せず、知識が能力や理解を含むかどうか、第1章第5節で述べたように決めていない。また本章は、研究上の達成がすべて、ある命題が真であると知ることには尽きるとは想定しない。証明、概念、解釈、モデル、実践的能力、人工物は、社会が表象し、理解し、推論し、行いうることを変えうる。それでも、そうした貢献の一部は命題として表すことができる。ある構成は可能である、ある手続きは特定された条件のもとである結果をもたらす、新しい概念はそれまで結びついていなかった現象をある仕方でも組織する、といった命題である。つまり命題の内容は、新たに観察された自然の事実に限られない (Thurston 1994)。たとえば、多面体についてのオイラーの公式の歴史では、その結果、すなわち頂点の数から辺の数を引き、面の数を足すと2になるという公式を、オイラーは立体の部分についての概念によって述べたが、その一つはまず自ら名づけなければならなかった。辺にはまだ固有の名がないとして、オイラーはそれを *acies* と呼んだ (Sandifer 2004, 2-3)。その結果が表していたのは、観察された自然の事実ではなく、立体の部分の数のあいだの関係である。この例は、前述のとおり、概念を整えることから生まれるものも命題として表現しうることの一例である。他方で、その命題が知識として確立したことまでは示さない。マスグレイヴ (Musgrave) とピグデン (Pigden) が百科事典の項目でまとめるラカトシュ (Lakatos) の見方では、非形式的な数学の予想とその証明は批判を通じてともに改良され、反例は隠れた補題を明るみに出して証明を改め、証明は予想が成り立つ条件を見いだして予想を改めるからである。そこでは、証明は欠陥を免れていなくても尊重に値しうる (Musgrave and Pigden 2026, §§2.1, 3.1)。さらに一般に、貢献が命題として述べられうるからといって、その命題が達成を尽くすことにはならない。ある出力について命題を書けること、その出力がその命題を支持すること、ある主体がその命題を知っていること、その出力そのものが知識であることは、それぞれ別の主張である。また、認識的に価値あるものがすべて知識だということにもならない。

このように本章は、あらゆる認識的達成を知識に還元しない。知識は、価値あるあらゆる過程の変化や達成の名前ではなく、理解、能力、概念の再編、経験的十全性、認識された無知は、それぞれの基準で評価される。それでも、知識が要求の厳しい目標成功であることに変わりはない。何か新しいもの (something new) や認識的改善 (epistemic improvement) なら問わすにおけるような、真理・帰属・非偶然性についての問いを、知識は避けられないからである。本章は、異なる認識論的説明の下で知識が何を要求するのか、知識が他の達成とどう関係するのか、そして各説明が人工システムによる研究において何を見えるようにするのかを見定める。

2. 過程における状態——どのような種類の認識的態度が問題なのか

命題的知識に関連する中心的な哲学的用法において、 p と信じるとは、およそ p が成り立っているとみなすことである。この導入的な注釈は、信念の理論よりもはるかに薄い。それは、信念が根本的に何であるかについての代表的な説——信念を、蓄えられた表象とみる表象主義、思考と行為にわたって表出される傾向性とみる傾向性説、知覚やほかの状態や行為とのあいだで果たす役割によって定まる状態とみる機能主義、合理的な主体として解釈するときには帰属できる状態とみる解釈主義、真理と合理性の規範に構成的に服する状態とみる規範的説明 (Schwitzgebel 2023) ——のどれも選ばない。信念条件は伝統的な分析の一部であり、イチカワとシュトイプ (Ichikawa and Steup) によれば、20世紀の知識の分析をめぐる文献の多くはその分析の批判から出発した。それでも現代の議論は、信念が命題的知識のあらゆる事例にとって必要であるかどうかについても争っている (Ichikawa and Steup 2026; Myers-Schulz and Schwitzgebel 2013)。

研究は確定した信念だけから成るのではないので、いくつかの隣接する態度が重要となる。命題は、その帰結を探るために一時的に仮定されることがありうる。モデルは、その理想化が文字どおりには信じられていないときでさえ、ある課題のために受け入れられることがありうる。確信は、カテゴリーカルではなく程度をもちうる。判断は肯定の行為を名指しうのに対し、信念は持続する状態を名指す。いくつかの説によれば、判断保留は、検討された問いに対する一つの態度であって、単にいかなる表象も欠いていることではない。さらに、候補は、信じられることなく思い抱かれる (entertained) ことがありうる。たとえば、ある歴史家は、アーカイブの各ファイルを要約するノートとは別に、自由に連想を広げ、「自己検閲せずに書く」ための「アイデア・ファ

イル」(Idea File)を設けていると述べている。そこに書いたものは完成した文章には現れないかもしれず、その歴史家はそこに、文書のなかの意外な細部や意味のある沈黙も書き留めている (Walkowitz 2009)。

これらの区別があれば、一見ひとつにまとまった研究パイプラインのなかで、中心にある状態の意味が段階ごとに変わってしまうのを防げる。仮説を生成することは、それを信じることを要求しない。互いに両立しない説明を比較することは、それらすべてを信じることを要求しない。確率を割り当てることは、合意された閾値を超えるカテゴリカルな信念を自動的に生み出すわけではない。制度的受容は、個々の成員が異なる信念度を割り当てているときでさえ、集合的行為を統御しうる。新しい概念は、固定された集合のあいだで確信を移すだけでなく、候補の空間そのものを変えうる。

このため、後の章で研究の過程全体を記述するときには、信念より広く、認識的態度や、コミットメントという語を使う。たとえば第6章は、ある結果がどのコミットメントを改める理由になるかを問う。これらの語は、命題的知識の分析における信念に取って代わるものではない。信念 (belief) が、研究のあいだに変化するあらゆる表象、確信、仮定、方針、公的地位の名前として使われるのを防ぐためのものである。

2.1 信念の担い手は結果の産出者である必要はない

研究を遂行する行為者または組織されたシステム A は、結果として生じた知識が帰属される主体である必要はない。社会 S は新しさの参照先を与え、誰のために知識が求められるかを特定するが、それによって知識の主体になるわけではない。人工システムが定理を導出し、実験を実行し、あるいはパターンを同定し、それを人間の共同体がのちに評価して知るに至る、ということがありうる。信念条件が採用される場合、その条件は知っていると思われる主体に適用される。それは、人工の産出者がその命題を信じていることを自動的に要求するわけではない。

この区別は、知識が人工エージェントや人工社会それ自体に帰属されるときに避けられなくなる。システムと命題 p との関わり方を、 p がシステムのほかの働きと次第に深く結びついていく順に並べてみよう。システムが p を表現する文を蓄える。促されるとその文を出力する。 p に高い確信を割り当てる。推論と計画において p を用いる。 p のゆえに行為を変える。そして反対の証拠が現れると p を修正する。第2節の冒頭に挙げた機能主義や傾向性説は、信念を、推論や行為のなかで果たす働きによって捉える。その下では、文を蓄えたり出力したりするだけでは、信念があることの弱い証拠にしかならず、推論、予測、行為、驚き、修正と安定して結びついていることが、より強い証拠になる。これに対して解釈主義や規範的説明は、合理的な主体として解釈できることや、規範に応じる主体であることのように、別の種類のものを求め、そのために異なる種類の表象、行為者性、心のあり方を要求するかもしれない。どの説をとるにせよ、蓄えや結びつきについてのこうした証拠が直接にかかわるのは、その状態が信念の役割を果たしているかどうかであって、その状態が何についてのものかではない。人工システムの状態が p の扱う対象についてのものであるかどうかは、第8章第2.2節で取り上げる。

これは、同じシステムを随意に信念者とも非信念者とも呼んでよいという許可ではない。システムのアーキテクチャと振る舞いが証拠を提供し、信念が何から成るのかについての理論が、その証拠が文字どおりの帰属にどう関わるかを決定する。したがって本書は、まず機能的な事実を記述し、それらが信念を構成するとされる際の追加的な説明を明示することができる。このように統合された仕方では p を保持するシステムが、そう保持することに責任を負うのか、あるいは理由に基づいて保持しているのかは、さらなる問いである。第7章第3.3節は、人工の行為者へのより厚い帰属が何を要するかを問うなかで、この問いを取り上げる。

人工社会に知識を帰属できるかどうかは、二つの段階で理論の選択に左右される。第一は、知識にそもそも信念が必要かどうかである。イチカワとシュトイプは、信念条件は真理条件よりわずかに論争的であるにすぎないと述べ、信念なしの知識は可能だと論じた哲学者が少数いると報告している。二人が紹介するラドフォード (Radford) の例では、イングランド史について問われた人が、自分の答えを当て推量にすぎないと思いながら、年号を次々に正しく答える。ラドフォードの説では、この人は自分の答えに確信をもっていないので、それを信じてはいない。それでも、正答が続くことはこの人がその事実を学び、忘れていないことを示すので、この人はそれを知っている。ただし、どちらの主張にも異論がある。暗黙の信念はあるとする応答もあれば、これほど確信のない答えは知識の表現ではないとする応答もある (Ichikawa and Steup 2026, §1.2)。知識第一の認識論も、信念条件を否定する立場ではない。たしかに知識第一の認識論は、信念に追加の成分を加えて知識を構成することはない。しかし、それが退けるのは信念条件ではなくその分析の順序であり、知ることは信じることを含意するという伝統的な見解とは両立する (Ichikawa and Steup 2026)。第二は、信念を必要とする見解のあいだで、信念が何から成るのかについて理論が分かれることである。第2節の冒頭に挙げた諸説の違いがこれにあたる。したがって、ある人工社会が p を知っていると主張されるならば、その社会が選ばれた信念条件を満たさねばならないか、あるいはその知識帰属が、その形での信念を要求しない説明によって支えられていなければならない。問題は知識の主体に関わるものであって、研究のあらゆる道具や参加者に課される普遍的な心的前提条件に関わるものではない。

3. 真理と可謬性

命題的知識の中心的な用法について、本章は事実性 (factivity) を保持する。すなわち、ある主体 X が p であることを知っているならば、 p は真である。これは実質的な作業上のコミットメントであるが、比較的控えめなものである。イチカワとシュトイプによれば、認識論者の大半は、偽であることは知られないという点で一致しており、かつて確信されていたが偽であった見解を当時の人々が「知っていた」と記述する日常の語り方を、文字どおり真なるものではなく一種の誇張として扱う (Ichikawa and Steup 2026)。

事実性を認めても、真理が、問えば真偽を返してくれる操作的な託宣 (オラクル) になるわけではない。 p が真であるかどうかは、研究者がそれが真であることを証明できるか、決定的に認識できるか、確信するに至りうるかと異なる。硬貨は、誰も観察していないときでも表を上にして落ちうる。主張は、利用可能な証拠が誤解を招くものであっても真でありうる。逆に、共同体は偽なる主張に対してその最良の方法を責任をもって適用しながら、誤りを検出できないままにすることがありうる。真理は事物がどうあるかに関わる。証拠的・制度的な手続きは、主体が事物のあり方と関係づけられるに至る可謬的な経路に関わる。

このことは二つの評価を切り離す。時点 t において、探究者たちは、利用可能な最良の証拠と手続きの下で p を受け入れることに根拠づけを与えられているかもしれない。遡及的には、 p が偽であれば、彼らは何ら非難に値することをしていなかったとしても、 p であるという事実的な命題的知識を保有していなかったことになる。第一の評価は、参加者が何を信じる資格をもち、あるいは合理的に許容されていたかに関わる。第二の評価は、その認知的な試みがその事実的な目標に到達したかどうかに関わる。

研究機関は、方法を設計し、証拠を集め、批判を組織し、信頼性を推定し、訂正を伝播させることができる。しかし、あらゆる結論に付された別個の真理ラベルを観察することはできない。人工システムも同じ立場にある。人工システムは、最終的な真理についての不可謬のテストを保有することなしに、証明、観察、介入、外れた予測、追試、物質的な破綻、批判に応答責任を負いうる。これらはいずれも、システムの見込みや主張に対して、その外から押し返してくるものである。以下ではこれを抵抗と呼ぶ。それらの抵抗は、道具、モデル、概念、背景的仮定によって媒介されている。しかし、それらによってコミットメントを修正されうるシステムは、テキスト上のもっともらしさや社会的承認のためだけに最適化されたシステムとは異なる仕方、その対象と関係している（第8章第2.3節は、この関係を世界への応答可能性として取り上げる）。上に挙げた抵抗のうち最後の批判も、社会的承認も、どちらも他者からの反応である。しかし違いがある。批判は、理由を伝えることも、世界の抵抗、つまり外れた予測や物質的な破綻のように世界がこちらの見込みどおりにならないことを伝えることもできる。これに対して社会的承認は、理由も世界の抵抗も伝えないまま得られることがある。

4. 真なる成功を偶然的ではなく認知的なものにするのは何か

イチカワとシュトイブが知識の分析についての百科事典の項目で示すように、ゲティア事例は、一般的な問題の範例的な例示を提供する。主体は、支持された真なる信念に到達しようが、その真理はそれにもかかわらず認知的に不適切な偶然の一致に依存していることがある（Ichikawa and Steup 2026, §3）。これらの事例は、知識の単一の代替分析を決定するものではない。それらは、真なる結果とそれに対する何らかの支持とが、ここで問題となっている要求の厳しい認知的成功をなご構成しえないことがあるのはなぜかを、見えるようにするのである。

ランダムな生成器が真なる命題を出力することはありうる。汚染されたベンチマークは、能力のないシステムを正確であるかのように見せうる。研究者は、偽なる中間的主張から真なる結論を推論しうる。妥当に見える説明が、正しい予測に責任を負うメカニズムとは無関係でありうる。これらの事例は、一つの薄いコミットメントを動機づける。

ここで中心的な目標として用いられる要求の厳しい意味での知識にとって、偶然的な真理は不十分である。関連する態度の真理は、その根拠、方法、有能さ、環境、あるいはその他の真理と結びついた条件に対して、何らかの適切な認知的関係に立っていないなければならない。

偶然的な真理の問題は、個々の信念者や作られた事例に限られない。たとえば、アルフレッド・ケンペ（Alfred Kempe）が1879年に発表した四色定理の証明は、1890年にパーシー・ヒューウッド（Percy Heawood）がその中心的な論証の一段がある種の地図では成り立たないことを示すまで、10年以上にわたって受け入れられていたようである。定理そのものは、1976年になってようやく、別の経路によって証明された（O'Connor and Robertson 1996）。真なる信念がその根拠と非偶然的に結びついていることを要求する説では、ケンペの証明にもとづいてその定理を受け入れた人々は、真なる信念をもってはいたが、それは依拠した根拠のうえでは知識ではなかったということになる。

上の薄いコミットメントは、真理が根拠や方法など「何らかの適切な認知的関係（appropriate epistemic relation）」に立つことを求めるだけで、それがどの関係かは言っていない。つまりこの言い回しは、問題を解決するのではなく、どこに問題があるかを示しているにすぎない。知識を正当化された真なる信念とみる伝統的な分析（第1章第5節）では、この場所を占めるのは正当化の条件である。認識論者のなかには、この条件を、信念が合理的であること、根拠づけられていること（warranted）、証拠によってよく支持されていることとして述べる者もいる（Ichikawa and Steup 2026, §1.3.1）。この分析への反例に対する応答のなかには、正当化の条件を保ったうえで、感受性や安全性のような条件（信念が正しかったのが偶然でないかを、実際とは少し違う状況を考えて確かめる条件。第4.1節）をさらに加える説もあれば、正当化の条件を、たとえば知識についての信頼性の条件に置き換える説もある（Ichikawa and Steup 2026, §§3-6）。そこで、この空所を埋めるために、適切な関係を何であれ「正当化」と呼ぶこともできる。しかし、候補となるあらゆる関係を正当化と呼べば、どの関係が求められるのかについての不一致に対処するのではなく、それを覆い隠すことになるだろう。オルストン（Alston）は、正当化の理論ではこれに近いことが起きたと論じる。フェルドマン（Feldman）が伝えるところでは、哲学者たちは、何を信じるべきか、よい証拠、信頼性、適切な機能などによって、正当化について大きく異なる説明を与えてきたが、オルストンはこの不一致を、**正当化された**（justified）という語が拾い出す信念の客観的な性質はないとすることで説明する。あるのはむしろ、十分な証拠をもつこと、その証拠に基づくこと、信頼できる過程によって形成されることなど、さまざまな認知的望ましき（epistemic desiderata）であり、オルストンはそれぞれの本性、実現可能性、重要性、相互関係を順に検討する（Feldman 2005; Alston 2005, xi-xii）。本章では、知識を厳しい目標として保ち、真なる態度を非偶然的なものにする諸関係を、採用する説明に応じて重みの変わる別々の望ましきとして扱う（第4.1節）。

哲学者のなかには、真なる信念とほとんど変わらないものを**知識**（knowledge）と呼ぶ軽い用法があることを認めつつ、認識論の中心的な問いでは、それより強い意味、つまり真なる信念以上のものを求める意味で知識を論じる者もいる（Ichikawa and Steup 2026）。非偶然性についての薄いコミットメントは、ここで検討する知識の目標のために、より強い意味を選ぶ。ただしそれは、軽い用法で「知っている」と言うことが言葉の使い方として誤りだと示すものではない。より強い目標が有用なのは、運よく真となる出力の頻度を増やすだけのシステムは、まさに本書の探究を動機づけている能力と妥当性確認の問題を残したままにするからである。

以下の小節は次の順で進む。第4.1節は、適切な認知的関係に何が求められるかについて競合するアプローチを概観し、第4.2節は、それぞれを採ったときに人工システムによる研究について何が帰結するかを表2.1で示す。第4.3節は、それらのアプローチの多くの内側で問われる基礎づけの関係を取り上げ、第4.4節は、それが前提にする問い、つまりシステムがもつ証拠とは何かを扱う。第4.5節は、偶然の生成器の事例によって、生成すること、知ること、貢献すること、研究することを切り離す。

4.1 認識的成功を位置づけるいくつかの仕方

目指す達成を知識に定めても、知識であるために何が要るのかを、どの理論にも偏らずに分解できるわけではない。知識について長く議論してきた認識論には、知識であるために何が要るのか、つまり、真なる信念に何が加われれば、成功が運ではなく知識になるのかについて、互いに競合するさまざまなアプローチがある。それらは、主体、内容、その根拠、世界、そしてより広い認識的環境のあいだの関係を、それぞれ実質的に異なる仕方でも組み立てる。そのため、どのアプローチをとるかによって、研究のどの特徴が見えてくるかが変わり、人工研究者についての判断も変わる。

内在主義的でアクセス志向のアプローチは、知る者とされる主体がどのような理由、証拠、反省的アクセスを保有しているかを問う (Ichikawa and Steup 2026, §1.3.2)。証拠主義的アプローチは、ある態度が主体のもつ証拠に適合しているかどうかを問う。たとえば、ミタグ (Mittag) が『インターネット哲学百科事典』の証拠主義の項目で示すフェルドマン (Feldman) とコニー (Conee) の証拠主義では、信念、不信、判断保留のいずれが正当化されるかは、その時点で主体がもつ証拠への適合だけで決まり、その証拠がどのように獲得されたかにはよらない (Mittag n.d., §§2, 3c)。証拠の獲得にかかわる規範がそもそも認識的評価に属するかどうかは、争われている (第5節)。これに対して外在主義的・信頼性主義的アプローチは、信念が正当化されているかどうかを、少なくとも部分的には、その信念を生んだ過程がどのような因果的経緯をたどったか、そしてその過程がどれだけ真理に導きやすいかによって決める。そのため、なぜその過程がうまくいくのかを主体自身が十分に説明できない場合でも、正当化を認めることができる (Goldman and Beddor 2021, §§1.1, 2.1)。この対比は、中身の見えない人工システムにとって重要である。そうしたシステムでは、結果を生んだ過程が信頼できるかどうか、人間の受け手がその結果を支持する理由を手に行っているかどうか、受け手がその結果に依拠してよい資格をもつかどうか、互いに食い違いうる。過程は信頼できるのに、受け手には説明できる理由が何もないこともあれば、受け手は理由を手に行っているのに、結果を生んだ過程は信頼できないこともある。このように食い違いうるからこそ、どのアプローチをとるかで判定が変わる。信頼性主義は過程が信頼できるかを、内在主義は受け手が理由を手に行っているかを見るからである。

徳信頼性主義的アプローチと達成アプローチは、真なる成功が、関連する主体に帰属可能な安定した認識的能力を顕在化しているかどうかを問う。これに対して徳責任主義的アプローチは、知的徳に、開かれた心 (open-mindedness) や良心性 (conscientiousness) のような涵養された性格特性を含め、そのなかには徳が対応する動機づけを要求するかどうかを問うものもある (Turri, Alfano, and Greco 2021)。人工システム自身の知識が問われるなら、このことは、そのシステムのなかの何かが認識的動機づけの役割を果たしているかどうかという、さらなる問いを提起する。様相的アプローチは、実際とは少し違う状況を考えて、信念が正しかったのが偶然でないかを確かめる。問い方は二つある。一つは、状況が少し違っていても、同じように信じていれば、やはり正しかったと言えるかを問う。言えるとき、その信念は安全性を満たす。もう一つは、もし p が偽だったら、 p を信じていなかったはずかを問う。そうであるとき、その信念は感受性を満たす (Ichikawa and Steup 2026, §5)。無効化 (defeat) に注目する説明は、いったん正当化された信念が、それに反する情報、すなわち主張の否定を支持する情報 (反駁の情報) や、主張とその根拠との結びつきを疑わせる情報 (掘り崩しの情報)、あるいは自分の判断そのものの信頼性を疑わせる情報 (高階の情報) を得たときに、その正当化を失うのか、どう失うのかを問う (第5節)。また、これらのアプローチの多くの内側では、さらに別の問いが生じる。主体があるコミットメントを、それを支持するとされる支えのゆえに保持しているのか、それとも単に別の経路で到達された結論と並んでその支えを保有しているだけなのか、という問いであり、第4.3節が基礎づけ (basing) の問いとして取り上げる。

ここまで見てきたアプローチは、内在主義から無効化に注目する説明まで、いずれも正しい出力と、偶然でない達成とを区別する。ただし、成功の担い手の境界をどこに引くかは、アプローチによって異なりうる。成功を、人工システムに帰すこともあれば、それを使う人間やほかの人工システムに帰すことも、両者を含む全体に帰すこともありうる。

知識第一の認識論は、条件をもう一つ加えるのではなく、説明の順序を変える。ウィリアムソン (Williamson) は、知識を説明されない出発点とし、ほかの認識的な概念を部分的にそれによって説明する。 p を信じるとは、おおまかには、 p を知っているかのように扱うことであり、正当化は主として知識が与える地位であり、自らの証拠は自らの知識であり、知っていることだけを主張すべきである (Williamson 2000, 9–10, 46–47, 185, 243)。知識を原始的なものとして扱っても、本節で区別した諸条件が評価にとって無用になるわけではない。ウィリアムソンは知ることを分析不可能とみなすが、その理由は二つある。一つは、ウィリアムソンの説では、知ることはそれ自体ひとつの心的状態であるのに、標準的な分析はどれも、真理という心的でない条件を含むことである。もう一つは、ゲティアの反例以後に出された分析が、複雑になるたびににより複雑な反例に退けられてきたことであり、その説によれば、これもこの見方から予想されるとおりである。それでもウィリアムソンは、上で様相的アプローチとして見た二つの条件のうち、知識には安全性が必要だが、感受性は必要ではないと主張している。感受性を必要としない理由は、次の例に見ることができる。見る距離をわずかに過小に見積もる者が、水面から1メートルの高さにある印について、それが20メートル未満の高さにあると見て判断するとしよう。この信念は感受的でない。印が20メートル未満の高さになかったとしたら、20メートルをわずかに超える高さにあったかもしれない、その場合にもこの者は20メートル未満だと判断したであろうからである。それでもこの者はそのことを知っている。1メートルの距離と21メートルの距離はまったく違って見えるので、この信念は、命題が偽であった場合に依拠したであろう証拠とはまったく異なる証拠にもとづいているからである。系統的な偏りがどれほどわずかでも、ゼロでない限り、偶然的 (contingent) な命題についての主体の信念のほとんどすべてが、同じ仕組みによって感受的でなくなりうる。さらにウィリアムソンの見方では、知ることの必要条件は、知識の定義を組み立てる条件の一つである必要はない。ここでいう定義とは、知識という概念を使わずに書ける条件を並べ、そのすべてを満たせば知識になる、という形の分析である (Williamson 2000, 30–31, 33, 147, 159–161, 307–310)。したがってウィリアムソンの説では、安全性は知識を定義する構成要素ではないが、必要条件ではある。安全性についての証拠は、これとは別の事柄であり、知識の帰属を評価するときに用いられる。結果がたやすく誤りえたことを示す証拠は帰属に不利にはたらくが、安全性の証拠だけで知識が確立されるわけではない。知識第一の説のうえでは、第4.2節の表2.1もこのように読める。各行が示すのは、採用した説が知識に求めるものと、帰属を支えるために確立すべきものであって、定義の部品ではない。

シミオン (Simion) の自然主義的な説は、認識的な領域を知識の生成というひとつの機能を中心に組織する。ここでいう機能は、来歴によって決まる機能 (etiologial function) である。つまり、探究がこの機能をもつのは、過去の知識の産出が認識的な便益をもたらし、その便益がこの実践の存続を説明する助けになるからである。そうした来歴は、世代を越えた選択によってだけでなく、継続的な選択や学習の過程によっても与えられうる (Simion 2025, 7–8)。この説を訓練された人工システムに当てはめると、そうしたシステムがその機能をもつかどうかは、その現在の出力がどれほど頻繁に真であるかだけ

でなく、その訓練と選択が何に報いたかにも部分的に依存する。この当てはめを進めると、現在の正答率が同じ二つのシステムでも、一方の訓練と選択が正しい答えに報い、他方がもっともらしい答えや評価者の承認に報いてきたなら、この説のもとでの知識の帰属は分かれる。ただし、訓練で掲げられた目的をそのまま来歴による機能とみなすことは、それ自体が一つの前提である。実際の選択が何に報いてきたかは、掲げられた目的と食い違いうるからである。自然化された知識第一主義的機能主義を集団へ拡張するバード（Bird）の議論は、第6節で扱う。

社会認識論はさらなる選択肢を導入する（第6節）。認識的地位は、証言、分割された専門知、組織化された記憶、批判的相互作用、あるいは集団によって実現される機能に依存しうる。探究中心的・理解中心的なアプローチ（第1章第3節、第3章）もまた、研究に関連する認識的变化が、確定した命題的知識の獲得に尽きることかどうかを問いに付す。

これらのアプローチと正当化との関係は一樣ではない。内在主義的・証拠主義的アプローチと過程信頼性主義は、何が信念を正当化するのかについての説であり、内在主義と外在主義の対立は、正当化をめぐる対立である（Ichikawa and Steup 2026, §1.3.2）。感受性と安全性は、知識に加わるさらなる様相的条件として提案されている（Ichikawa and Steup 2026, §5）。知識についての信頼性主義の説は、正当化の条件の代わりに信頼性の条件を置く（Ichikawa and Steup 2026, §6.1）。ソーサ（Sosa）の説のような徳の説では、知識とは適切な（apt）信念、つまり巧みに生み出されたがゆえに真である信念である。そのため知識は正当化を含むが、その正当化は巧みさ（adroitness）、つまり信念が技能によって生み出されていることとして理解される。ただし、求められるのは技能が成功を説明することなので、正当化は、成功が有能さの功績であることから切り離せる構成要素ではない（Ichikawa and Steup 2026, §10）。基礎づけは、支持が手元にあるだけの信念と、その支持に基づいて保持されている信念とを分ける関係にかかわる（第4.3節）。知識第一の説は、上で見たように、正当化を知識から説明する。第4節の初めに置いた薄いコミットメントは、これらのどれも選ばない。適切な認識的關係が、これらの説のどれかの意味での正当化なのか、正当化に条件を加えたものなのか、正当化に置き換わる条件なのかは、決めずにおく。

これらは、一つの最大限の定義へと積み上げられるべきモジュールではない。また本書は、何らかのさらなる分析が可能となる前に一つの説明を選ぶ必要もない。これらは条件的に用いることができる。アクセス可能な理由が要求されるならば、不透明性と移転についての一群の問題、つまりシステムが結論に至った理由が受け手から見えないという問題や、その理由が受け手自身の理由として引き継がれるかという問題が帰結する。信頼できる外的な結びつきで十分であるならば、別の一群が帰結する。有能さ、様相的頑健性、理解、あるいは社会的組織化が中心的であるならば、成功の所在と意味はまた変わる。したがって、哲学的立場は、認識上の成功を異なるところに求めることの帰結を辿るための資源であって、一つの隠れたパイプラインの別名ではない。それぞれの立場をとったときに何が求められ、どのような見かけの成功を知識と取り違えやすいかは、第4.2節の表2.1で立場ごとに示し、第9節で人工システムによる研究に求められることとしてまとめる。

4.2 条件的な帰結

第4.1節で区別したアプローチは、知識に何が求められるかについて互いに競合する説である。表2.1は、そのそれぞれについて、そのアプローチを採ったときに何が帰結するかを示す。左の列は、各アプローチを、それが何を重く見るかによって示す。中央の列は、その場合に知識に何が求められ、知識の帰属を支えるには評価者が何を確かめる必要があるかを示す。右の列は、特徴的な偽陽性を示す。つまり、そのアプローチの下では知識に数えられないのに、評価では知識と取り違えられやすい見かけの成功である。行ごとに読めば、どのアプローチを採るかによって、人工研究システムに何を作り込み、何を確かめる必要があるかがどう変わるかが分かる。ただし、この表は各説を厳密に述べたものではなく、見通しを得るために単純化したものである。各行は、いくつもの説からなる系統を一つの代表的な要求にまとめており、偽陽性もいくつかありうる例の一つにすぎない。同じ系統のなかで説がどう異なるかは、表の後の段落で、信頼性主義と徳のアプローチについて示す。また右の列の偽陽性は、システム自身が知っていると思われる場合を例にとっている。人間や共同体がシステムを通じて知るようになる場合には、中央の列の各要求はその人間や共同体の信念にかかり、システムの出力は、その信念にとつての証拠として、あるいはその信念を形成した過程の一部として効いてくる。たとえば事後的な理由づけも、それが示す論証が健全であれば、つまり推論が妥当で前提も真であれば、その人間が実際にもつ理由になりうる（第4.3節）。

表2.1 認識論のアプローチごとに帰結すること（単純化した見通し）

採用された説明が強調するのが……	ならば、知識が求め、その帰属を支えるには確立する必要があるのは……	特徴的な偽陽性
アクセス可能な理由	関連する主体に利用可能で、その主体が使用できる理由	実際に作動した経路とは無関係な、説得力のある事後的な理由づけ
証拠への適合	主体が保有する証拠に釣り合った態度	反対の証拠があるにもかかわらず、選択的な検索によって生み出された高い確信
過程の信頼性	関連する環境における、主体の信念の真理促進的な形成または維持	課題分布を越えて一般化された、一つのベンチマーク・スコア
様相的安全性または感受性	知っていると思われる主体の信念が、重要な近傍の可能性において偽とならないこと、あるいは主張が偽であったなら抱かれなかったであろうこと	ほぼ同一の条件または汚染された条件の下での反復的成功
有能さ（徳信頼性主義）	主体またはシステムに帰属可能な安定した認識的能力ゆえの成功	運のよい一回の試行、あるいは決定的な選択が隠れた人間によって行われたパイプライン
知的性格（徳責任主義）	開かれた心や良心性のような特性と、それらの動機づけの役割を果たすものを表出する探究	探究の遂行を統御することなく報告のなかで再現される、留保や自己批判
無効化への感応性	反駁的情報と掘り崩し的情報への適切な応答	作動しているコミットメントは変わらないまま、ユーザー向けの報告だけが訂正されたもの
知識第一の機能主義	知識の生成を機能とする能力または組織されたシステムの適正な作動	知識を条件の組に分解する従来のチェックリストに、その機能についての独立した説明なしに、名前だけ「知識第一」を付けたもの

なお、表には三つのものを含めていない。一つは基礎づけである。基礎づけは、ほかと並ぶ一つの強調点ではなく、信念とその根拠とのあいだの関係であり、表のアプローチの多くが何らかの形で求めるものである。そのため、どのアプローチを選ぶかは別に、第4.3節で扱う。アクセス可能な理由と証拠への適合の行が、理由や証拠が手元にあり、態度がそれに合っていることだけを述べているのはそのためであり、信念がそれに基づいて保持されていることは第4.3節で加わる。二つめは社会的な説であり、これには別の表がある（第6節、表2.2）。三つめは探究中心的・理解中心的なアプローチであり、第3章で検討する。表の中央の列にある要求は、そのアプローチを採った場合にだけかかる条件付きの要求である。すべてを足し合わせて、あらゆる人工研究者が満たさなければならない一つの基準の一覧にするものではない。

そうした帰結の一例として、信頼性の行を詳しく見よう。信頼性主義は、信頼性を生み出す要因への反省的アクセスなしに知識を認めうる（Goldman and Beddor 2021, §§1.1, 2.1）。ただし、信頼性は過程の類型に属するものであり、一つの個別の過程は、信頼性の異なる多くの類型にあてはまる。たとえば、薄暗い部屋で遠くの時計を読んで時刻を信じた人の過程は、「視覚」とも、「薄暗いところでの視覚」とも、「薄暗いところで遠くの小さな数字を読むこと」とも記述でき、それぞれの信頼性は異なる。どの類型が判定を決めるのかが一般性問題（generality problem）であり、信頼性主義者たちはこれにいくつかの応答を提案してきた（Goldman and Beddor 2021, §2.3）。人工システムを通じて形成された信念の場合、それがシステム自身のコミットメントであっても、システムの出力に頼る人間の信念であっても、一般性問題があるために、信頼性主義的な判定は、それを生み出した過程がどう類型化されるかに依存することになる。モデル単独か、検索と人間による選別を含めたモデルか、所与のタスク分布の上でのパイプラインか、である。隠れた人間の選別がモデルの誤りを取り除いているなら、全体は信頼できて、モデル単独はそうでないかもしれない。ベンチマークを選ぶことも、部分的にはそうした類型を選ぶことである。ベンチマークの成績が示すのは、一つのタスク分布の上での一つのタイプの信頼性であって、システム一般の信頼性ではない。大塚も、実際にはAIのモデルはベンチマークによる試験を通じて事例ごとに正当化されており、そのため信頼性が個々のモデルに結びつけられるようになったと指摘している（Otsuka 2026, "Understanding the Machine Virtues"）。そのため、人工システムが信頼できると主張するには、何を単位としているのか、つまり人間の選別をその単位に含めているのか、どのタスク分布で確かめたのかを示す必要がある。

統計的検定は、一般性問題の具体的な一例である。大塚（Otsuka）は、ひとつのデータセットが、ある実験計画のもとでは帰無仮説の棄却に、別の実験計画のもとではその保持に至るという停止規則の問題を、一般性問題の一つの形態であると論じる。大塚の例では、硬貨を20回投げた表が6回出る。20回投げると決めた計画のもとでは、硬貨は公正だという帰無仮説を有意水準5%で棄却できないが、表が6回出るまで投げ続けると決めた計画のもとでは棄却できる。観察された結果と同じかそれより極端な結果が出る確率が、前者では0.0577、後者では0.0318と、計画によって異なるからである。つまり同じ投げの列が、判定の異なる二つの検定過程の事例になる。大塚が取り上げる頻度主義の応答の一つは、同じ実験が異なる検定として記述されうること認めたい。実験者に検定の過程をあらかじめ特定させ、それに従うことを求める。そのとき判定は、その過程に相対的にのみ正当化され、データを見たあとで別の計画に切り替えることは、検定のハッキングとみなされる（Otsuka 2023, 99–103）。この要求を類比によって人工システムに広げると、それは、評価される過程の類型を結果の検討に先立って固定し、あとのタイプの付け替えを、中立的な記述し直しではなく報告されるべき変更として扱うことになるだろう。

ほかのアプローチは、別の人間の介入や評価の失敗を際立たせる。徳アプローチは、有能さを顕在化する成功に重みを置く。ジョン・トゥリ（John Turri）、マーク・アルファノ（Mark Alfano）、ジョン・グレコ（John Greco）が示すソーサ（Sosa）の説では、成功は、その生起だけでなくその正しさが、それにふさわしい条件のもとで行使された有能さに帰属するとき、適切（apt）である（Turri, Alfano, and Greco 2021, §5）。様相的アプローチは、その信念があまりにも容易に偽でありえたのではない、あるいは偽であったとしてもなお保持されていたのではないかを問う。ソーサは、適切さを安全性から切り離す。その説では、遂行者の有能さやその行使の条件が容易に失われた場合、遂行は安全でなくても適切でありうる。逆に、成功が確保されていても、それが遂行者の有能さのおかげでなければ、遂行は安全であっても適切ではない。適切さと安全性のどちらも他方を含意しない、というのがソーサの立場である（Sosa 2005, §§8–9）。トゥリらの紹介によれば、ソーサはまた、成功がそのおかげである有能さは社会的に座を占めうる、つまり個人でなく社会的な集まりのうちにありうるとする。その説では、証言によって受け取った信念の正しさは、わずかに受け手の証言を信頼する有能さに負い、より十全には伝達の連鎖に連なる人々に集合的に座を占める有能さに負う。タッチダウン・パスがクォーターバックの有能さを顕在化しつつ、より十全にはチームの有能さを顕在化するのと同じである。受け手が得るのは部分的な功績だが、それで知識に足りうとソーサは論じる（Turri, Alfano, and Greco 2021, §7）。ソーサの適切さと安全性の区別に照らすと、モデルの誤りを補正する隠れた人間の選別は、出力を安全にしながらか、その成功をモデルの功績にはしないことがありうる。そのとき成功がモデルの功績か、道具を含めたモデルの功績か、より広い人間と機械の組織の功績かは、有能さがどこに座を占めるとみなされるかに依存する。これは、上で論じた過程の類型化と並行する選択である。大塚も、深層学習モデルについて関連する結論を引き出している。深層学習モデルの信頼性は、ベンチマークで事後的に、モデルごと、テストセットごとに確かめられるので、名前のついた個々のモデルに帰されるようになる。大塚は、信頼性主義もそうしたモデルを受け入れうることを認めつつ、その現在の正当化を考える手がかりが徳認識論にあるかもしれないと示唆している（Otsuka 2023, 132–135）。

4.3 基礎づけと作動しているコミットメント

主体は、 p を信じるよい理由を保有しながら、その理由のゆえに p を信じているわけではない、ということがありうる。たとえば研究者は、ある結論を支えるデータを手にしてしながら、その結論を、先輩の研究者がそう言ったから信じているかもしれない。認識論では、この違いを、正当化の二つの概念で表すのが標準的である。命題的正当化（propositional justification）は、主体がある命題を信じる十分な理由をもっているかどうかにかかわり、信念的正当化（doxastic justification）は、実際の信念が適切な仕方であるか保持されているかどうかにかかわる。両者の関係のよくある捉え方では、命題的正当化のほうがより基本的であり、信念的正当化は、主体の信念がその命題的正当化に適切に応じている、あるいはそれに基づいている、ということである。イチカワとシュトイブによれば、正確な関係は争われているが、両者が食い違いうることには異論がない。上の研究者の場合がそうである。またイチカワとシュトイブは、知識の伝統的な分析が求めているのは、ふつう信念的正当化のほうだと考えられているとも述べている（Ichikawa and Steup 2026, §1.3.3）。基礎づけ関係（basing relation）とは、ある理由が、ある信念がそのゆえに保持されている理由であるときに、その理由と信念とのあいだに成り立つ関係である。基礎づけは、信念が正当化されている、つまり信念的正当化があるための必要条件ではあるが十分条件ではない、と一般に考えられている。また

それは、主体がもつよい理由のうち、その信念の正当化に寄与するものと寄与しないものとを分ける (Korcz 2026)。つまり、基礎づけを問うとは、よい支持があるかどうかではなく、信念がその支持のゆえに保持されているかどうかを問うことである。

この区別は、第4.1節の諸アプローチとは次元が違う。諸アプローチのうち、内在主義、証拠主義、信頼性主義などは、何が信念を正当化するのかについての説である (第4.1節)。主体がアクセスできる理由、主体のもつ証拠、信頼できる過程などが、その答えの候補である。どの説を採っても、正当化するとされるものを主体がもっていることと、信念がそのゆえに保持されていることは区別できる。上で見たよくある捉え方では、前者が命題的正当化であり、後者の基礎づけが加わると信念的正当化になる。信頼性主義では、正当化するのは信念を実際に引き起こした過程であるので (Korcz 2026, §5)、正当化するもののゆえに信念が保持されているという要素は、はじめから正当化の条件に含まれている。したがって基礎づけは、真なる信念に何かが加われれば知識になるのかという問いへの、諸アプローチと並ぶもう一つの答えではなく、それぞれのアプローチの内側で問われる関係である。表2.1が基礎づけを外したのはそのためである。同じ理由から、結論を支えるよい理由をもっていることと、その理由のゆえにその結論を保持していることとの区別は、どのアプローチを採るかを定める前に、それとは別に検討できる。

たとえば証拠主義者のフェルドマンとコニーは、よい理由をもつことと、その理由のゆえに信じることとの区別を、独自の二つの概念で表している。一つは正当化である。フェルドマンとコニーの説明では、態度は、主体がもつ証拠に適合するとき正当化されている。もう一つは十分な基礎づけ (well-foundedness) である。態度が十分に基礎づけられているのは、それが正当化されているに加えて、主体が、その態度に適合する証拠の一部に基づいてそれを保持しており、しかもその支持を打ち消すような、自分のもつほかの証拠を無視していないときである (Feldman and Conee 1985, 24-25, 33 n. 21)。この二つは、上で見た命題的正当化と信念的正当化に近い。

本小節は、この区別が人工システムによる研究に何を求めるかを問う。基礎づけは、知っているとされる主体の信念と理由とのあいだの関係である。そして第2.1節で見たように、研究を行うシステムは、その主体である必要はない。システム自身に知識を帰属させるなら、問われるのはシステム自身のコミットメントの基礎づけであり、問いは、結論の傍らにある理由のうち、どれが実際にその結論を生んだのか、ということになる。コミットメントは、第2節で述べたように、本書が信念より広い語として使うものであり、ここでは、システムが保持し、出力し、後に改めうる結論を指す。見出しの「作動しているコミットメント」は、システムが何を出力し、後にそれをどう改めうるかを実際に左右しているコミットメントを指し、傍らに表示されたり保存されたりしているだけの結論や理由とは区別される。これに対して、人間や共同体が知識の主体であるなら、問われるのはその主体の信念の基礎づけであり、システムの出力は、その主体にとっての証拠や情報源の一つになる。社会的な主体は第6節以下で扱う。出力に頼る主体は、ほかの人工システム、たとえば混合社会や人工社会である *S* の人工の成員であることもある。その場合、その受け手が出力から受け取るものは受け手自身の作動しているコミットメントになり、本小節の問いは受け手の側で生じる。

次の三つの事例は、この点で互いに異なる。AIの報告書は、*p* についての優れた証拠を含んでいながら、システムは無関係な近道のゆえに *p* を産出した、ということがありうる。システムが *p* を知っているとされるなら、その作動しているコミットメントを左右しているのは、その証拠ではなく近道である。ユーザーは、傍らに表示された証拠を無視しつつ、制度的な威信のゆえにAIの結論を信じることがありうる。この場合の主体はユーザーであり、その信念は証拠ではなく威信に基づいている。リポジトリは、結論とその根拠の両方を含みながら、それらを結びつけるいかなる作動する過程ももたないことがありうる。この場合には、信念が何かに基づいているような主体が、まだいない。第7節の読まれていない発見に近い状況である。いずれの場合も、支持と結論が同じ場所にあることは、態度や出力が適切に基礎づけられていることを立証しない。

システム自身がその結論を知っているとされる場合、基礎づけは、システムをどう評価するかに直接かわる。評価者がふつう手にするのは、システムが出力した結論と、それに添えられた論証、引用、データである。それらを吟味して分かるのは、せいぜい、システムがその結論を支えるよい理由をもっているかどうかであり、これは命題的正当化の問題である。基礎づけが問うのはその先、つまりシステムがその理由のゆえに結論を出したのかどうかであり、信念的正当化はこれに左右される。そしてこれは、出力を見るだけでは分からない。知識の伝統的な分析が求めるのはふつう信念的正当化のほうだと考えられているので、そうした条件を課す説では、論文が説得的であることは、システムがその結論を知っていることを示さない。この違いは、第3章第5.3節が区別する十全性の問いのうち初めの二つ、つまり、ある説明が出力を産出した過程を忠実に記述しているかと、利用可能な根拠がなぜその主張の受け入れを支持するのかを明確化しているかとの違いに近い。一方、知識の主体が出力に頼る人間であるなら、同じ隔たりは人間の側に現れる。説得的な論文を読めば、人間は理由を手にする。しかし、その人間の信念がその理由に基づいているのか、上の二つ目の事例のようにシステムの威信に基づいているのかは、また別の問題である。

システムが主体である場合、この違いは後の修正にも現れる。結論が、添えられた理由に基づいていないなら、その理由が誤りだと分かっても、システムが結論を改めるとはかぎらない。表2.1の「無効化への感応性」の行の偽陽性、つまり、ユーザー向けの報告だけが訂正され、作動しているコミットメントは変わらないままである場合が、これにあたる。逆向きの変化もありうる。結論が最初は近道や偶然によって生じたとしても、システムが後に独立した証拠を得て、その結論を保つか改めるかがその証拠に応じて決まるようになれば、基礎づけの説によっては、その時点で結論の地位は変わりうる。問われるのは、結論が最初にどう生じたかだけでなく、いま何によって保たれ、何に応じて改められるかである。第4.5節の偶然的生成器の事例では、共同体の信念について同じことが成り立つ。

では、システムの結論が証拠に基づいている、あるいはシステムが証拠に応じて結論を改める、と主張するには、何を示せばよいのか。結論が生成された後に文書を添付しても、支持を結論の傍らに置くだけである。それを示す手段としてまず考えられるのは、結論がどこから来たのかの記録、つまり来歴 (provenance) である。どの入力、モデル、計算、チェック、介入が、システムの作動しているコミットメントを生み、その後の修正を左右したのかを示す記録である。この要求は、何が実際に結論を左右したかという因果の言葉で述べられている。そのため、これがもっともよく合うのは、理由が適切な仕方では信念を引き起こしていることを基礎づけの条件とする、基礎づけ関係の因果説である。これに対してメタ信念説 (doxastic theories) では、理由についての信念、つまりその理由がその信念にとってよい理由だという適切なメタ信念があれば、信念はその理由に基礎づけられている。この説では、理由が信念を適切な仕方では引き起こしていなくても、信念がよい理由に基礎づけられていることがありうる (Korcz 2026, §§1, 3, 5)。そうした説を採るなら、何が結論を引き起こしたかだけでなく、システムのなかにそうしたメタ信念の役割を果たすものがあるかどうかとも問題になる。ただし、どちらの説でも、

支持と結論が同じ場所にあるだけでは基礎づけにはならない。そのため、上の事例が示す失敗は、どちらの説からも失敗と認められる。ここで区別できることは三つある。一つは、基礎づけが成り立つための条件である。二つめは、それが成り立つと帰属するための証拠であり、三つめは、その証拠を残す仕組みである。来歴の記録は三つめにあたり、帰属の有力な証拠を与えるが、記録がないことは、基礎づけの関係がないことと同じではない。入力を変えたときに結論が変わるか、理由が退けられたときに結論が改められるかを確かめる検査も、帰属の証拠になりうる。どの証拠で帰属を支えられるかは、採る基礎づけの説と、そのシステムに対して行える検査とによる。

4.4 システムがもつ証拠と検索の蓄え

証拠への適合としての正当化（第4.1節）と、証拠に基づく基礎づけ（第4.3節）は、どちらも、さらに別の問いへの答えを前提にしている。主体がもつ証拠とは何か、という問いである。証拠への適合は、態度が主体のもつ証拠に適合しているかを問い、基礎づけは、態度がその証拠のゆえに保持されているかを問う。本小節が問うのは、そのどちらでもなく、両者が前提にしているこの問いであり、扱うのは、システム自身が知っているときとされる場合である（第4.3節）。そうしたシステムの場合、参照できるものの一部は、データベースのような検索の蓄えにありうる。蓄えのなかのものがシステムのもつ証拠に数えられるかどうかによって、正当化についての判定も、基礎づけについての判定も変わりうる。フェルドマンとコニーの二つの概念（第4.3節）も、どちらも主体がもつ証拠を基準にしている。ところが二人は、何かを証拠としてもつとはどういうことかは不明確だと認めている。たとえば、促されてはじめて思い出せる命題が、もっている証拠に数えられるかどうかである（Feldman and Conee 1985, 32 n. 2）。データベースのような検索の蓄えを用いるシステムでは、この不明確さが、システムをどう組織し評価するかの問いになる。たとえば、蓄えのなかに結論に反する証拠があり、蓄え全体としては結論を表明された程度までは支えないのに、システムが結論に有利な文書だけを検索し、確信をもって結論を出したとする。蓄えのなかのものをすべてシステムのもつ証拠に数えるなら、この結論はシステムの証拠に適合していない。表2.1の「証拠への適合」の行の偽陽性がこれにあたる。検索されたものだけを数えるなら、結論はシステムのもつ証拠には適合しうる。その場合、欠陥は結論の保持の仕方ではなく、証拠の集め方のほうにある。証拠の集め方が認知的評価にどうかかわるかは、第5節で扱う。この比較からは、もう一つのことも出てくる。証拠への適合を求める説のもとでは、蓄えをシステムのもつ証拠に数えるように境界を広げることは、システムの認知的地位を高めるとは限らない。それまで外にあった反証もシステムのもつ証拠になるので、結論を改めない限り、同じ出力が正当化を失いうるからである。手元に置く情報を増やすことは、それだけでは認識上の強さにならない。

ウィリアムソンの説は、主体がもつ証拠の境界を別の場所に引く。その説では、自らの証拠とは自らが知っていることであつて、知る位置にあるにすぎないことではない。ウィリアムソンは、後者を数えれば、全体としてはアクセスできない証拠の集まりのうえで仮説が確からしくなってしまうと論じる（Williamson 2000, 202–203）。この見方では、知ることは活動ではなく状態であり、人は多くの命題を考慮することなく知っている。記憶に情報を蓄える生き物についてのウィリアムソンの単純なモデルの一つ（第8.2節で再び取り上げる）では、取り出しに時間がかかっても、その生き物は記憶に記録されたすべてを知っていると約定される（Williamson 2000, 199, 226–227）。ウィリアムソンの説を採るなら、それは検索の蓄えをどう組織するかという問いを、システムがその蓄えの記録していることを知っているのか、それとも知るに至る位置にあるにすぎないのか、という問いに変えるだろう。

システムが頼る蓄えが、システムの信じていることの一部を保持しうるかどうかは、拡張された認知、つまり認知が脳やシステムの外の道具や記録にまで及ぶうるかをめぐって争われている問いであり、第5章第6.1節はそれを決着させずに示している。記憶の認識論は、同じ問いを記憶による知識について取り上げる。マシュー・フライズ（Matthew Frise）の報告では、J・アダム・カーター（J. Adam Carter）とイエスパー・カレストルプ（Jesper Kallestrup）は、ノートに蓄えられた情報が主体の傾向的な記憶知識の一部でありうるので、主体はそれを参照するたびに学び直しているのではないと論じる。これに対してコーケン・マイケリアン（Kourken Michaelian）は、この論拠は成り立たない機能上の類似に拠っていると論じる。この論拠は生物の記憶が純粋に保存的であることを求めるが、いま優勢な見方では、想起することは記録を単に再生することではなく、新しい内容を生成することだからである（Frise 2023, §§1.1, 3.2）。蓄えられた記録を返すのではなく内容を生成する部分と、預けられたものを返す蓄えとをあわせもつシステム、たとえば検索の蓄えを備えた学習済みのモデルを考える。マイケリアンが生物の記憶と外部の記憶のあいだに引く対比をそうしたシステムに移すなら、その対比はシステムの内側に入りうる。そのため、蓄えの記録がシステムの知っていることに含まれるかどうかは、システムの境界を蓄えを含むように引くことでは決まらない。それにかかわるのは、むしろ、記録がシステムのすることにどう入り込むかである。第2.1節で見た信念の機能的な説では、文が蓄えられていることは信念の弱い証拠にすぎず、推論、行為、修正への安定した統合のほうが強い証拠になる。そうであれば、システムが検索できても、その推論にも修正にも影響しない記録は、システムが信じたり知ったりしていることの一部というより、見直すために残されたノート（第5章第6.1節）に近い。したがって、蓄えの中身をシステムが知っていることみなす評価には、検索できることだけでなく、そうした統合の証拠が要る。社会について対応する問い、つまり、どの成員も読んでいない結果を社会の基盤が保持している場合の問いは、第7節で取り上げる。

4.5 生成すること、知ること、貢献すること、研究すること——偶然の生成器の事例

結果が人工システムから出てくるときに区別しておく必要のある述語を、ひとつの極端な事例が切り離す。ランダムな生成器がたまたま新しい真なる定理を出力したと仮定しよう。その出来事は、真なる候補内容を生み出す。もし誰もそれを、生成器の圧倒的に多数の偽なる出力から区別できないならば、それはまだ要求の厳しい意味での知識を生み出してはいない。もし有能な数学共同体がのちにその定理を証明するか、別の経路で得られた証明を確認するかして、それを知るに至るならば、候補の偶然的な起源が、共同体のその後の知識を偶然的なものにする必要はない。伝統的な分析において正当化の条件が担っていた役割の一つは、当て推量知識に数えられるのを排除することであり、ゲティア事例は、正当化された真なる信念でさえ、知識と両立しない仕方ですぐに運になることがありうると示唆する（Ichikawa and Steup 2026, §§1.3, 3）。どちらの場合も、問題になる運は、信念とその根拠あるいはその真理とのつながりのなかにあり、候補がどのようにして考慮されるに至ったかにあるのではない。これは、候補を生む経路と、それを受け入れることを根拠づける経路とは別でありうる、というよく知られた点である（第3章第5.1節、第5章第4節）。ここで偶然でないことが求められるのは、共同体の信念とその証明との関係である。

その場合、生成器は、知識を保有することも、よく組織された探究を遂行することも、安定した研究能力を顕在化することもなしに、知識の産出にとって因果的に重要でありうる。この事例は六つの述語を切り離す。すなわち、真なる内容を生成すること、それを知ること、他の主体の知識に貢献すること、

研究を遂行すること、研究能力を保有すること、そして認識の達成についての功績に値することである。大事なのは、これらの述語を区別しておくことである。内容が正しく、しかも検査済みである場合にも、真なる内容を生成することと貢献することとを分ける論者がいる。ウェザロール (Weatherall) とウルフソン (Wolfson) は、証明の認識的価値についての草稿で、形式的な正しさは証明が認識的価値をもつための必要条件でも十分条件でもない論じ、その価値を、成果物そのものではなく、数学者が証明を生み出し、確かめ、理解するようになる活動のうちに置く。そうであれば、「正しい証明をただ生成することは、それ自体では数学への貢献ではない」。ウェザロールとウルフソンは、数学者が述べ、確立したいと考えている定理の証明を計算機が助ける証明支援と、計算による証明生成とを区別し、機械が生成した証明は、どれほど健全であっても、二人が集団の活動として扱う数学に貢献するのは、人間の数学者がそれに取り組むことを促す限りにおいてである、とする (Weatherall and Wolfson 2026, 4, 13–15, 35–38)。デデオ (DeDeo) は、真理ではなく価値の判断について、一つの点を付け加える。ある定理が別の定理にとって重要だという判断は、結局のところ、前者が解消する行き詰まりを誰かが経験したことに支えられている。ところが自動化された方法は、そうした経験を機械的な過程の背後に隠してしまい、その経験に裏づけられない価値の代理指標だけを残す (DeDeo 2024, §2)。偶然の生成器は、真であることが偶然である内容について六つの述語を切り離した。社会的な活動としての数学について書かれたこれらの説では、正しく検査済みの結果を生成することと数学に貢献することも、切り離されうる。どちらも、その活動を人間の数学者たちのものとしており、ウェザロールとウルフソンはそれを人間による発見として描く。S が混合あるいは人工であれば、同じ切り離しが成り立つかどうかは、証明への取り組みが、そしてデデオの説では価値の判断を支える行き詰まりの経験が、その人工の成員において実現しうるかどうかにによる。どちらも、成員が人工である共同体は取り上げていない。研究者としての地位というより強い問いは、第7章第3.2節の帰属の分析に属する。

この事例は、作業的特徴づけがどう当てはまるかも示す。作業的特徴づけは、S にとって新しい認識の達成をもたらそうとする試みを研究とする (第1章第1節・第3節)。生成器の出力は共同体の後の知識に因果的に寄与したが、ある達成に因果的に寄与することは、それだけではその達成への組織された試みではない。そのため、出力した定理が後に証明されたからといって、生成器の動作が研究になるわけではない。生成器が寄与したかどうか、共同体の探究が研究であったかどうか、生じた知識を誰が担うかは、それぞれ別に帰属させる必要がある。

5. 検証は何に寄与するのか

研究実践は、答えを提案し、次いでそれを検証するものとして語られうる。認識論の多くの説は、知識を、真理と、主体と内容とのあいだの何らかの認識的に適切な関係とによって記述する。そこで、検証が正当化を生み出し、正当化が信念を知識へと転換する、という形で、二つの語彙を対応づけることもできそうに見える。本節は、この対応づけがどこまで成り立つかを考える。

そのために、検討を三つに分けて進める。一つは、活動である検証を、正当化をはじめとする、結果がもちうる地位から区別することである。検証がうまくいっても、結果がしかるべき地位をもつとはかぎらない。二つめは、それとは逆向きの食い違いを取り上げることである。結論が手元の証拠によく支持されていても、その証拠を集めた探究は、結果の地位とは別の基準に照らしてまざり行われていることがある。三つめは、節の問い、つまり検証が知識の帰属に何を寄与するのかに戻り、テストから可能な知識帰属に至る道筋を段階に分けることである。検証が直接に与えるのは最初の二つの段階、つまり結果と、その証拠の関連性である。それが態度を変えるには基礎づけの関係 (第4.3節) や更新の関係が要り、その態度が知識であるには採用する説が求めるもの (第4.1節・第4.2節) が要る。どの段階でも失敗はありうる。とくに結果が不利なとき、その結果がどの要素にかかわるのかを、仮説は補助的な仮定とともにしか試されないというデュエムの論点と、無効化の認識論によって見る。第5.1節は検証の多様さを、第5.2節は確認、つまり結果が何をどれだけ支持するのかを、第5.3節は限られた例として統計的方法を取り上げる。

本書は検証を広い意味で用いる。検証とは、探究者が候補となる答えを真なるものとして扱うべきかどうかを見定めようとする活動の一族を指す。この語は、決定的な確立を約束するために用いられているのではない。異なる学問分野や文献は、むしろテスト、確認 (confirmation)、検証 (corroboration)、証明、妥当性確認、監査、追試、批判について語り、それぞれ異なる対象と成功条件をもつ。重要な共通の機能は、不確実性の下での候補の認識的地位の評価であって、一つの普遍的な手続きではない。

検証は探究者が行うことである。これに対して、正当化、根拠づけ、信頼性、安全性、知識は、可能な地位や関係を名指す。ここでの根拠づけは、主張を受け入れることの認識的な支えを広く指す。信念の正当化はその一つの形だが、根拠づけは、結果や主張の受容についても言う。また根拠づけは、信念がその根拠に基づいて保持されているという第4.3節の基礎づけとは別の語である。明示的なテストは、正当化された信念にとって普遍的に必要なわけではない。知覚、記憶、証言、制度への依存は、認識的な支持を提供しうる。また、テストに合格することは十分でもない。テストは誤った量を標的にしたり、欠陥のある道具に依存したり、不適切なモデルを通じて分析されたり、その答えを漏洩させたり、それが支持する条件を越えて解釈されたりしうる。たとえば、探査機ガリレオの高利得アンテナは、大気中と真空中での展開試験に合格していたが、大気中ではピンのむき出しのチタンが酸素と反応し続けたために摩擦が展開を止めるほどには上らず、真空中の試験ではピンとソケットが互いに動いていなかった。ところが宇宙では、打ち上げ後に上段ロケットの振動がピンとソケットを真空中で互いに動かしたので、むき出しのチタンのピンがソケットにかじりつき、いくつかのリブを留めた。そのため、1991年に展開を指令されたアンテナは一部しか開かなかった。後の調査の報告は、試験の条件がこの問題を明らかにするのに十分でなかったと結論している (Johnson 1994, 375–377)。

検証と正当化の区別は、活動がうまくいっても、結果がしかるべき地位をもつとはかぎらないことを示した。証拠の集め方については、逆向きの食い違いも生じる。結論は利用可能な証拠に適合していながら、その証拠を生み出した探究が選択的であったり、早まって打ち切られたり、都合の悪い観察が現れにくいように組織されていたりすることがありうる。逆に、責任ある証拠探索が、利用可能な証拠が不十分なままであるために判断保留に終わることもありうる。証拠収集の規範についての現代の研究は、これらの通時的な問いを、ある態度が現時点で保有されている証拠に適合しているかどうかという共時的な問いに還元すべきではないと論じる (Flores and Woodard 2023)。この区別が研究にとって重要であるために、正当化のあらゆる理論にそれが組み込まれている必要はない。それは、よく支持された結果とよく遂行された探究とが、なぜ同じ達成ではないのかを示す。したがって人工システムにとって、検索方針、情報源の除外、測定の選択、能動的な情報探索、停止規則は、最終的な推論がその入力条件として非の打ちどころのないものであるときでさえ、認識的に重大な帰結をもち続けうる。

いま挙げたもののうち停止規則については、それが認知的に重大な帰結をもつかどうかそのものが争われている。実験をどの規則で終わらせたかは、集めたデータが何を示すかには影響しない、とする見方があるからである。ロメイン（Romeijn）はその両側を示している。標準的なバイズ推論が満たす尤度原理のもとでは、仮説についての推論にかかわる情報はすべて、尤度関数、つまりそれぞれの仮説のもとで観測されたデータが得られる確率に含まれるので、実験を終わらせた規則はデータの含意を変えない。頻度主義者がこの原理に抵抗するのは、検定の判定を正当化する誤り率が実験計画に依存し、そのため、第4.2節で述べた停止規則の問題のように、同じデータが、ある計画のもとでは棄却を根拠づけ、別の計画のもとでは根拠づけないからである（Romeijn 2025, §§3.2.2–3.2.3）。ただし、証拠収集についての論点は、この争いに決着をつけることを必要としない。停止規則が、収集されたデータの含意を変えないとしても、それは、どの証拠が収集されたのか、そして探究が責任をもって遂行されたかどうかに影響しうる。

探究を責任をもって遂行することがそれ自体認知的な基準であるかどうかは、さらに別の論争である（第4章第5節も参照）。とはいえ、以下のどの見解に照らして読んでも、よく支持された結果とよく遂行された探究との違いは残る。第4.1節でその証拠主義を紹介したフェルドマンとコニーは、責任をもって証拠を集めることにはある認知的な長所があると認めるが、その結果としての態度の正当化がそれに左右されることは否定する。二人がコーンブリス（Kornblith）から採る例では、同僚の決定的な反論を耳に入れることさえしない物理学者も、以前に正当化されており、その証拠が変わっていないなら、自分の理論を信じることに於いて正当化されたままである。批判に開かれているのはその物理学者の性格である（Feldman and Conee 1985, 21–22, 31–32）。ソースタッド（Thorstad）は、探究に独特に認知的な規範はそもそも存在しないとする。信念について独特に認知的な規範を支持する考慮は、活動である探究には引き継がれず、さらに証拠を集めるか、再確認するか、打ち切るかは、全体を考慮した理由によって決まり、真理や知識の見込みはその理由の一つである、と論じる（Thorstad 2022, §§3–5, 6.3, 7）。

カウピネン（Kauppinen）はさらに進む。ソースタッドが、認知的な考慮を、合わさって全体を考慮したうえでの「べき」になる理由のひとつに数えるのに対し、カウピネンは、問いを追求する理由はすべて実践的であり、問いに答えることが何をやる助けになるかについての理由から伝わるものであって、したがって探究の認知的規範は存在しないと論じる。そして、探究者が証拠の集め方について実際に認知的な批判に値する場合、その欠陥は、信念や行為についての標準的な認知的規範にあるとする。カウピネンがフローレスとウッダード（Flores and Woodard）から翻案する事例では、ある人が、実際には信頼できない一つの情報源を参照し、たまたま真の答えに至る。カウピネンの説では、その人は、探究の規範に訴えなくても認知的な批判に値する。その人はおそらく、行為についての認知的規範というものがあるとすれば、それに違反している。情報源の信頼性を知らず、それを信じることに正当化されてもいないのに、それを行為の理由として扱っているからである。ただしカウピネンは、欠陥がむしろ、証拠の支えのないその信念をもつこと自体にある可能性も認める。情報源の信頼性について、よいが誤解を招く証拠をもっている変形では、その人は批判に値しない（Kauppinen 2026, §§3–4）。

これらの見解のあいだで異なるのは、第二の評価、すなわち探究についての評価が認知的なものかどうか、したがって第9.1節の、責任ある証拠獲得を求める要求が認知的な要求なのか実践的な要求なのかである。認知的な要求であれば、探索や停止の決定は、結果の用途とは別に、真理や知識にかかわる基準で評価される。実践的な要求であれば、どこまで探し、いつ止めるのが適切かは、探究の目的や費用、何がかかっているかに照らして決まり、同じところで止めることも、ある目的にとっては適切で、別の目的にとってはそうでないことがありうる。探究者が人工システムであれば、その目的や費用はシステム自身のものとは限らない。システムの目標や予算を定めた者が決めていることもあれば、システムが自ら定めていることもある（第7章第4.3節）。誰の目的が探究をどこまで進めるかを左右するのかは、配分の問いの一部である（第10章）。カウピネンの見解のもとではそれは実践的な要求となるが、システムが証拠を集めるにあたって依拠する、情報源の信頼性についての支持のない仮定は、なお認知的な批判を受けうる。その批判が向かうのは、システム自身がそうした規範の主体であればシステムであり、そうでなければ、その仮定をシステムに組み込んだ者である。

探究全体の評価はひとまず脇に置き、節の問い、つまり検証が知識の帰属に何を寄与するのかに戻ろう。テストから可能な知識帰属への橋渡しは、その矢印が自動的であるかのように装うことなく表現することができる。

目標主張 H + 方法 M + 補助的コミットメント A

-> 結果 R

R + 背景知識と評価基準 N

-> H または別のコミットメントに対する証拠的関連性 E

E + 基礎づけまたは更新の関係 B

-> 変化した認知的態度 D

採用する知識の説のもとの D の地位

-> 可能な知識帰属

行ごとに読めば、この図式は四つの段階を分けている。第一の段階では、目標とする主張 H が、方法 M によって、テストが依拠する道具、条件、前処理などの補助的コミットメント A のもとで試され、結果 R が出る。たとえば、H をあるモデルがある能力をもつという主張、M をベンチマーク、A を、そのベンチマークがその能力を測っているという前提や、その問題がモデルの訓練データに混入していないという前提だとして。第二の段階では、結果は、背景知識と評価の基準 N、たとえばどれだけスコアなら意味があるか、ほかのモデルの成績はどうかに照らして読まれてはじめて、証拠になる。そのとき証拠は H にかかわり、混入が疑われれば、H ではなく A にかかわる。第三の段階では、証拠が態度を変えるのは、誰かのコミットメントが実際にその証拠のゆえに改まる場合だけであり、それを担うのが関係 B である。第4.3節の基礎づけの関係や、信念の度合いを証拠に応じて改める更新がそれにあたる。第四の段階では、変化した態度 D が知識であるかどうかは、採用する説がそれに何を求めるかによる（第4.1節・第4.2節）。第三と第四の段階が問うのは、知っていると言われる主体の態度である。システムに知識を帰属させるならシステム自身のコミットメントであり、人間や共同体がシステムを通じて知るようになるなら、その人間や共同体の信念である（第4.3節）。

最後の行は、すべての説に共通する条件から知識を組み立てるものではない。知識第一主義では、先行する態度とその根拠は、知識を定義するために組み合わせられるのではなく、知識との関係で評価される。また、最後の行が「可能な」知識帰属であるのは、知識を帰属させるのが評価者だからでもある。評価者が知識を帰属させる証拠を十分にもたないことは、それだけでは主体が知識を欠くことを示さない。この図式は評価すべき問いを位置づけるものであり、それらを決着させるアルゴリズムではない。

すべての矢印は、失敗の実質的な発生箇所である。上の探査機ガリレオの事例は第一と第二の矢印にかかわり、第5.2節は確証について、第5.3節は統計的方法について、第二の矢印に立ち戻る。以下は、結果が不利なときの第二と第三の矢印にかかわる。方法と補助仮定は結果を生み出すのを助けるが、結果はそれ自体の認識的な宛先を携えてはいない。外れた予測は、焦点となる仮説、境界条件、道具のモデル、前処理の選択、背景理論、あるいはそれらの何らかの組み合わせにかかわりうる。デュエム (Duhem) による物理学のテストの分析は、限定的だが決定的な論点を示している。すなわち、実験に直面するのは孤立した仮説ではなく、否定的結果はより大きな理論的配置を巻き込みはするが、それ自体では欠陥のある要素を同定しない (Ariew 2026)。後の過程の分析では、探究者がどこを修正するかをどのように決めるのかを検討する (第6章第2節と第4節)。ここでの論点は、検証の結果を主仮説の変化へと機械的に対応づけることはできない、ということである。

無効化 (defeat) の認識論は、信念がもっていた正当化を、新しい情報がどのように失わせうるかを扱う。それは、どの要素に欠陥があるのかを決めはしないが、反対の証拠が主張にかかわる仕方を区別することで、この論点をさらに細かくする。ポロック (Pollock) の区別では、反駁的 (rebutting) 証拠は、主張の否定を支持する。たとえば、効果がないとする慎重な追試は、ある薬が効くという主張を反駁する。これに対して掘り崩し的 (undercutting) 証拠は、主張の否定を支持することなく、主張とそのために提示された根拠とのあいだのふだんの結びつきが、その場合には成り立たないと疑う理由を与える。たとえば、サンプルが偏っていたと分かれば、そこから引き出された推論は掘り崩されるが、結論が偽であることまでは示されない (Koons 2025, §1.1)。高階の証拠は、主張に直接にはではなく、探究者とその証拠にかかわる。それは、探究者自身の有能さ、信頼性、あるいは証拠の状況を疑わしいものにしうる。たとえば、分析を行った者が、その分析が誤りを生みやすいことで知られる条件のもとで行われていたと知る場合である (Roush 2017, §3)。そうした証拠が一階の主張への確信をどれだけ手放すことを合理的に求めるのかは、争われている。ラウシュ (Roush) の報告では、フェルドマンとホロウィッツ (Horowitz) は、自分の証拠が p を支持しないと信じながら p を確信することは合理的でありえないとする。フェルドマンによれば、「 p 、しかし私の証拠は p を支持しない」という信念は、真であると同時に合理的であることができない。後半が前半の合理性を掘り崩すからである。またホロウィッツは、そうして確信を保つ主体は、自分が正しかったことを運がよかったとしか説明できないと論じる。これに対してウィリアムソン、ウェザーソン (Weatherson)、ウェッジウッド (Wedgwood) は、疑いが主体自身の有能さにかかわる場合には、そうした水準の分裂 (level-splitting) は合理的でありうると論じる。それぞれの説では、ある命題が証拠に照らして高い確率をもつと同時に、その確率が低い見込みも高いことがありうる。合理性の規範は、それに従わなかったことを示す証拠を主体がもつだけで適用されなくなるわけではない。そして、実際に成り立つ支持関係は、主体がその関係について正しく信じているかどうかにかかわらず、信念を正当化しうる (Roush 2017, §3)。表2.1の「無効化への感応性」の行で何が適切な応答とみなされるかは、それに応じて異なる。

デュエムの論点を人工研究者に移せば、看板となる仮説に付されたスカラー値の損失だけでは、不利な結果がどこを改める理由になるのかの診断を尽くせない、ということになる。道具のモデル、サンプリングの仮定、前処理の手順、背景的な規則性も、修正の候補になるからである。損失をスカラー値で表しながら、修正の候補を別に保持する人工の構成は、これと両立する。ここまでは、デュエムの論点だけで言える。無効化の区別は、そこに二つの修正先を加える。掘り崩し的な証拠は、結論の否定を支持することなく、主張とその根拠を結び支持関係を疑う理由を与える。高階の証拠は、評価した者、つまり探究者や妥当性確認者の有能さの見積もりを改めることを求めうる。また無効化は、結果が不利なときに限られない。合格したテストの後で、たとえばサンプルが偏っていたと分かったときにも起こる。ただし、これらの区別は、無効化された関係を突き止める手順までは与えない。それを突き止めるには、依然として領域知識、実験の技能、そして場合によってはさらなる介入が必要である。以前に証拠とされたものそれ自体も、修正を要しうる。ウィリアムソンは、ひとたび条件づけに用いられた命題がその後ずっと確率1を保つ標準的なベイズの更新では、のちの証拠がそれまで証拠であったものを掘り崩すことを表せないと論じ、証拠の記録を確率とは別に保ち、命題が証拠としての地位を失いうるようすることを提案する (Williamson 2000, 214, 219–220)。この議論に立てば、証拠を加えることしかできない更新をもつシステムは、そうした喪失を表せないことになる。

これらの区別がシステムに何を求めるかは、誰が検証を組み立て、不利な結果のときにどこを改めるかを誰が決めるのかによって異なる。ほかの者、つまり人間の研究者やほかの人工エージェントが設計したチェックをシステムが実行し、結果を読んで何を改めるかをその者たちが決めるなら、結果がどの要素にかかわるのかを見定める仕事はその者たちの側に残る。その者たちがその仕事をできるためには、システムの報告は、看板となる仮説についての合否だけでなく、何を、どの補助仮定のもとで試したのかを示す必要がある。その者たちが人工エージェントであれば、報告が示す必要のあるものは、それらがシステムからどれだけ独立しているか、たとえばそれらとシステムがひとつの基盤システムのインスタンスどうしではないかにまで及ぶ。批判者が検出を加えるのは、生産者と同じところで誤らない限りにおいてだからである (第11章第3節)。これに対して、何を試し、何ををもって合格とし、どこを改めるかをシステム自身が決めるなら、これらの区別はシステム自身が引き受ける課題になる。第6章第4.2節は、研究を行うのが機械である場合について、この課題を論じる。第7章第4.2節と第4.3節は、選抜、評価、訂正を誰が供給するのか、そしてシステムが供給された基準を適用するだけなのか、それを選び、作り、改めもするのかを、自律性の比較として区別する。この違いは、知識の主体が誰かという問いとは別の軸である。ほかの者がシステムを通じて知り、検証を組み立てたのはシステムである場合を考える。その者が結果の診断を自ら引き受けるときには、システムがこの課題をどう果たしたのかを報告から読み取れる必要がある。アクセス可能な理由を求める説のもとでは、受け手には自分で使える理由が要る。ただしその理由は、結果がどう確かめられたかについてのものでもよいし、証言についての還元主義の説のように、当の用途でシステムが信頼できることについてのものでもよい (第11章第5.2節)。システムの報告は、そうした理由の源の一つになる。これに対して、過程の信頼性で足りるとする説のもとでは、受け手は、検証の構造をすべて読めなくても知りうる。証言による信念の正しさをより十全には伝達の連鎖に座を占める有能さに負わせるソーサの説 (第4.2節) をこの場合に当てはめても、同じである。そのとき帰属を支えるのは、報告を読めることではなく、システムの検証の仕方が信頼できることについての証拠である。システム自身に知識を帰属させるなら、さらに、その作動しているコミットメントがそうした修正に応じて実際に改めることが求められる (第4.3節)。

以上をまとめれば、検証が知識の帰属に寄与する仕方は次のようになる。ある結果が特定の研究主張に対してもつ証拠的関連性は、測定とそれが測ろうとするものをとらえているかどうか、サンプリング、解釈、推論の目標を含む諸関係に依存する。利用可能な証拠が態度を変えるのは、何らかの実際の基礎づけ関係または更新関係を通じてのみである。その態度は、システムに知識を帰属させるならシステムの作動しているコミットメントであり、ほかの者がシステムを通じて知るようになるなら、その者の信念、その者が人工システムであればその作動しているコミットメントである。よく支持された態度が知識となるのは、それが真であり、かつ採用する説が課すいかなる追加条件をも満たす場合に限られる。

5.1 検証は多様だが、それで機能を失うわけではない

本節の初めでは、検証を、共通の機能をもつ多様な活動の一族とした。本小節は、その多様さが機能の共通性を損なわないこと、また共通の機能が手段ごとの違いを消さないことを示す。

検証の出来事は、証拠を生成し、証拠が主張を支持するかどうかをテストし、無効化要因を露呈させ、過程の信頼性を推定し、試されているものが本当に目標としていたものなのかを問い直し、あるいは局所的な結果を社会的批判に利用可能に示す。それは、焦点となる主張への確信を変えないまま、道具や方法への確信を大きく低下させることがありうる。また、評価されている主張そのものを作り変えることもありうる。

機能上の共通性は、認識的な効力の違いを消し去らない。類推、パイロット結果、ランダム化試験、独立の追試、演繹的証明、歴史的資料の比較、解釈的批判は、いずれも何が真剣に受けとめられるかを変えうる。それらは、何にかかわるのか、どの代替案を弁別するのか、その無効化可能性 (defeasibility)、そして提供しうる地位の種類において異なる。それらすべてを信念の更新と呼べば、共通する一つの効果は記述できても、それらの効果がなぜ異なるのかは見失われるだろう。人工研究者にとってこれは、工程に「検証」と名づけた段階があることや、その可否の表示があることが、それだけでは何が確かめられたのかを教えない、ということである。第6章第3.1節は、これらの違いのうち、問いのどの部分にかかわるか (かかわる先) と、それをどれだけ強く決着させるか (決着の強さ) を使って、チェックを比べる。

検証はまた、別個の段階として生成の後に続く必要もない。候補形成はすでに、先行する証拠、整合性、類推、予備的チェック、そして何が精査に耐えるかについての予期を用いている。数学的証明は、同じ対象のなかで結果を構成し、かつその根拠づけを提示しうる。批判は、内容を認定するのではなく、内容を変えうる。大まかに見れば、提案することと評価することは異なる役割である。より細かく見れば、それらは、何をどこまで確かとみなすかを少しずつ組み替えていく一つの流れのなかで入り混じっている。第5章第4節は、その時間的・実践的な構造を展開する。

5.2 確証：結果は何を、どれだけ支持するのか

図式の第二の矢印、つまり結果と背景の基準から証拠的関連性に至る関係は、確証理論 (confirmation theory) が扱ってきた関係である。証拠が仮説をどのように、どの程度支持し、あるいは仮説に不利にはたらくのか、という問いである。クルーピ (Crupi) は、自らが比較的標準的とする分類に従って、主な立場を三つの系統に分ける。一つは、ヘンペル (Hempel) の理論のような、事例による確証である。二つ目は仮説演繹法 (hypothetico-deductivism) である。これは、仮説が補助仮定とあわせて証拠を含意するとき、証拠は仮説を確証するとする立場である。それを強める案には、証拠があらかじめ予測されていたことを求めるもの (予測主義)、証拠が仮説を組み立てるのに使われていないことを求めるもの (使用新奇性)、仮説が偽であれば証拠が得られにくかったことを求めるもの (消去主義。厳しいテストへの訴えはその目印とされる)、仮説が簡潔で統一的な仕方では証拠を取り込むことを求めるもの (単純性)、仮説が証拠を説明することを求めるもの (説明主義) などがある。三つ目は、支持を確率で表すベイズ的確証理論である (Crupi 2025, §§1–3)。確証理論が扱うのは、証拠が仮説に与える支持という関係であって、活動としての検証ではない。検証の一族のうち、テストや追試のように、結果が仮説を支持するかどうかを見定めようとする活動は、この関係を問う。しかし一族には、証明や解釈的批判のように、その働きがこの関係には取まらないものも含まれる (第5.1節)。確証理論のこれらの立場のあいだの論争のいくつかは、本書の別の箇所に見れる。予測と事後的な当てはめ、厳しいテストは候補の評価のなかで (第5章第4.2節)、外れた予測の責任をどこに置くかについてのベイズ的な扱いと誤り統計的な扱いは第6章第2.3節で、チェックの独立性は第6章第3.2節と第11章第2節・第3節で、ベイズ統計と頻度主義統計の対比は第5.3節で扱う。本小節は、それらの議論が暗黙のままにしている一つの区別と、それに関係するいくつかの結果が人工システムによる研究に何を含意するかを取り上げる。

ベイズ的確証理論のなかで、クルーピは二つの確証の概念を区別する。確固さ (firmness) としての確証では、仮説は、証拠のもとでありそうであるほど確証されている。確率的関連性 (relevance) としての確証では、証拠は、仮説の確率を上げるかぎりでは仮説を確証する。ここでの「関連性」は、確率が上がるという意味であり、図式の第二の矢印の「証拠的関連性」とは別の語である。まれな病気についての検査が陽性になれば、その病気である確率は上がるが、それでも低いままでありうる。確率的関連性の意味ではその病気は確証されるが、確固さの意味では確証されない。クルーピは、この区別を「きわめて基本的」でありながら「ときに見落とされる」とするサモン (Salmon) の言葉を引き、この区別は、哲学だけでなく人工知能や推論の心理学でも、繰り返し強調し直さなければならなかったと述べている (Crupi 2025, §3.3)。クルーピの提示では、確率は信念の状態を表す。ここではそれを、主張を評価する者の確率として読むことができる。そのため、それを用いることは、人工システム自身が信念の度合いをもつことを前提にしない (第2節・第2.1節、第6章第1節)。

この区別は、人工システムによる研究にとってまず、合格したチェックが何を報告しているかにかかわる。ベイズ的な概念のもとでは、チェックに合格したことが示しうるのは、その結果が主張への支持を上げたということであり、そのあと主張がどれだけ支持されているかは、チェックの前にそれがどれだけありそうだったかにも依存する。クルーピ自身はこの場合を論じていないが、この区別を、候補を少ない費用で大量に生成するシステムに当てはめると、一つの帰結が出てくる。生成される候補の大半が偽であり、チェックの前に候補どうしを見分ける手がかりがなければ、一つひとつの候補はチェックの前にはありそうにない。そのため、偽の候補も一定の確率で合格するチェックでは、合格した候補に残る支持は、チェックの前からありそうな少数の候補を試す探究で同じように合格した場合よりも弱い。この帰結は、知っているとされるのが誰であっても成り立つ。システム自身が知っているとされるなら、合格がシステム自身のコミットメントをどれだけ支持するかにかかわる。ほかの者がシステムを通じて知るようになるなら、候補が合格したという報告はその者にとっての証拠の一部であり、どれだけの候補が生成され、そのうちどれだけが偽でありそうだったかを知らずに報告を読む者は、合格が主張

をどれだけ支持するのかを見誤りうる。消去主義や誤り統計の見方は、合格を別の仕方でも評価する。仮説が偽であったならその合格がどれだけ起こりにくかったか、によってである (Crupi 2025, §2.4)。どちらの見方でも、ある仮説がチェックに合格したという報告は、それだけでは、その合格が何をどれだけ支持するのかを示さない。そして、誤ったときの害が大きいほど受け入れに多くの支持が要するとする帰納的リスクからの論証によれば、その支持がある用途にとって十分かどうかは、さらに別の判断である (第11章第4節)。第6章第3.1節のチェックの決着の強さも、したがって一通りでなく報告される。

クルーピの概観にある、さらに四つの点が、そうした主張をどう確かめるかにかかわる。第一に、素朴な仮説演繹法では、仮説と矛盾しないどんな証拠についても、その証拠が仮説を確認することになる補助仮定が存在する。クルーピはこれをデュエムの論点と結びつけ、光の粒子説の論者の多くが、補助仮定を作り替えて波動説の論者たちが導いた事実を取り込んだものの、それを確認とは主張しなかったことを指摘している (Crupi 2025, §2.3)。これを、前処理、除外、モデルの構成要素をデータを見た後で選べるシステムに当てはめると、そうしたシステムは、その選択が制約されていない程度に応じて、素朴な仮説演繹法なら確認と数える一致を、しばしば作り出せることになる。予測主義と使用新奇性の見方では、そうした選択がデータを調べる前に固定されていたかどうか、確認を主張するために示すべきことの一部になる (第4.2節のあらかじめ特定された検定の過程と、第4.3節の来歴を参照)。消去主義の見方で示すべきなのはむしろ、そうした選択をしたうえで、仮説が偽であったならその一致が起こりにくかったかどうかである (第5章第4.2節)。ペイズ的な扱いでは、そのかわりに補助仮定が、それがどれだけありそうかによって重みづけられる (Crupi 2025, §3.5)。第二に、使用新奇性の見方では、仮説を組み立てる制約として用いたデータは、その仮説を確認することができない。クルーピは、シュルツ (Schurz) に従って交差検証をこの考えの統計における日常的な応用として挙げ、この見方自体が争われていることも報告している (Crupi 2025, §2.4)。この見方に立てば、システムが候補を生成したり調整したりしたベンチマークとの一致は、候補をほとんど、あるいはまったく確認しない。選抜から取り置いた経路での評価が重要になる理由の一つがここにある (第11章第8節)。第三に、確率的関連性の意味では、ある結果が仮説を確認するなら、その結果が得られないことは仮説を反確認する。そのため、どちらの結果も支持と数える評価、つまり「表なら私の勝ち、裏ならあなたの負け」型の判断は、この意味では一貫しない。二つの結果のうち仮説を確認するのは、高々一方だからである (Crupi 2025, §3.3)。したがって人工研究者の主張の評価は、どんな結果であればそれに不利と数えられたのかを問うことができる。第四に、確率的関連性の意味では、事前にありそうになかった証拠ほど、つまりしばしばより多様な証拠ほど、それを含意する仮説を強く確認する (Crupi 2025, §3.6)。一つの基盤システムを何度も実行したときの一致は、実行どうしが同じところで誤りうる限り、仮説が真であってもなくても起こりやすいので、この意味での多様な証拠には数えにくい (第11章第5.4節)。ワイスバーグとライスマン (Reisman) は、モデル化の結果が頑健でありうる三つの水準を区別する。モデルのパラメータの変更に耐えること、その数学的構造の変更に耐えること、そしてモデルが組み立てられている表象の枠組み全体の変更に耐えることである (Weisberg and Reisman 2008, 9–11, 15–17, 25)。実行やインスタンスが一つのモデル化の枠組みを共有しているなら、それらの一致が示するのはせいぜいその枠組みの内部での頑健性であり、枠組みを変えることで試される頑健性ではない。

5.3 限定的な例示としての統計的方法

第5.1節で見た検証の多様さは、歴史学と自然科学のような分野の違いにだけ現れるのではない。統計学という一つの方法の族の内部でも、何が結論を支えるとみなされるかは、方法ごとに異なる。大塚は、統計的な導出は、その確率が何を意味するかを定めなくとも数学として進めうるが、採用された解釈が、その結果が何を示すと有意味に言えるかを限定すると論じる (Otsuka 2023, 42–43, 176–178)。大塚は統計的推論を、存在論的な前提——データの背後にある自然の斉一性と、パラメトリックな推論においては選ばれた分布族——の上に立つものとして扱い、その諸学派を、正当化の概念において異なるものとして読む (Otsuka 2021, 3–7, 13)。

大塚の読みでは、ペイズ的な更新は、ある信念の度合いを、行為者のほかの信念の度合いからの妥当な推論によって正当化するので、内在主義に似ている。他方で、その事前分布を出発点として確保しようとする試みは、基礎づけ主義に似ている。基礎づけ主義とは、信念のなかには、それ以上の正当化を必要とせず、ほかのすべての信念を正当化する最終的な拠りどころとなる「基礎的な」信念がある、とする立場である。第4.3節の基礎づけ (basing) とは別の概念である (Otsuka 2023, 53–57, 61–66)。大塚の説明では、ネイマン＝ピアソン (Neyman–Pearson) の検定は、これとは異なり、検定手続きの信頼性を通じて判定を正当化しうる。ノージック (Nozick) の二つの追跡条件は、 p が真でなかったなら主体は p を信じなかっただろうこと、そして p が真であったなら主体は p を信じただろうことを求める。前者は第4.1節の感受性にあたる。大塚は、真の帰無仮説を棄却しない確率である信頼係数を、対立仮説について検定が前者の条件をどの程度満たすかの尺度として読み、偽の帰無仮説を棄却する確率である検出力を、後者の条件をどの程度満たすかの尺度として読む。そうして検定を信頼性主義に対応させる (Otsuka 2023, 88–94)。大塚は、赤池情報量規準によるモデル選択を、この規準がモデルをデータ生成過程への近さではなく期待される予測性能によって評価するので、プラグマティズム的なものと読む。また、深層学習モデルの現在の正当化を考える手がかりが、徳認識論 (virtue epistemology) にあるかもしれないと示唆する (Otsuka 2023, 118–124, 129–135)。

大塚は、これらの対応関係を、完全な類比ではなく単純で理想化されたモデルとして提示しており、この比較がペイズ統計学者を内在主義者にし、内在主義者をペイズ主義者にするということを否定している (Otsuka 2023, 55–56, 179–180, 182)。大塚はさらに、ペイズ統計学も頻度主義の統計学も、それが似ている一つの認識論の内部にとどまって満足することはできないと論じる。大塚の説明では、事後モデル検査は、事前分布、尤度、実験計画、測定についての仮定からなるネットワーク全体をデータに照らしてテストする。そのためペイズ的な実践は、行為者の信念の度合いのあいだの関係を越え、基礎づけ主義的な描像から離れていく (Otsuka 2023, 67–72)。大塚によれば、検定の判定を正当化する誤り率は、判定そのものが確かめられない分布と実験計画についての仮定に依存する。したがって、過程の信頼性に責任を負おうとする頻度主義者は、それらの仮定を別の手段で精査しなければならない (Otsuka 2023, 95–98, 106)。

より持続的な教訓は、形式的に正しい操作がその認識的役割を決定するわけではない、ということである。大塚が指摘するように、計算された事後分布や低い p 値は、それ自体では研究対象の現象について何も語らない (Otsuka 2023, 105)。さらに、頻度主義の見方では、同じデータが、ある停止規則のもとでは仮説の棄却を根拠づけ、別の停止規則のもとでは根拠づけられないことがありうる。尤度原理を擁護する側はこれを否定する (第4.2節と第5節)。事前分布、尤度、サンプリング計画、停止規則、損失関数、誤りの基準、測定モデル、目標の定義は、それぞれ、どのような結論が根拠づけられるかを変えうる。方法を選択する、あるいは適用する人工システム自身が結論を下すとされるなら、そのとき何を結論することが根拠づけられるかは、そのシステムが

その手続きを、問い、データ生成条件、仮定、そして誤りのコストへとどう関係づけるかに依存する。ほかの者がシステムの出力から結論を引き出し、その出力がその者の結論をどれだけ支えるのかを自ら評価するなら、システムが手続きをそれらにどう関係づけたのかが見えなければならない。過程の信頼性で足りるとする説のもとでは、受け手はそれが見えなくても知りうるが、帰属を支える評価者には、システムのそうした関係づけ方が信頼できることについての証拠が要る（第9.1節）。そのシステムが統御的な基準を批判し修正しうるかどうかは、認識的自己統治（第7章第4.3節）についてのさらなる問いであって、統計を用いるたびに求められる前提条件ではない。

6. 知識について社会的なのは何か——新しさの参照先、産出、保有、保持

第2節から第5節は、知っているとされる者の態度がどのようなものでなければならないかを問うた。研究において、知識はそれに加えて社会的でもあり、しかもその意味は一つではない。知識が新しいとされる参照先は社会的であり、知識の産出は集合的でありうる。知識の保有は、集合的であることも、そうでないこともありうる。本節はこれらの意味を分ける。まず参照先 *S* を取り上げ、次にファガンの区別によって産出と保有を分ける。そしてその区別を、保持を含む六つの関係へと広げ、社会的知識の説をそれらの関係に沿って比べる（表2.2）。そのうえで、これらの説が人工システムによる研究に何を含意するかを、*A* と *S* の組み合わせごとに見る。最後に、社会的なものについて強さの異なる三つの主張を区別し、知識が社会的であることが、真理を合意に委ねることと読まれないようにする。

変数 *S* は、第1章第6節が学習や再発見の例とともに述べた観察から生じた。知識の場合でいえば、研究が新しい知識を産出したと言われるとき、主張されている新しさは、一人の人物の私的な学習の新しさではない。そのため知識の主張については、*S* は、第1章第1節で導入した、互いに結びついた二つの役割をもつ。一つは、求められる知識が、実現したときに、それに相対的に新しいとされる参照先としての役割であり、もう一つは、その知識が誰のために求められるかを定める役割である。どちらの役割も、知識を *S* に帰属させるものではない。誰がその知識を知っているか、それが *S* の知識になったかどうか、なるとすればどのようにしてかは、以下で比較する説に委ねられる。また、どちらの役割も、*S* の認識的状況が求められる種類の知識を含みうることを前提にするが（第1章第2節）、社会がどのようにして知りうるのかを教えるものではない。*x* は *p* を知っている における個人変数に集団を代入しても、それが一人の巨大な人物になるわけではないし、集団の知識が成員の知識へと還元されるかどうかが決着するわけでもない。

日常的な言説で省略されている参照先は、人類でもありうるが、実際には、証言・教育・記録・器具・標準・制度からなるより広いネットワークのなかに位置する、関連する専門家共同体である。社会の以前の認識的状況をどう捉えるにせよ（表2.2）、すでに知られていることは社会ごとに異なるから、新しさの主張は、その参照先が明示されているときにはじめて評価できる。ある命題は、ある研究室にとって新しく別の研究室には周知であることもあり、ある分野にとって新しく隣接する学問分野ではすでに確立されていることもあり、人間にとって新しく人工の共同体には周知であることもあり、ある人工社会にとって新しく別の人工社会にとってはそうでないこともありうる。その違いは新しさだけにかかわるとは限らない。たとえばハヴスタッド（Havstad）は、*Laboratory Life*（『ラボラトリー・ライフ』）の民族誌的研究を同じ研究所で再現する研究のなかで、自らが調査した研究室による総説の一文——標的遺伝子のネットワークが生理的応答を規定するという文——は、その調査の時点で、その研究室の内部ではほぼ確立された事実として、競合する研究室からは争われているものとして、そしてその分野の外の科学者からは思弁的なもの、あるいは誤ったものとして扱われるだろうと示唆している（Havstad 2020, 25–26）。これらの集団のあいだで異なるのは、その文に与えられる地位であり、本章はその地位を知識から区別する。

残る二つの意味、産出と保有は、ファガン（Fagan）が区別している。ファガンは、単純な定式ではともすれば混同されかねない二つのテーゼを分ける。第一は、集団が科学的知識を産出するために必要でありうるというテーゼである。ファガンはその点を三つ挙げる。探究の多くの場合に、方法論上の規範を満たすこと、理論的概念を発展させること、そして探究の結果を科学的知識として妥当性確認することに、集団が必要となる（Fagan 2012, abstract）。これは探究の多くの場合についての主張であって、すべてについての主張ではない。第二の、さらに踏み込んだテーゼは、集団がそれ自体として科学的知識を保有するというものである。ファガンは、このテーゼへの支持は「より両義的」だとする。それを支持するとされる事例は、「成員の信念の関係的な複合体」によっても説明できるからである（Fagan 2012, abstract）。ロンジーノの報告によれば、ファガンはバードの説に異を唱えて関係的な見方をとる。この見方では、共有された知識は、個人を超えて存在する社会的な実体がもつ知識ではなく、個人のあいだの認識的な関係の関数である（Longino 2025, §5）。二つのテーゼのこの区別は、集団心への即座のコミットメントを要求することなく、知識が社会的であるという主張に分析的な力を与える。

受容は、二つのテーゼが分かれる一つの点である。共同体による受容は、結果を妥当性確認するものでありうるので、ファガンの三つ目の点で産出に属しうる。それが、共同体がその結果を還元できない集合的な信念として保有することまで示すかどうかは、別の問いである。ドゥイグ・ウユグン・トゥンチ（Duygu Uygun Tunç）は、研究の協働が出版や発表のかたちで結果を公に表明することは、その証拠を生み出した分散的な過程とは別のものだと言っている。同じ研究の過程を、結果を表明するための独裁的な手続きとも、多数決の手続きとも、合意にもとづく手続きとも組み合わせることができる。そしてウユグン・トゥンチの報告では、CERN では出版の決定は専任の出版委員会が下し、成員の信念の関数にはなっていない（Uygun Tunç 2023, 10 and n. 18）。ウユグン・トゥンチはまた、自らの議論では前提としないものの、集団は信念をもちえないが、命題を集団の見解として受容することはできるという立場にも触れる。その根拠は、信念は自動的かつ非自発的に形成されるが、集団の見解はそうではないという点にある（Uygun Tunç 2023, 10 n. 16）。この根拠は、第3章第6節がエルギンから取り上げる区別と合う。その区別では、コーエンの意味での信念はおおむね非自発的で傍観的であり、受容は自発的で行為者的である（Elgin 2025, 11, 35–36）。エルギンは、認識的な共同体を、自らが拠って立つ方法と基準を考案し、集合的に是認するものとして描き、成員の大半が自律的に働いていなければ、単なる合意はほとんど何も加えないとする（Elgin 2025, 121–122）。ウユグン・トゥンチの事例は、出版の決定を成員の信念から分けている。ファガンの二つのテーゼに当てはめれば、この区別だけでは、公の肯定が集合的な保有を確立するかどうかは決まらない。

知識を産出する生態系は、一致することも分離することもありうるいくつかの関係を含んでいる。それらは知識の獲得を構成し、可能にし、伝達し、維持し、あるいは促進しうる。しかし、それらは、**産出**（production）と呼ばれるひとつの出来事をそれぞれ別の意味で言い表したものではない。

- 候補となる内容とその根拠が局所的に生成される。
- 書かれた主張、形式的証明、データセット、数理的あるいは物理的なモデル、実演など、他者が点検できる形式で、結果が明確に表現される。

- ・一つ以上の主体が、関連する態度と地位を獲得する。
- ・資格ある参加者が結果を批判し、あるいは妥当性確認する（第11章が批判と認定として検討する関係）。
- ・結果が集団の現に働いている推論・意思決定・行為の構造に入る。
- ・結果が、検索・無効化・訂正・撤回が可能な形式で保持される。

本書は、ファガンが引く、知識を産出することと保有することとの区別を広げて、この六つの関係を分けている。これらは分析上の次元であって、必須のパイプラインではない。批判は公開前に起こりうる。主体は、他者が点検できる形式で結果が明確に表現される前に、その結果に対する関連する態度を獲得しうる。人工システムの内部状態がそうした態度を実現しうるかどうかは、採用する信念の説による（第2.1節）。暗黙的でありうる理解は、第3章第6節で取り上げる。また、結果が点検可能であっても、対立する理論のあいだを判定するものとは限らない。たとえば、電子の質量が速度によってどう変わるかの測定は、当時、電子の対立するモデルと、それらに結びついていると見なされたより広い理論とのあいだを判定するものとされていた。しかしローレンツ（Lorentz）の理論とアインシュタイン（Einstein）の理論は、この変化についての予測を含め、経験的な予測ではすべて一致しており、1908年に示されたデータが支持するように見えた予測も、その時点ではローレンツだけでなくアインシュタインの予測でもあった（Janssen and Mecklenburg 2004, 2, 31）。両者は基礎において異なりながらその予測を共有していたので、その予測を支持する測定は両者のあいだを決めることができなかった。ある歴史家の見方では、アインシュタインの理論と、ローレンツおよびポアンカレ（Poincaré）の理論との本質的な違いは、時間と空間の概念の変更にまで進んだかどうかであり、1910年前後までは、ローレンツ-ポアンカレの理論がなお広く用いられていた。エーテルを捨てるべきかどうかが物理学者のあいだで問題になったのも、その歴史家の推測では1910年よりのちのことである（Hirosige 1964, 107）。器具やプロトコルは、広範な信念が形成される前に結果を実践へと統合しうる。アーカイブは、かつてそれを理解していた人々が亡くなった後も結果を保持しうる。出版は結果を社会的に不活性なままにしておくこともある。たとえば、1982年に南極の昭和基地で測定された異常に低いオゾン量は、1984年の報告書に印刷され、そこでは1966年以来の最低値が四つの要約事項の二番目に挙げられていた。観測者は後に、その結果を日本国内と国際シンポジウムで発表しても特に反応はなかったと回想しているが、そのシンポジウムの印刷された論文集は、その観測者がオゾンの減少を最初に報告したと認められる根拠となった（Chubachi 1984, 13, 18; Chubachi 1993, 817）。

異なる説は、知識を異なる地点に割り当てるか、あるいはそれらの地点がひとつの尺度をなすことを否定する。表2.2は、こうした説を八つの系統に分けて示す。中央の列は、各系統において何が社会的知識を担い、どの関係を通じてそれを保有するかを示し、右の列は、人工システムが産出した結果にとって何が帰結するかを、*S* が人間であっても人工であっても成り立つ形で示す。二つの場合の違いは、各説の紹介の後の段落で分ける。ただし、この表は各説を厳密に述べたものではなく、見通しを得るために単純化したものである。各行は、表の後の段落で名を挙げる一つの代表的な説によってその系統を示している。また、二つの行は互いに異なる立場をまとめている。成員説・実働成員説の行は、十分な数の成員を数える説と、該当する領域で意思決定の権限をもつ実働成員だけを数える説の両方を含み、集団態度説の行は、集団の信念と集団の受容の両方を含む。

表2.2 社会的知識を担いうるものと、その保有の関係（単純化した見直し）

説の系統	担い手の候補と保有関係	人工システムによる結果にとっての帰結
個人保有説	産出が分散していても、適切な位置にある個人が知っている	結果が <i>S</i> の知識になるには、証拠または正当化された依存を通じて、 <i>S</i> の個々の知者に届かなければならない
成員説・実働成員説	十分な数の成員、あるいは該当する領域で意思決定の権限をもつ成員が関連する態度をもち、その根拠のあいだの整合性のような集団レベルの条件に服する	<i>S</i> の適切な成員へ届けることが社会的獲得の成功の一部となり、集団内の別のところにある反証がそれを無効化しうる
関係説	集団の信念は、成員の信念と、そのあいだの認知的な関係とからなる体系である。集団は産出に必要でありうるが、それによってそれ自体として知識を保有するわけではない	結果は、単に寄託されるだけでなく、成員のあいだの認知的な関係を通じて成員に結びつけられなければならない
集団態度説	構造化された集団が、集団レベルの規範のもとで信念または受容を形成する	集約、権限付与、整合性、責任が特定されなければならない、受け入れられた立場が信念であるかどうかそのものが争われている
分散的知識なき分散認知説	より大きなシステムが認知的作業を遂行するが、知識は個人に帰属可能なままである	記録や計算機の構成要素は、知者になることなく産出に加わりうる。研究を行う人工システム自身が知りうるかどうかは、この説では決まらない
知る主体なき知識説	科学的知識は、個人の知る者や集合的な知る者ではなく、信頼できる過程の公的に認定された結果に帰属される	記録の地位は、誰かの現在の態度ではなく、認定と公的な受容にかかっている
分散説・集団機能説	適切に組織された集合体が、知ることに結びついた認知的機能を実現する	成員、記録、検索、行為が協働して社会的な知る者を実現しうる
批判的文脈説	受け入れられた内容は、対象に合致し、特定の社会的条件のもとで批判を生き延びたとき、その共同体にとっての知識となる	可視性だけでなく、公共的な場、批判の受容、標準、権威関係が重要となる

ここまでと以下で論じる説が、表の行を順に代表する。すなわち、ウエグン・トウチの個人保有、ラッキー（Lackey）の実働成員説、ファガンが記述する関係的な見方、ギルバート（Gilbert）が定式化する共同受容説、分散的な知ることを伴わないギアリー（Giere）の分散認知、ポパー（Popper）の知る主体なき認識論のギアリーによる翻案、バードの機能主義、ロンジーノ（Longino）の批判的文脈の経験主義である。互いに比べると、これらの説は、上に挙げた六つの関係のどれを社会的知識にとって決定的とみなすかによっても分かれる。第1章第2節の言葉でいえば、結果が*S*の認知的状況に入ったと言えるのはどの関係が成り立ったときかについて、説が分かれるということである。たとえば、個人保有説と成員説は、主体による態度の獲得を最も重く見る。批判的文脈説は成員による受容とあわせて批判を、知る主体なき知識のギアリーによる翻案は認定と公的な受容を、分散説・集団機能説は結果が現に働いている推論と行為の構造に入ることとその保持を、それぞれ最も重く見る。

ウユグン・トゥンチは、分散した科学システムが集合的に証拠と知識を産出しう一方、知識を保有する主体は個人のままでありうると論じている。その説では、共同研究が集合的に成し遂げるのは、ある命題を支持する証拠を確立することである。それは信念を形成する過程ではないので、集団の信念を生まない。また、個人に保有を認めない説をとれば、最も成功した共同研究からさえ誰も学ばないことになってしまう (Uygun Tunç 2023, 2, 9)。ギアリーも同様に、分散認知が分散的な知ることを含意する必要はないと論じている。信念、知識、行為者性の帰属は、分散認知システムの人間の構成要素に限定されるべきだというのである。ギアリーは、そうした帰属をシステム全体に広げることを高次の解釈とみなし、それが正当化されるのは何らかの理論上の利点がある場合に限られるとしたうえで、その利点はないと判断する。ハッブル望遠鏡のシステムについて、システム全体が何かを見つけると予期したのか、見つけたと信じたのかと問うても、そうした問いは「あまり意味をなさない」。これに対して、帰属を人間に限れば、歴史家から人類学者まで、科学を研究するすべての者が共有できる理論的枠組みが得られる (Giere 2007, 315–319)。これらの見解は、集成的産出を、システム全体が知っていることの否定と両立させる。ギアリーの立場はさらに別の選択肢も指し示す。知る主体なき認識論というボパーの考えを翻案して、ギアリーは、科学という制度が科学的知識を産出するのであって、それを知る集合的行為者を指定する必要はないとする。個人がその結論を知っていると主張できるのは、その結論が信頼でき受け入れられた手続きによって認定されたことを知っているからである (Giere 2007, 318)。ボパー自身の定式はより簡素である。ボパーによれば、主観的な意味での知識は心的な意向や思考過程からなるのに対し、客観的な意味での知識は、思考過程ではなく思考内容、すなわち言語で定式化された理論の内容からなる (Popper 1978, §X)。認定と公的な受容という条件はギアリーのものである。そこでウユグン・トゥンチは、ギアリーの説明を主体なき説明に分類し、それを退けて個人による保有を支持する。その理由として、客観的な意味での知識という概念は、知識がどこにあるのかについて何も語らないことを挙げる。物理的宇宙論を理解する者が一人もいなくなれば、出版物が保存されていても、誰かがそれを読むまでは、宇宙が膨張していることは知られていないことになる、というのである (Uygun Tunç 2023, 12–13)。

ギアリーが応答している説のひとつは、クノール＝セティナ (Knorr Cetina) のものである。ウユグン・トゥンチがこのやりとりをまとめるところでは、二人は、協働の実験で知識がどのように産出されるかを、分散認知として同じように記述するが、そうして産出された知識の認識主体を誰とするかで大きく分かれる (Uygun Tunç 2021, 3–4)。クノール＝セティナは CERN の衝突型加速器の実験を分子生物学の研究室と比べ、実験が認識的主体 (epistemic subject) としての個人を消去すると論じる。ここで認識的主体を、それを通じて知識が調達される存在と定義している。論文はすべての成員の名をアルファベット順に載せて公表され、知識を生み出す能力は実験に移る。クノール＝セティナの論では、共同研究が自らと交わす会話が、個人の主体を集散的で分散した主体に置き換える。ただし、生み出されるもののすべてを知っている成員はいない。対照的に、分子生物学の研究室では、著者の慣習と課題の割り当てが個人の貢献を区別し続け、個人を認識的主体として保つ (Knorr Cetina 1999, 236–237, 249–250, 303–304)。クノール＝セティナは、実験が企業のような行為者として振る舞うことを否定し、個人と研究所が実験の内部でも存続し、それが最もはっきり見えるのは新しい実験が形成される時期だと報告する (Knorr Cetina 1999, 268, 271–274)。同じまとめによれば、ギアリーはクノール＝セティナを、実験そのものを知る者として指定していると読み、そうした認識的行為者性の拡張を、存在論的なコミットメントが過大だとして退ける (Uygun Tunç 2021, 3)。しかし、クノール＝セティナは認識的主体を知識の調達によって定義するので、そのテーゼは、ファガンの二つのテーゼのうち一つ目、すなわち集団が知識の産出に必要でありうるというテーゼに最も直接にかかわり、保有の問いを未決のまま残すものと読める。

バードは、保有についての集成的な説を展開している。自然化された知識第一主義的機能主義によれば、知ることは信じることの一種ではなく、適切に機能している認知システムが生み出す状態である。ここで認知システムとは、行為を導き、推論や記憶に引き継がれる事象的な表象状態を生み出すことを機能とするシステムであり、機能とは、自然選択によるものであれ人工物としてのものであれ、そのために選択されてきた働きを指す。この種の機能は異なる素材や配置によって実現されるので、集団も、適切に組織された認知システムを実現することによって知りうる。しかもその説明は、帰属を命題ごと

に個々の成員が知っていることへと還元しない (Bird 2024, §§2–3)。

共同受容説は、ギルバートがその代表的な定式を与えているものであり、ジェフ・ダン (Jeff Dunn) がラッキーの論じる形をまとめるところでは、成員たちが p を共同で受け入れることに同意しているとき、そしてそのときに限り、集団は p と信じているとする (Dunn 2021)。ダンの読みでは、ラッキーは、成員が信じていること、正当に信じていること、知っていることを単に足し合わせる見解も、集団の信念、正当化、知識が成員のそれから切り離されて浮遊することを認める見解も、ともに退ける。共同受容説に対しては、ラッキーは、それでは集団の嘘を説明できないと論じる。たばこ会社の取締役たちが、喫煙は有害だと信じながら、安全だと共同で受け入れ、公表するとしよう。会社は嘘をついているように見える。しかし、取締役会の共同受容だけで会社が喫煙は安全だと信じていることになるなら、この説では会社の主張を嘘と呼べなくなる。ダンが伝えるラッキーの診断では、共同受容説は集団の信念をあまりに意のままになるものにし、そのため嘘をついているという非難をあまりに容易に逃れさせる (Dunn 2021)。ラッキーはまた、自ら根拠の脆さ (base fragility) と呼ぶものをもつ集団の信念も退ける。それは、成員のかかなりの部分の信念の根拠が、別のかかなりの部分の信念の根拠と衝突している状態である。ラッキーによれば、そのような状態は、認識的に評価できるとは限らない。また、異なる理由が異なる計画を求めるので、集団の今後の行為についての熟議に組み込むこともできない (Dunn 2021)。ダンが引くラッキーの説明では、集団が p と信じるのは、その実働成員のかかなりの割合が p と信じ、しかもその信念の根拠を合わせると実質的に非整合でない信念の集合が得られる場合であり、またその場合に限られる (Dunn 2021)。ここで実働成員とは、カレストルブがトゥオメラ (Tuomela) から受け継ぐ意味では、意思決定の権限をもち、それゆえ集団の状態の内容に責任を負う成員である (Kallestrup 2024, 571)。ジェシカ・ブラウン (Jessica Brown) が引くラッキーの集団の正当な信念の説明では、集団が p と正当に信じるのは、次の二つが満たされる場合であり、またその場合に限られる。一つは、その実働成員のかかなりの割合が、合わせると整合的な信念の集合が得られる根拠にもとづいて p と正当に信じていることである。もう一つは、 p に関連する証拠を完全に開示し、成員の個人としての、また集団としての認識的な規範的要求に従ってその証拠について合理的に熟議しても、それらの根拠に加えると p を十分に確からしくしない信念の集合が得られるような、さらなる証拠が出てこないことである (Brown 2023, 2)。したがってこの説では、該当する領域に実働成員をもたない集合体、たとえば人類には集団の信念はなく、そのように特定された S は、結果がそれに相対的に新しいとされる参照先の役割を果たす一方で、保有を担うのはその内部の成員、下位集団、組織である。第1章第2節の前提、つまり S の認識的状況がその知識を含みうることは、この場合、それらの成員、下位集団、組織を通じて満たされる。科学の一分野がこの意味での実働成員をもつかどうかは、ここでは決めない (第7節と比較)。集団の主張についてのラッキーの説は、集団の信念についての説とは異なり、集団の権限によって語る代弁者に依拠する (第11章第5.1節)。

カレストルプ (Kallestrup) はもうひとつの中間的立場をとる。ある種の集団的知識は成員の知識へと還元できないが、それでも個人的特徴に錨を下ろしたままである。カレストルプの説では、集団の規約は、成員が個人としては従わない認識的標準を定めうる。たとえば、伝聞証拠を考慮しないよう指示された刑事陪審は、被告人が有罪であることを知りうる。これに対して陪審員は、信頼できる情報源からの、しかし誤った反対の伝聞をそれぞれ得ているので、誰もそれを知らない。それでも陪審の知識は、成員の状態や過程と、集団の構造のなかで成員が互いに立つ関係とによって決まっている (Kallestrup 2024, 578–579, 581–582)。

ロンジーノの批判的文脈的経験主義は (Longino 2025, §5)、社会的なものを、ここまでの説とは別のところに置く。ここまでの説が主に問うのは、誰が知識を保有するのかである。これに対してロンジーノは、社会の組織が、主張を受け入れることを支えるもの、つまり根拠づけそのものに入り込むと論じる。証拠としての関係は背景仮定に媒介されるので、科学の共同体の成員どうしの、あるいは異なる共同体のあいだの批判的な相互作用だけが、そうした仮定の受け入れを主観的な選好が恣意的に左右することへの歯止めになると論じる (Longino 2025, §5)。ロンジーノの実質的な説明のもとでは、批判の場、批判の受容 (uptake)、公共的に承認された標準、そして知的権威の調整された平等 (どの観点も共同体の批判的な相互作用に寄与しうる一応の能力をもち、その対等な地位は、批判にかかわらないことや批判に応じないことで失われうることは、すでに完了した私的達成を単に伝達するのではなく、科学的客観性を構成するのに寄与する (Longino 2025, §5)。この意味での批判の受容は共同体に課される要求であり、共同体のなかでの信念の分布は、その内部で行われる批判的な議論に感応するかたちで、時間とともに変わらなければならない (Longino 2025, §5)。ドゥティル・ノヴァエスとデデ (Dutilh Novaes and Dede) が伝えるところでは、ロンジーノはこれらの条件を、現実の共同体が程度の差をもって満たす、理想化された共同体の特徴として示す (Dutilh Novaes and Dede 2025, §2)。ここではこれらを、あらゆる知識にとっての理論中立的な条件のリストとして採るのではない。これらは、社会的組織化が根拠づけそのものに入り込みうる重要なあり方のひとつを示している。

ロンジーノは、誰が知識を保有するのかも論じている。レイ (Wray) が伝える初期の説では、知る者は、個人を超えた集団ではなく、相互依存する個人である (Wray 1998, §IV)。エイギによれば、ロンジーノの合致 (conformation) は、評価する観点によって異なり、程度をもつ成功の関係であり、真理はその極限の場合である (Eigi 2016, 49)。エイギの説明では、知識は共同体とその企てに指標づけられるが、受容が恣意的になるわけではない。共同体は実効的な批判の条件を満たさなければならないからである (Eigi 2016, 47–49, 138)。本章が根拠づけについてのロンジーノの社会的条件に依拠するところでも、知識の事実的な用法は保持する。

これらの説が人工システムによる研究に何を含意するかは、研究を行う行為者 A と、知識がそのために求められる社会 S が何であるかによる。ここでは二つの組み合わせを取り上げる。 A が人工のもので S が人間である場合と、 A も S も人工のものである場合である。混合社会はそのあいだに位置する (第8章第1.2節)。 A が人工のもので S が人間の共同体である場合、説は、 A の結果が S の知識になるために何が必要かについて分かれる。証拠や正当化された依存を通じて S の個人の知る者に届くこと、その実働成員に届けられること、共同で受容されること、受け入れられた手続きによって認定されること、組織された認知システムに統合されること、公共的な場での批判を生き延びて受け止められること、である。論者が述べる形では、どの説のもとでも、 A が出力を生成して公開するだけでは足りず、何を加える必要があるかは説ごとに異なる (バードの説を広げたとき、妥当性確認を経て通常の経路で利用できる記録で足りうるかどうかは、第7節で見る)。これは、第1章第2節の三つの判断のうち三つ目、結果が S の認識的状況に入ったかにかかわる。説はまた、 A がその知識の担い手に含まれうるかどうかについても分かれる。こちらは二つ目の判断、達成がどこで実現したかにかかわる。ウユグン・トゥンチとギアリーの見方では、人間と機械からなる組織が全体として知ることはなく、知識は個人に置かれる。マシュー・J・ブラウン (Matthew J. Brown) とウユグン・トゥンチが伝えるところでは、ギアリーは、信念、知識、行為者性の帰属を人間の構成要素に限るが、それは、ハッブル望遠鏡のように、人間が器具や計算機を用いて研究を行うシステムについて、理論的な選択として擁護されたものである (M. J. Brown 2015; Uygun Tunç 2021, 3–4)。どちらの見方も、それ自体が研究を行う人工システムである A を検討していないので、そうした A が個人として知る者になりうるかどうかは、これらの見方では決まらない。それは第2節の信念の説にかかっている。クノール＝セティナのテーゼは、産出についてのテーゼとして読めば、どの成員も産出されたもののすべてを知らないほど産出が協働のなかに分散しうることを示すが、誰がその結果を保有するのかは決めない。バードの集団機能説では、記録や検索も、知る集合体として適切に組織されたものの一部になりうるので、問いは、 A がそうした集合体の構成要素として組織されているかどうかになる。 A が、 S が保有するようになる知識に貢献するかどうかと、 A 自身が知る者であるか、その一部であるかどうかは、別の問いである (第2.1節、第1章第2節)。この場合、これらの説のどれも、 A 自身が知る者であることを求めない (第8章第1.1節)。ただし、別の関係を通じてそれが問題になることはある。システムの報告が、人間がそれを通じて知るようになる主張とみなされるなら、主張と証言の説は話し手としてのシステムに条件を課し、ウィリアムソンの知識の規則のもとでは、その規則がかかわる知識はシステム、あるいはそれが代弁する集団のものである (第11章第5.1節)。また、求められる達成が技能のようにシステム自身のうちに実現するものであれば、システムがその最初の担い手である。

S それ自体が人工社会である場合には、第二の問い、すなわち A 自身が知る者であるかという問いは、より広い形をとって、脇に置けなくなる。すなわち、人工の成員が、 A がその一つであれば A も含めて、そもそも知る者でありうるか、という問いである。 S の二つの役割はどちらも、 S の認識的状況が求められる種類の知識を含みうることを前提にしており (第1章第2節)、ここではその状況は人工の成員、その組織、その記録からなる。そのためこの問いは、 S のどの成員もまだ結果を知らない段階でも、また A が S の成員であるかどうかにかかわらず生じる。そのうえで、結果が S の知識になったかどうかは、担い手を要する説では、その担い手が人工の成員かその組織のなかに見いだされるかどうかにかかり、知る主体なき知識のギアリーによる翻案では、 S の内部での認定にかかる。個人保有説、成員説、関係説では、問いは、人工の成員がその結果を信じているかどうかに戻る (第2節)。関係説ではさらに、成員のあいだに認識的な関係が成り立っているかどうかとも問われる。共同受容説では、問いは、人工の成員が共同受容に加わりうるかどうかになる。バードの説では、社会をなす人工の成員が全体として適切に組織された認知システムを実現しているかどうかになる。知る主体なき知識のギアリーによる翻案では、人工の成員による認定が認定に数えられるかどうかになり、成員がひとつの基盤システムのインスタンスであるときには、それがとりわけ問われる。同じ基盤のインスタンスどうしは同じところで誤りうるので、互いの認定が独立の点検を加えないかもしれないからである (第11章第3節)。ロンジーノの説では、それらの相互作用が、知る共同体に必要な多様性と公共的な批判の基準を伴っているかどうかになり、複数のインスタンスや多数決はそれだけではそれを与えない。ギアリーが行為者性を人間に限ることを前提として受け取れば、そうした社会には知識を帰属させうる成員がいないことに

なる。しかしギアリーはそれを、人間が研究を行うシステムについての理論的な選択として擁護しているので、この場合を決着させるものではない。第8章第1.3節がこれらの違いを展開する。

最後に、ここまでの説が述べる知識の産出の社会的な依存は、真理が合意によって創出されることを含意しない。両者を混同しないために、本書は、強さにおいて異なる少なくとも三つの主張を区別する。

1. 探究は、社会的に継承された言語、概念、器具、証言、労働、記録に因果的に依存している。
2. 証拠的地位、客観性、あるいは知識は、組織された社会的実践によって部分的に構成される。
3. 共同体が受け入れるものは何であれ、それによって真となる。

個人主義的認識論も第一の主張を受け入れうる。ロンジーノの批判的文脈説のように、社会的な組織が根拠づけや保有に入るとする表2.2の説は、第二の主張の諸形態を擁護する。どちらも第三の主張を含意せず、本章で論じたどの説も第三の主張をとらない。社会的承認は、偽なる命題を真にすることなしに、根拠づけと保有に影響を与えうる。証拠、失敗した介入、抵抗する素材、その他の世界を巻き込む制約は、引き続き重要である。人工システムによる研究にとって、この区別は、人工の成員のあいだの合意やそれらにわたる多数決を、成員の作りにかわらず、結果を真にするものと読むことを防ぐ。第二の主張について問われるのは、人工社会の組織された実践が根拠づけを部分的に構成しうるかかどうかであり、第8章第1.3節と第11章第3節がこれを取り上げる。

S を選ぶことは新しさの相対性を決着させるが、保有を決着させはしない。作業的特徴づけは、保有を本節で比べた説に委ねている。そのため、結果が S の知識になったという主張は、自らの依拠する説明を十分に明らかにしているときにはじめて評価できる。第8章第1節は、 A と S を分け、担い手を成功の側で問ひ、 S が人間か、混合か、人工かで場合を分けるときに、 S の役割に立ち戻る。本章では、第7節が本節で比べた説をひとつの事例で試し、第9節が要求としてまとめる。第11章は、それらを認定と批判のなかで取り上げる。

7. 読まれていない、人工システムによる発見

本節は、ここまでの説を、ひとつの事例でまとめて試す。結果の産出と、 S による保有とが離れる事例である。結果を産出するのは人工の A であり、人工システムの信念についての説によつては、 A 自身がその結果を知っているかもしれない（第2節）。 S は人間社会であり、その成員はまだ誰も結果を読んでいない。そのうえで事例を強め、結果が承認された妥当性確認を通過し（第5節）、検索、使用、訂正に組み込まれるようにする。すると、第6節で分けた関係のうち、なお欠けているのは主に、成員が結果に対する態度を獲得することになる。第6節で比べた説は、この欠如が問題になるかどうかで分かれる。判定はまず人間の S について述べ、そのあとで人工または混合の S に移す。

意図的に困難にした事例は、次のとおりである。ある人工システムが、人間社会にこれまで知られていなかった真なる命題を導出し、強い根拠と完全な記録を保存し、その結果を公共のデータベースに寄託する。いかなる人間もそれを読んでおらず、信じておらず、それに基づいて行為してもいない。人間社会はその結果を知っているのだろうか。

物理的な書き記しは存在する。人工システムの信念と根拠づけについての適切な説明のもとでは、そのシステム自体がその結果を知っているかもしれない。また、信念を要求しない説のもとで、その知識をもつこともありうる（第2節）。たとえばハンフリーズ（Humphreys）は、自動化された方法に合わせて信頼性主義の定義を修正し、ある装置が事実 F についての知識をもつのは、その装置が F の全体を表す表象を含み、その表象が対象について成り立ち、信頼できる過程、すなわち正確な表象を高い割合で生み出す過程がその表象を形成した場合であり、かつその場合に限る、と提案する。ハンフリーズは、この定義は表象が人間にとって接近可能な場合には理にかなっているが、表象が不透明でその正確さが分からない場合には問題をはらむと付け加える。そのときには、コンピュータ支援の数学についてのバージ（Burge）の権原（entitlement）のアプローチ、すなわち通常の視覚的知覚のような一部の情報源を原初的な認識的権威とみなすアプローチのようなものに頼る必要があるかもしれない、とも述べる（Humphreys 2020, 24–25）。これに対して、サイモンズ（Symons）とアルバラード（Alvarado）は、それより前に、このアプローチをコンピュータ支援の証明からコンピュータ・シミュレーションへ持ち込むことに反対している。ウィンズバーグ（Winsberg）とアルバラードが、コンピュータ・シミュレーションについての百科事典の項目でこの議論をまとめるところでは、シミュレーションは、バージの言う透明な伝達者、すなわち結果の根拠づけが、それが行う演算の根拠づけと同じ種類のものであるような伝達者ではない。シミュレーションのもつ特徴によって、別の種類の根拠づけ、すなわち経験に由来する根拠づけが、その過程に入り込むからである。そのためシミュレーションは、既定として信頼されるのではなく、それ自身の根拠づけと承認の手続きをもたなければならず、それは、シミュレーションの構築を可能にした方法の承認とは独立に追求されなければならない（Winsberg and Alvarado 2026, §2.3）。サイモンズとアルバラードが論じているのはシミュレーションであり、学習された表象ではない。しかし、ハンフリーズの提案に当てはめると、二人の説では、正確さが分からないために欠けているものを権原が補うことはない。そのため、表象が不透明な装置にその知識を帰属させることは、かわりに、現実のデータとの比較など、その装置自身についての評価に依拠することになる。これは、下で事例を強化するときに加わるのと同じ種類の妥当性確認である。こうした定義のもとでは、寄託したシステムは、装置が知識をもつと同じ仕方、信念なしにその知識をもちうる。それでも、人間社会がその結果を知っているかどうかは別の問いとして残る。その記録は、人間の知識への貢献の候補である。書き記しがあることも、システム自身が知っているかもしれないことも、記録が候補であることも、それだけでは人間社会がすでに知っていることを確立しない。社会の認識的組織から切り離されたファイルは、人間の探究の観点からは、気づかれていない真理と虚偽の無際限に大きな集積と区別がつかないままでありうる。

限定された結論は次のとおりである。

単なる書き記しや公共の寄託は、人間の社会的知識にとって理論中立的な十分条件ではない。

この言明は、保存が知識を部分的に構成することは決してありえないという主張よりも弱い。事例を強化してみよう。その記録は、*S*の制度が承認する手続きによる妥当性確認を通過している。その来歴は認証されている。それは正準的なインフラストラクチャーに索引づけられ、関連する問いが生じたときには自動的に検索され、推論と行為に用いられ、訂正や撤回を伝播させる仕組みに接続されている。それでも、いかなる人間も命題そのものをまだ読んでいない。

ここで哲学的な判定は分かれる。個人保有説は、人間は知る位置に十分にあるものの、まだどの人間も知っていないと言うことができる。ラッキーのような成員に基づく説明は、集団について同じ判定に達する。ラッキーの説を、当該領域に実働成員をもたない人間社会全体に当てはめれば、そこに集団的信念はなく（第6節）、判定に関わるのは、その中で実働成員をもつ集団である。それらの集団のどの実働成員もその結果を信じていないので、どの集団もまだそれを信じても知ってもいない。共同受容説は、成員たちがその結果を集団の見解として立てておくことに共同でコミットしたかどうかを問う。分散認知説は、全体に知識を帰属させることなく、そのインフラストラクチャーを真正の認知的資源として分類できる。知る主体なき知識を認める説明は、結果の地位を誰の態度からも切り離す。しかしギアリーの説は、関連する共同体が主張を公表し受け入れるまでそれを科学的知識とみなさない（Giere 2007, 318）。この事例に当てはめると、この条件は、読まれていない結果はまだその資格を満たさないだろうことを示唆する。ただし同じ箇所は、受け入れられた手続きによる認定が知っているという主張を支えることも認めており、強化した事例における承認された妥当性確認は、それに近づきうる。ポパー自身の例はこの事例に近い。計算機が計算して印刷し、図書館に送られ、誰にも読まれな一連の対数表は、ポパーが客観的知識と呼ぶものを含む。あるものを本にするのは、それが理解されうという可能性であって、実際に読まれることではないからである（Popper [1972] 1979, 115–116）。その定式では、寄託された記録は、真であれ偽であれ、すでに客観的知識に属する。しかし客観的知識はそこで、誰かが知っていると言主張することから独立なものとして定義されているので（109）、この定式は本節の問い、すなわち人間社会がその結果を知っているかどうかには答えない。バードの集団機能的な説を、誰も一度も読んだことのない記録に広げれば、組織されたシステムがその結果を検索し、使用し、修正できるのだから、その結果はすでに社会の知識の一部であると言うかもしれない。これに対して、以下で論じるバード自身の事例は、保持にかかわる。批判的文脈説は、ロンジーノの定義によれば、より確定的な判定を与える。

ビューター（Büter）が引用する定義では、受け入れられた内容が共同体にとっての知識となるのは、成員が企てを遂行できるほど意図された対象に合致し、その共同体で認知的に受容可能であるときである。ある時点で受容可能であるには、その時点で共同体にとって明らかなデータによって支持されていなければならない。また、その支持を成り立たせる推論と背景仮定は、共同体に利用できる限り多くの視点からの批判を生き延び、共同体は、批判の場、批判の受容、公的な基準、知的権威の調整された平等によって特徴づけられていなければならない（Longino 2002, 135–136, as quoted in Büter 2012, 84）。この事例に当てはめると、ロンジーノの定義からは、読まれていない結果は、自動化された妥当性確認の質がどうであれ、まだ社会の知識ではないことになる。どの成員もそれを受け入れていないからである。そこからさらに、結果が受け止められた後にその妥当性確認が受容可能性に寄与しうるかどうかは、妥当性確認そのものがロンジーノの条件のもとで批判にさらされてきたかどうかに依存することになる。

基本の事例と強化した事例の違いは、**休眠した記録**（dormant record）と**生きた認知的能力**（live epistemic capacity）の違いである。前者は単に存続するだけである。後者は、検索、有能な使用、依存関係の追跡、訂正と統合されており、強化によって加わったのは、第6節の関係のうちこれらにあたるものであった。この区別は、記録を生きた能力にするために何が要るかを名指す。しかし、それが知識の境界にあたるかどうかは、説によって異なる。学術論文がもはや誰にも覚えられていないというバード自身の事例は、結果にいま誰も気づいていないという点で、いまの事例に近い。N博士が査読を経た結果を公表する。それを読んだ者は皆、亡くなるか忘れてしまう。何年も後に、別の科学者が索引を通じてそれを見つける。バードは、その結果は必要とする者に通常の経路を通じて利用可能であり続けたので、その間ずっと科学共同体の知識の一部であり続けたとし、これを因果的に孤立し不活性であった真正の知ることとして擁護する（Bird 2024）。この見解では、妥当性確認され索引づけられた記録は、休眠しているあいだも社会的知識であり続けうる。ラッキーは、科学共同体がN博士の結果をずっと知っていたという主張に異を唱える。集団的知識は成員の知識に還元できなくとも個人的特徴に根ざしているとするカレストルプは、この事例ではラッキーの側に立ち、ラッキーの議論に着想を得た反論を次のように述べる。知っていることと知る位置にあることとのよく知られた区別は、アクセスされた情報と単にアクセス可能な情報との違いにかかわる。そのため、どの成員も結果にアクセスしていない集団は、それを知る位置にあるにすぎないと記述するほうが自然である。カレストルプ自身の論じるところでは、N博士の結果は誰の記憶にも、その拡張にも蓄えられていないので、共同体はそれを傾向的に信じてさえおらず、それを知る傾向性をもつ、つまり知る位置にあるにすぎない。図書館に収められていることや学術誌のウェブサイトに上げられていることは、それだけでは知識にならない（Kallestrup 2024, 583–585）。こうして、読まれていない、人工システムによる発見は既存の論争のなかに入るが、入る地点は異なる。バードの事例では、結果は休眠する前に査読を経て読まれていたので、この事例は保持にかかわる。どの成員も一度も読んだことも受け入れたこともない記録が社会的知識になりうるかどうかはさらなる問いであり、強化した事例はまさにこの点でバードの事例と異なる。バードの見解では、休眠と生きた能力の区別は、共同体の知識が終わる場所を標示しない。ラッキーの説明のひとつの読みでは、生きた能力でさえ、実働成員が結果を信じるまでは、当該領域で実働成員をもつ集団を知る位置にとどめるだけである。検索、有能な使用、訂正が保有を構成するとする説があれば、それはこの区別を、認知的資源が知識となる地点として扱うだろう。

この論争は、知識が行為とどう結びつくにかかっている。カレストルプがラッキーに帰するもう一つの反論は、二つの原理を組み合わせる。一つは、行為者が p を知っているなら、 p であるかのように行為することはその行為者にとって認知的に合理的である、という原理であり、ラッキーはこの向きでそれを受け入れている。もう一つは、ラッキーが**集団・成員行為原理**（Group/Member Action Principle）と呼ぶもので、集団がある行為をするのは、少なくとも一人の成員がその行為に因果的に寄与する何らかの行為をする場合に限られる、という原理である。問題は、どの成員ももっていない知識に基づいて、集団がどのように行為しうるのである。どの成員も結果に気づいていない共同体にこれを当てはめると、成員たちが、結果に偶然でない仕方にかかわる形で共同体の行為に寄与しうるのはどうしてなのかが謎になる、とカレストルプは結論する（Kallestrup 2024, 572 and n. 15, 585–586）。バードは、ラッキーの事例を翻案してこの反論を示す。指導的な専門家が、N博士の忘れられた論文がすでに確立していたことを単なる勘で主張し、分野がその主張に基づいて行為する、という事例である。そのうえでバードは、この結びつきが成り立つのはその知識に基づいてなされる行為についてだけだと応じる。人は p を知っているても、その知識が行為に何の役割も果たしていなければ非合理に行為しうるので、共同体の非合理的な行為は反論にならないのである（Bird 2024）。カレストルプが伝えるところでは、ラッキーはまた、集団が知っていることと知っているべきだったことをバードの説が区別でき

ないことを示すために、一つの事例を挙げる。同じ警察部隊の三人の警官が、それぞれ一つずつ事実を知っている。ある男の子が行方不明だと届けられたこと、その子がその朝何を着ていたか、そして同じ服装の子が大人の男と一緒に近くの地区を歩いていたことである。ラッキーの読みでは、バードの説に従えば、署は、目撃された子が行方不明の子である見込みが高いことを知っていることになる。ところが、それに基づいて行為しなかったことについて、どの警官にも責任があるとは思われない。アヴラム・ヒラー (Avram Hiller) と R・ウルフ・ランドール (R. Wolfe Randall) は、この事例を変形する。各警官は自分の知っていることを中央のシステムに入力し、システムは入力を照合して、それらが組み合わさると結論を印字する。そうなれば、どの警官も印字された結論に容易に接近できるので、二人は、集団が行為しなかったことに責任を問うのが直観にかなうと考える。そして、集団の構造がその種の情報に接近するように設計され、実際に適切に機能しているので、集団はその結論を知っていると論じる。カレストルプは、変形した事例の署に責任があるとしても、それは署が結論を知っているからではなく、知っているべきだったからかもしれないと応じる。知識は傾向的な信念を伴い、生物的でない保管庫がそうした信念を支えるのは、個人が記憶のなかで情報を蓄え取り出す仕方の機能的な類似物として働く場合に限られる、というのである (Kallestrup 2022)。自動化されたこの変形は、強化した事例に近い。カレストルプの応答は、集団そのものがどのようにしてその結論を傾向的に信じるのかを問う。その説明がなければ、照合が集団に与えるのはせいぜいその結論を信じる傾向にとどまる。カレストルプが自然な候補として挙げるのは、集団の水準での拡張された認知過程であるが、これはヒラーとランドールが検討したうえで採らなかったものである (Kallestrup 2022)。この応答でも、上に述べたラッキーの説の読みと同じく、自動の照合と使用は、集団による記録の統合がそうした過程にあたるのでないかぎり、それだけでは結果を集団の知識にしない。そのうえでこの応答は、結果を照合するシステムをもつ集団が、知っているべきだったことについてはなお答えを求められうることを付け加える。

以上の判定は、人間の S についてのものである。人工の成員が記録を自動的に検索し、推論と行為に組み込むような、人工社会あるいは混合社会である S に移すと、上に示した諸説は、すべてが同じように動くわけではない。個人保有説では、問いは、誰かがその結果を読んだかどうかではなく、人工の成員による自動の検索と使用が、その成員がその結果を信じていることにあたるかどうかになる。これは第2節の信念の説へ戻る問いである。ラッキーの成員に基づく説では、問いは、その社会の人工の実働成員がその結果を信じているかどうかになり、また、その検索が、単にアクセス可能な情報ではなく、アクセスされた情報にあたるかどうかになる。行方不明の子どもの事例の自動化された変形は、これを決めない。そこで入力を照合するシステムは成員に含まれていなかったからである。この説では、実働成員とは該当する領域で意思決定の権限をもつ成員なので、どの人工の成員がそれにあたるかは、人工の S のなかでその権限がどう分配されているかによる。これは第12章第5節が取り上げる問いである。共同受容説では、問いは、人工の成員が共同受容に加わりうるかどうかになり、知る主体なき知識のギアリーによる翻案では、人工の成員による認定が認定に数えられるかどうかになり、成員がひとつの基盤システムのインスタンスであるときには、第6節で述べた理由から、それがとりわけ問われる (第11章第3節)。ロンジーノの定義のもとでは、人工の成員がロンジーノの意味での知る共同体の成員でありうるなら、つまり第8章第1.3節で見る多様性と公共的な批判の基準を伴う共同体をなしうるなら、成員による受け入れという条件は、人工の成員によって満たされうる。その場合、結果が受け入れられたかどうかは、人間の誰かがそれを読んだかどうかではなく、人工の成員がそれをどう扱うかにかかる。これに対して、バードの集団機能的な判定は、このようには変わらない。その判定はもともと、成員の誰かが結果に気づいているかどうかではなく、組織されたシステムが結果を検索し、使用し、修正できるかどうかにかかっていたからである。人工の S についても、問いは、人工の成員が全体として適切に組織された認知システムを実現しているかどうかのままである (第6節)。ただし、バードの説では機能はのために選択されてきた働きなので、その組織が、行為を導く事実的な表象状態を生み出すために選択され、あるいは設計されたのか、それとも、たとえばチェックに通ることや承認を得ることのために選ばれたのかも問われる。第4.1節のシミオンの説の場合と同じく、選択が実際に何に報いてきたかは、掲げられた目的と食い違いうる。

結果を産出したシステム自身が、人工または混合の S の成員であることもある。その場合、採用する信念の説のもとでシステムが結果を知っているなら、個人保有説では S の成員がすでに知っていることになり、結果は産出された時点で S の認識的状況に入っていることがありうる。ラッキーの説では、システムの信念が集団の信念に数えられるのは、システムがその集団の実働成員である場合に限られ、しかも集団が結果を信じるには、実働成員のかかなりの割合が信じていなければならない。ロンジーノの説では、システムが受け入れることはたかだか一つの成員の受け入れであり、受容可能性には、結果が共同体のなかで利用できる他の視点からの批判を生き延びることが要る。バードの説では、判定はなお、組織されたシステム全体にかかる。ここから、人工社会に知識を認める道は集団機能説だけではないことも分かる。個人保有説のままで、誰を知る個人に数えるかが変われば、人間の誰も読んでいない結果について、成員の誰も知らないとは言えなくなる。ただしそれには、機械が名目上の成員ではなく、採る信念の説の意味で知る成員であることが要る (第2節)。

基本の事例をさらに切り離した極限、すなわち S のどの成員も、人間であれ人工であれ、結果を観察・検索・評価・使用できない場合には、 S がその結果を知っているかに答える上の諸説は、まだ知らないという点でおおむね一致する。ポパーの客観的知識は、上で見たとおり、別の問いに答えている。公共への寄託である基本の事例そのものは、この極限に達するとは限らない。寄託は通常の経路を通じて利用可能だからである。ただし、バードの説では、知ることは適切に機能している認知システムが生み出す状態であり (第6節)、上で見た忘れられた論文も査読を経ていた。そのため、休眠していても知識でありうるという判定を誰も読んでいない記録へ広げるとしても、それがよく当てはまるのは、ただ寄託されただけの記録より、 S の制度が承認する妥当性確認を通過した、強化した事例の記録である。残るのは、そのように完全に切り離された真なる記録を何らかの意味で知識と呼ぶかどうかだけであり、それがどう決まっても、結果が S の認識的状況に入るためにつくるべきものは同じである。すなわち、結果が妥当性確認、推論、行為、訂正に入っていける経路である。この経路は、研究を行ったシステムの内部にあるとは限らず、システムと S のあいだのインフラや、 S の側にあることもある。それでもこの境界事例は、アップロード (upload) が **社会が知っている** (society knows) と同一視されるのを防ぐので、哲学的に重要であり続ける。

これに対して、そうした経路がある強化した事例では諸説が分かれ、その分かれ方は、結果が S の認識的状況に入ったと主張するプロジェクトが何をつくらなければならないかに関わる。上のように拡張したバードの説では、経路があることで、結果はすでに S の認識的状況に入っているといえるかもしれない。ラッキーの説では、そのプロジェクトは、さらに結果を S の成員に届け、信じられるようにしなければならない。集団が知ることを通じて入ったと主張するなら、届け先は、 S のうち当該領域で実働成員をもつ集団の実働成員になる。ロンジーノの説では、その条件のもとで批判を経て、成員に受け入れられるようにしなければならない。したがって、そのプロジェクトが何をつくらなければならないかは、採用する保有の説によって変わる。第6節で述べた

ように、結果が S の知識になったという主張が、自らの依拠する説明を十分に明らかにしているときにはじめて評価できるのは、そのためである。活動が S のための研究であるかどうかは、次の段落で見るように、結果が、その保有の説の求める条件を満たして S の認知的状況に入ったかどうかには依存しない。

この事例は、第1章で立てた二つの区別を具体的に示している。人工システムの活動が人間社会にとって新しい知識を産出することへ向けて組織されているならば、それは、人間が誰もその結果を獲得しないうちから、作業的特徴づけの意味での研究でありうる。**試み** (attempt) が研究活動をその成功から切り離すからである (第1章第3節)。そこでの例、たとえば決定的な結論を出せない実験や崩れた証明の戦略は、達成が欠けていることによって両者を分けていた。これに対してここでは、達成はあるかもしれないが、まだ S のうちにはない。そのため、成功の側そのものが分かれる。第1章第2節の三つの判断でいえば、第一の、活動が人間社会のための研究であるかは、成り立ちうる。第二の、人間社会にとって新しい知識がどこかで実現したかも、採用する信念の説のもとでシステムが知りうるなら、システム自身のうちで成り立ちうる。第三の、その結果が人間社会の認知的状況に入ったかは、採用する保有の説に依存し、まだ成り立っていないこともある。人工または混合の S でも、産出したシステム自身が成員であり、採用する説のもとでその知ることがすでに数えられる場合を除けば、判断は同じように分かれ、第三の判断は、上で移した諸説によって決まる。

8. 知識の手前と周辺における認知的変化

研究は知識に届かないまま終わることが多い。また、成功した場合でも、研究が達成するものが知識であるとは限らない。本節は、こうした成果を、知識だけを数えるのでも、あらゆる変化を改善として数えるのでもなく、どう記述し評価できるかを問う。本節が取り上げる文献は、この問いを、おもに一つの限定された形で論じてきた。こうした成果を進歩と数えてよいか、という形である。とくに科学的進歩の説のあいだの争いは、おおむね、知識の蓄積だけを進歩と数えるのか、それ以外の成果も進歩と数えるのかをめぐっている。本節は、この限定された形での議論をたどる。そうすることで、こうした成果を評価するときに何を区別しておく必要があるかが見えてくるからである。

本節はまず、研究活動を、知識に至るという成功から切り離す (第8.1節)。第8.2節は、研究が知識に届かなかった場合を扱う。そこで生じうる知識の手前の変化として、無知を明示化することを中心に、問いを立てること、証拠を集めること、確信の度合いが変わることを取り上げ、それらが進歩と数えられるかを問う。そこで引く議論は、科学に限らない認知的進歩についてのものである。第8.3節は、成果が知識そのものではない場合を、科学的進歩の説とあわせて扱う。偽である理論が真理に近いことや、問題を解決することのような、知識の周辺にある達成である。理解も、知識の一種とはみない立場では、ここに入る。理解が知識の一種かどうかは、第3章第3節で扱う。本節の終わりでは、これが人工システムによる研究の評価にとって何を意味するかを述べる。

8.1 研究は知識に到達せずとも適切に遂行されうる

第7節は、知識がすでに実現しているかもしれないが、まだ S の認知的状況に入っていない事例で、研究を成功から切り離した。本小節は、第1章第3節が出発点とした場合、すなわち求める知識にそもそも届かない場合を扱う。中心的主張が偽と判明したり、結果が決定的でないままに終わったりする場合である。**試み** (attempt) が研究活動をその成功から切り離すので、そうした探究も、それによって研究であったことをやめはしない。また、研究であったかどうかの判断は、結果を待たずに下しうる。その判断は、活動が成功したかどうかではなく、活動がどう組織されていたかにかかるからである (第1章第3節)。ただし、成功から独立であることは、着手の時点で判断が定まることまでは意味しない。活動が途中で研究へと組織し直されることもあれば、その組織的な方向づけが後の記録からはじめて分かることもある。

そこで本書は、知識を目標とする場合について、少なくとも次の六つの評価を区別しておく。

1. ある活動が、作業的特徴づけの意味で S のための研究であるかどうか。
2. それが、利用可能な状況に相対的に、責任をもって、あるいは有能に遂行されているかどうか。
3. S にとって新しい、目標となる知識が、 A を含むいずれかの担い手のうちに、あるいは知る主体なき知識を認める説のもとでは認定された結果のうちに、実現したかどうか。
4. その知識が S の認知的状況に入ったかどうか。これには第6節の保有の説が異なる答えを与える。
5. 目標における失敗が、副次的な認知的貢献を生み出すかどうか。
6. その貢献が重要であるかどうか、あるいはそのコストに見合うかどうか。

第一、第三、第四の評価は、第1章第2節の三つの判断にあたる。

適切に遂行された探究が目標において失敗することはありうる。拙く遂行された探究が運によって真なる答えに到達することもありうる。失敗した実験は、方法、補助仮定、あるいは元の問題の定式化における限界を明らかにしうる。その副次的な結果が知識であるかどうかは、それ自体の評価を要する。それが重要であるかどうかは、さらに、目的や影響を受ける当事者とそれがどう関係するかによる。ここでの用法では、それが知識であるかどうかは、研究から生まれたかどうかにも、重要であるかどうかにも左右されない。

8.2 知識の手前の変化：無知の明示化と、それが進歩かどうか

知識に到達できないことは、単なる進歩の欠如にとどまらないことがありうる。社会は、ある事柄について誤っていることもあれば、その事柄が未解決であることに気づいていないこともあれば、問いには気づいているがもっともらしい答えを同定できないこともあれば、問いと、それに答えるためになお必要な証拠との両方を特定できることもある。これらは、いずれもまだ目標となる知識を含んでいないとしても、実質的に異なる認知的位置である。

ヴィルホルト (Wilholt) は、全面的な無知のなかで、意識された無知 (conscious ignorance)、すなわちおよそ主体がその答えを知らないと認識している問いを、不透明な無知、すなわち知らないことを知らないものから区別する。誤りや不透明な無知から意識された無知への移行をソクラテ斯的進歩 (Socratic progress) と呼ぶ。意識された無知はそれ自体、実質的な概念的・認識的有能さを要求する (Wilholt 2020, 2-6)。意識された無知のうち、もっともらしい候補となる答えが利用可能でないと主体が自覚している場合に限り、ヴィルホルトは深い無知 (deep ignorance) という呼び名を用いる。この場合を、ブロムバーガー (Bromberger) が p 窮地 (p-predicament) と呼んだものと同一視する。ある主体がある問いについて p 窮地にあるのは、その主体の見解のもとでは問いに正しい答えがあるのに、同じ見解のもとで決定的な反論のない答えを一つも思いつけないとき、かつそのときに限る (Wilholt 2020, 9-10)。ヴィルホルトは、そうした問いは研究者にとって特に魅力をもつと論じる。それを乗り越えれば、新しい問いと結びつきが開かれると見込まれるからである。ただしその結論は、深い問題と受け取られることが、その問いを研究の課題に含める方向に働く、というところにとどまる。そうした問いは、たとえば利用できる方法のどれもそれに取り組む手立てを与えないときには、見送られうる (Wilholt 2020, 10-11)。ヴィルホルトのソクラテ斯的進歩の説明は、まだ答えについての知識である必要のない認識的改善を、精確に描いた例である。また、本章の達成の分析を、問いを理解し、扱えるものにするということについての後の説明に結びつける助けにもなる (第4章第3節)。

ただし、この論点が示すことには限りがある。それは、答えられていないあらゆる問いを価値ある研究上の貢献にするものでも、無知を知識と同一視するものでも、問いを分節化することが研究にとって十分であると示すものでもないということである。重要性、取り組みやすさ、新しさ、分節化の質は、それぞれ別の問題である。

本節がヴィルホルトの説明から受け取るのは、次の対比である。一方は、知識が欠けていることがどこにも表されていない状態である。他方は、欠けていることが、記録への書き込み、誤差の見積もり、立てられた問いなど、何らかの形で表されている状態、つまり明示された認識的限界である。後者だけが、どこに注意を向け、何を調べ、何を批判するかの手がかりになる。意識された無知は、いつも「自分はこれを知らない」というはっきりした信念の形をとるとは限らない。重要なのは、欠けていることがこうした何らかの形で表されているかどうかである。

作業の記録は、そうした限界を明示したまま保つことができる。たとえば、ワイス (Weiss) は1972年、レーザー干渉計による重力波アンテナの設計研究で、装置の雑音源を一つずつ見積もり、必要な知識が欠けている箇所を書き留めた。地面振動のスペクトルは100 Hzより高い周波数では測定されておらず、地表付近の高周波の重力勾配についてはデータがなかった (Weiss 1972, 64, 67)。主張された発見の解体も、そうした限界を明示することから始まる。2009年の報告がレトロウイルスXMRVを慢性疲労症候群と結びつけた後、複数の研究所の作業部会は、臨床検体を調べる前に、どれにウイルスを加えたかを伏せた試料を用意し、各研究所の検査法がそのウイルスを検出できるかを確かめた。つまり、検査法で何が分かり、何が分からないかを、まず明らかにした。その後の盲検の多施設研究は、関連を見いださなかった。この研究の設計には参加したすべての研究所が事前に合意しており、著者には当初の主張者も含まれていた (Simmons et al. 2011a; Simmons et al. 2011b; Alter et al. 2012)。

そうした限界を、持続的な研究の対象にしてきた分野もある。たとえばクノール＝セティナは、自らが観察した衝突型加速器の実験が、本人の言う否定的知識 (negative knowledge) を培っていたと報告する。これを非知識から区別し、知ることの限界、知ろうとするとときに犯す誤り、知ることを妨げるもの、そして関心がなく知りたいとも思わないものについての知識とする。ある物理学者は、系統誤差、つまり偶然のばらつきではなく装置や解析の仕方由来する測定の不確かさを、自分たちの無知を測る方法だと述べた。また、陽子の内部構造を記述する約45の競合する理論が解析に異なる入力を与えたとき、それらから得られる値の広がり、理論のあいだの意見の相違のまま残されるのではなく、理論的な系統不確かさとして結果に明示された (Knorr Cetina 1993, 269-273)。クノール＝セティナはこれをその文化の選好として示し、分子遺伝学のような分野はそれを共有しないとする (Knorr Cetina 1993, 270-271)。この対比の分子生物学の側は、第6章第4.3節で取り上げる。

どこにも表されていない欠如と、明示された限界とのこの区別は、人工システムによる研究に直接の帰結をもつ。閉じた問いや出力した答えの数で評価すると、証拠がまだ決着をつけていない問いを早々に閉じてしまうシステムほど、高く評価されかねない。そうしたシステムは、出力をもっともらしい主張で埋め尽くす一方で、それに頼る者にとって、不確実性、欠けている証拠、両立しない候補、定式化されていない問題を見えにくくすることがありうる。逆に、システムは、一見決着のついた答えが支持を欠いていることを示したり、どの代替案がなお生きているかを同定したり、以前には分節化できなかった限界を問いうるものにしたことによって、Sの研究上の位置を改善するかもしれない。したがって人工研究システムは、開かれた問い、判断保留の理由、失敗した操作化、欠けている証拠の関係を、答えの生成後に残された空欄としてではなく、第一級の認識的状态として保持する必要があるかもしれない。ただし、そうした状態をシステム自身の記録に保つことは、まだSの位置の変化ではない。意識された無知への移行がSにおいて起こるには、表された限界がSの認識的状况に入らなければならない。Sが人間であれば、限界は、人間の成員が問いとして受け取れる形で彼らに届かなければならない。Sが人工であれば、人工の成員がそれを自らの探究における開かれた問いとして扱えなければならない。第4～6章は、そうした状態が探究のなかでどう形成され、変わっていくのかを検討する。ここでの主張は、それらが認識的進歩にとって重要でありうるということだけである。それらが進歩を構成するのか、進歩を促すだけなのかは、本小節の終わりで見るように、進歩についての説によって分かれる。

ウィリアムソンの議論は、システムが自分の知らないことについてどう記録を、どこまで完全にできるかにかかわる。ウィリアムソンは、ほとんどどんな非自明な条件についても、それが成り立つあらゆる場合にそれが成り立つと知る位置にあるわけではないと論じ、とりわけ、何かを知らないでいながら、自分がそれを知らないことを知る位置にないことがありうると論じる (Williamson 2000, 13-14, 107)。情報を蓄える生き物についてのウィリアムソンの単純なモデルでは、蓄えたものを見渡せない生き物は、何かを知らないでいながら、それを知らないことを知らないことがありうる。その一つでは、記憶からの取り出しにいくらでも時間がかかりうるので、その生き物は、記憶が欠けているのか、まだ届いていないだけなのかを区別できない。ウィリアムソンはこうした欠陥を、通常の意味での合理性の欠陥ではなく、自己知の欠陥と記述する (Williamson 2000, 226-227, 233-234, 316-317)。ここで効いているのは、蓄えを一度に見渡せないことそのものより、取り出しにかかる時間に上限がなく、記憶に欠けていることとまだ届いていないことを区別できないことである。これらの議論が正しいとして、この条件を満たす人工システムに当てはめると、システムが自らの開かれた問いと欠けている証拠についてどう記録を、システム自身の点検だけでは完全であると保証できないことになる。これに対して、有限の蓄えを調べ終えたことを確かめられる人工の構成には、この論拠はそのままでは当てはまらない。ただしその場合にも、点検で確かめられるのは蓄えに表されていることだけであり、上で見た不透明な無知のように、どこにも表されていない欠如は、蓄えを調べ終えても現れない。

ここでの用法では、過程の語彙は、あらゆる変化を改善へと変えはしない。ロス（Ross）は、知的共同体は答えを知る前に進歩しようと論じる。すなわち、価値ある問いを定式化したり、関連する証拠を獲得したり、正しい答えへの確信を高めたりしうる。その理由として、進歩として広く認められている出来事が、共同体がまだ割れているあいだに起こることを挙げる。ガリレオが、月の表面は平らでないことを示す望遠鏡による証拠を報告した後も、共同体はそれを信じていても知ってもいなかった。それでも共同体が進歩しなかったとするのは、ロスによれば、あまりに大きな修正を迫る見方である。また、こうした出来事が進歩とみなされるのは後の知識を促進するからにすぎない、という見方も退ける。その基準では、研究グループのためにコーヒーをいれることまで進歩に数えられてしまうからである（Ross 2021）。これに対してハリス（Harris）は、問いを定式化することや証拠を集めることは進歩を促進するのであって、進歩を構成するのではないと反論する。共同体は、多くの問いを定式化したり、実りなく証拠を集めたりするだけでは進歩しないだろう。ただし、真なる命題への確信が増すことは、真理を中心とする説を信念の度合い（信念度）によって定式化し直せば、進歩として受け入れられるかもしれない（Harris 2024, 142–143）。ハリスの論点を上のヴィルホルトの説に移すなら、その論点は、ヴィルホルトがソクラテスの進歩と呼ぶもののうち、不透明な無知から問いを認識することへの移行に当てはまり、その移行を、進歩を構成するものではなく促進するものとみなすことになるだろう。誤った信念を手放すことにそれが当てはまるかどうかは、それほど明らかではない。信念度についてのハリスの譲歩が、その場合を覆いうるからである。したがって、ソクラテスの進歩についての上の主張は、進歩のあらゆる説のもとではなく、ヴィルホルトの説のもとでなされている。

この論争は、進歩の理論に決着をつけないまでも、一つの制約をもたらす。問いの形成、証拠の獲得、信念度の変化、認められた無知は、候補となる認識的変容である。特定の変化が進歩を構成するかどうかは、それが価値ある目標、正しい答え、誤りの露呈、あるいは探究能力の改善とどのような関係にあるかに依存する。より多くの確信、より少ない不確実性、より多くの情報、より多くの問いは、それだけでは十分ではない。

8.3 知識の周辺の達成と進歩の説

本小節では、知識の手前の変化から、知識の周辺の達成へと移る。まず取り上げるのは、歴史的には成功したが偽である理論である。本章は、命題的知識について事実性（factivity）をとる。つまり、知られていることは真である（第3節）。そのため、偽である理論は、どれほど成功していても、それ自体としては知られていたことにならない。では、その理論が成し遂げたことをどう記述すればよいのか。いくつかの答え方がある。知られていた真なる限定的主張を同定することもできるし、偽である理論のうちその成功に責任のあった部分を近似的に真とみなすこともできるし（第6章第3.2節）、文字どおりの信念と実践的受容とを区別することもできるし、近似的真理や経験的十全性に訴えることもできるし、理解や問題解決能力を知識の蓄積に尽きない達成として扱うこともできる。本小節は、これらの選択肢のいくつかを取り上げる。そのなかには、バードの限定的主張、ニーニルオト（Niiniluoto）の真理近似性、ラウダン（Laudan）の問題解決、デルセン（Dellsén）の理解が含まれる。事実性は、これらの分析のどれをとるかを決めない。事実性が定めるのは、本章で命題的知識という語をどう使うかだけである。こうした理論のより広い認識的達成をどう記述するかは、これらの分析と後の検討に委ねられる。

これらの選択肢のいくつかは、科学的進歩の説として展開されてきた。デルセンは、その文献を概観して四つの説を区別する。それぞれの説は、固有の種類の認識的達成を進歩の中心に置く。真理近似性（truthlikeness）の増大、問題解決の有効性の増大、知識の蓄積、そして理解の増大である。最後のものはデルセン自身の説で、本人はこれを、理解を意味するギリシア語にちなんでノエティック（noetic）説と呼ぶ。デルセンはさらに、最ももつとらしい説は、複数の種類の達成を進歩に数える多元的なものかもしれないと付け加える（Dellsén 2018, §1）。この整理では、進歩は、知識や理解と並ぶもう一つの達成としてではなく、いずれかの種類の達成が増大し、蓄積し、あるいはより十全に達成されることとして分析されている。本書の用語でいえば、進歩の諸説が争っているのは、どの達成の増大を進歩とみなすかである。第1章第1節で述べたように、本書の作業文は研究活動の方向づけを特徴づけるものであり、科学的進歩の理論として役立つことを意図していない。とりわけ、認識的達成を通じて研究を特徴づけることは、進歩とは知識の蓄積であるとするバードの認識的説（epistemic account）に、本書をコミットさせない（Dellsén 2018, §§1, 4）。第3章第4節は、デルセンの説を含む理解に基づく見解を、理解と知識が目標および進歩の尺度として異なるのか、異なるならどう異なるのかを検討するために取り上げる。デルセンは、進歩と科学の目的とのあいだには密接な概念的つながりがあるように思われると述べる。おおよそ、Xが科学の目的であるのは、Xが増大し、蓄積し、あるいはより達成に近づくときに科学が進歩する場合であり、その場合に限る、というものである（Dellsén 2018, §6）。このつながりが成り立つとすれば、これらの説のひとつを採用することは、第1章第1節が確立しない種類の大域的な目的を確立することになる。以下の段落は、これらの説のあいだで選択することなく、人工システムによる研究の評価にかかわる区別をそれらから引き出す。ただし、これらは科学的進歩の説として提出されたものであり、作業的特徴づけのほうは科学に限られない。そのため本書は、これらの説から取り出した区別を研究一般に用いるが、ある説の尺度が科学以外の分野にも合うかどうかは、また別の問いである。

歴史的には成功したが偽である理論は、とりわけ、進歩を知識の蓄積とみるバードの説にとって難題になる。偽である理論は、それ自体としては蓄積された知識になりえないからである。しかし、一連の偽なる理論が真理に近づいたり、次第に重要な問題を解決したり、理解を改善したりすることはありうる。進歩を理論が真理に近づくことにみる真理近似性の説は、この難題を通り抜ける一つの経路を表す。別の経路を表すのは、進歩を問題解決の有効性の増大にみる問題解決的な説、理解の増大にみるノエティックな説、そして下で見るように、研究問題を解くことだけでなく定義することにも進歩をみる機能的な説である（Niiniluoto 2024, §§3.2, 3.6; Nickles 2017, §1.5; Dellsén 2021, 11259–11262; Shan 2019）。問題解決的な説を例にとると、その経路は次のようになる。デルセンが指摘するように、ラウダンの説は、進歩が事実的であることをいっさい求めないように意図して作られている（Dellsén 2025, 2128）。そのためラウダンの説は、どれも後に偽と判明するかもしれない理論の継起を、進歩的なものとして数えることができる。シャン（Shan）の機能的な説は、本人がクーン（Kuhn）とラウダンに帰する問題解決的な考え方を改め、研究問題の解決だけでなくその定義をも進歩の構成要素に数える。この説によれば、ダーウィン（Darwin）のパンゲン説は、それが措定したジェミュールは存在しないにもかかわらず、不完全ではあれ有用な問題と仮説の枠組みを与えることで、遺伝の研究における進歩をなした（Shan 2019, §§2–4; Dellsén 2021, 11266 を参照）。これらの説は進歩についての互いに対立する分析であって、知識の定義に対する交換可能な修繕ではない。

最初に挙げた真理近似性の説に戻ると、ニーニルオトはこの説を、進歩を真理近似性の増大として分析する形で定式化している。真理近似性は、ある認知的問題に対する最も情報量の多い真な答えへの近さとして測られ、そのため、ある偽なる理論から別の偽なる理論への一歩も進歩的でありうる。ニーニル

オトは、その答えが知られていないあいだは認識されないままでありうる実際の進歩と、利用可能な証拠のもとでの真理近似性の期待値によって判断され、証拠が変わるにつれて改訂されうる推定された進歩とを区別する (Niiniluoto 2024, §3.5)。進歩とは知識の蓄積であるとするバードは、真理や真理近似性だけで進歩を定める説では、非合理な方法によって正当化なしにたまたま得られた、真である、あるいは真理に近い信念まで進歩に数えられてしまうと批判していた。これに対してニーニルオトは、チェヴォラーニ (Cevolani) とタンボロ (Tambolo) にならって、推定された進歩がすでに正当化に場所を与えていると論じる。正当化は実際の進歩を成り立たせるものではないが、推定された進歩は証拠と正当化に相対的であり、実際の進歩の主張は真理近似性の期待値に訴えて正当化できるからである。また、近似的に真な理論を、その理論が近似的に成り立つという完全に真な主張として扱うことで偽なる理論を処理するバードのやり方で、真理に近づく理論の系列と真理から遠ざかる系列とを区別できるのかを、さらなる問題として挙げる (Niiniluoto 2024, §3.6)。本小節の初めに挙げた、歴史的には成功したが偽である理論についての選択肢と照らし合わせると、この問題はその最初のもの、すなわち知られていた真なる限定的主張を同定することにかかわる。理論が近似的に成り立つというバードの主張は、その種の主張だからである。

真であることと、それを確かめられることとの、第3節で見た分離は、科学的進歩についての事実的な説明 (factive accounts) に向けられた一つの異論にも関わる。デルセンが伝えるところでは、ラウダンは、事実的な説明のもとでは科学者自身が進歩に認識的に接近できなくなると論じた。最も経験的に成功した理論でさえ偽と判明するかもしれないので、科学者は、後の理論が先行する理論より事実上忠実かどうかを見分けられない。そしてラウダンによれば、そのように接近できない進歩は、科学における合理的な意思決定を導きえない (Dellsén 2025, 2128, 2140)。ニーニルオトの言い方では、ラウダンは、真理を知ることが「ユートピア的」な課題だという古典的な懐疑論の議論を繰り返している (Niiniluoto 2024, §2.4)。デルセンはこれに対し、合理的な選択に必要なのは、利用可能な各行為の方針が (下で定義する意味で) 進歩をどれだけ促進するかについての合理的な信念度であって、実際に促進するかどうかへの接近ではないので、進歩の事実的な基準もおお選択を導きうると応じる (Dellsén 2025, 2128, 2140–2141)。

上で見たニーニルオトの実際の進歩と推定された進歩の区別は、ここで保たれている事実的な枠組みの内側から同じ異論に応じるものと読むことができる。一方ライヘンバッハ (Reichenbach) は、それより前に、その枠組みの外から、同じ形の困難に取り組んでいたと見られる。ライヘンバッハは、真理を高い重みの理想化とみなし、知識を真な命題の体系ではなく措定 (posit) の体系とみなした。そして、帰納は成功するかどうかを知らなくても正当化できると論じた。頻度の極限が存在するなら、帰納を続けてその推定を正していけば、推定値は、極限からの幅をどれほど小さく定めても、いずれその幅の内に入り、そこにとどまるはずだからである。ただしライヘンバッハが認めたように、収束は遅すぎて、人間の一生のうちには観察されないこともありうる (Reichenbach 1938, 187–191, 348–351, 360–362, 403–404)。さらにライヘンバッハは、この議論で正当化されるのが帰納だけではないことも認めた。ここでいう帰納は、これまでに観察された頻度をそのまま極限の推定値とする規則である。観察された頻度に、観察の回数が増えるにつれて0に近づく補正を加える規則でも、その推定値は同じ極限に近づくからである。ライヘンバッハはそうした規則のうちで、リスクが最小であるという実際の理由から帰納を選んだ (Reichenbach 1938, 354–356)。そうであれば、この種の長期的な保証だけでは、ある時点でどの方法に従うかは決まらないことになる。それを決める導きこそ、ラウダンの異論がその欠如を突いたものであり、デルセンの合理的な信念度への訴えが応じようとするものである。人工システムの学習や更新の規則についての保証がこうした長期的な形をとる場合にも、同じ理由から、その保証だけでは、システムがある時点で何を結論することが根拠づけられるかは決まらなと考える。

真理近似性の比較は、問題がどう立てられるかにも依存する。オディ (Oddie) とチェヴォラーニは、ミラー (Miller) に由来する次の議論を報告している。真理近似性を理論と真理との類似の度合いで測る説 (ニーニルオトの説もその一つである) では、同じ二つの理論を別の原始述語で述べると、その順位が逆転しうる。オディとチェヴォラーニの例では、天気を「暑い」「雨が降る」「風が吹く」の三つの述語で記述することも、「暑い」「ミネソタ的」(暑いとき、かつそのときに限り雨が降る)「アリゾナ的」(暑いとき、かつそのときに限り風が吹く)の三つで記述することもできる。どちらの語彙も他方から定義できるが、前者で数えると誤りの少ない理論が、後者で数えると誤りの多い理論になりうる。二人が述べる二つの応答のうち、一つは、「暑い」や「雨が降る」は実在をその自然な区切り目 (関節) で切り分けるが、「ミネソタ的」のような寄せ集めの述語はそうではない、とする。この応答をとると、真理近似性を推定するには、どの述語が実在をその関節で切り分けるのかも推定する必要が生じる。述べ直された理論が同じことを言っていることを否定するもう一つの応答は、異なる概念で組まれた理論どうしの比較を難しくする (Oddie and Cevolani 2022, §1.4.3)。そう考えると、ニーニルオトの意味での推定された進歩は、ある認知的問題とそれを述べる概念とに相対的であり、それらの概念の改訂 (第4章第4節、第5章第3節、第6章第5節) は、何が比較されているのかを変えうる。ミラーの議論を、表象の選択も担う人工システムに当てはめると、さらに次のことが考えられる。システムが成果を述べる語彙を自ら選ぶなら、同じ成果が、選ばれた語彙のもとでは真理により近いと評価されることがありうる。そのため、表象を改めたことで対象をよりよく捉えたのか、評価の尺度のほうを変えたのかは、別々に問う必要が生じる。上の一つめの応答に立てば、前者を推定するには、どの語彙が実在をその関節で切り分けるかの推定も要る。

さらなる区別は、時間的な貢献に関わる。データセット、形式体系、測定器具、失敗した実験、あるいは分類は、認識的前進を構成することもあれば、後の前進を促進することもあれば、その仕事着手される時点で前進を促進すると合理的に期待されるだけのこともある。デルセンは、第二と第三の述語に精密な形を与える。その説明では、ある出来事が進歩を促進するのは、それに続くうる発展における進歩の量を、各発展の客観的確率 (chance) で重みづけして足し合わせたものが、その出来事が起きなかった場合のそれを上回る程度に応じてである。ある行為者が、ある出来事が進歩を促進すると期待することが合理的であるのは、同じ比較を行為者自身の整合的な信念度で行ったときに、その出来事の側が上回る程度に応じてである。デルセンの例では、しくじった実験が科学者たちを再実験の成功へと導いたとしても、最初から正しく実施していればさらに先へ進めたのであれば、その実験は進歩を促進していない (Dellsén 2025, 2131–2135)。これらの述語は、互いに、また実際に後に続いた結果とも食い違いうる。区別しておけば、これらの述語は、有用な人工物に、その因果的な未来におけるあらゆる後の結果の功績を自動的に帰すことはなく、また、予見できない出来事がその後の使用を妨げたというだけで責任ある試みを非合理とすることもない。

人工研究システムの成果を、一つの進歩スコアで表すことを考えてみよう。そのためには、新しい知識、排除された可能性、改善された方法などを、それぞれどれだけ重く数えるかを決める必要がある。ところが、まさにこの点で、上で見た進歩の説は互いに異なる。そのため、一つのスコアが成果の進歩そのものを表すと主張するなら、その主張は、それだけで進歩の説のどれかを選ぶことになる。限られた側面だけを測る代理の指標や、さしあたりの判断のためにまとめた値であれば、そこまでのコミットメントは生じない。進歩の説を選ばずにシステムの成果を評価するなら、スコアではなく、変化の種類ど

とに分けた記録を保つことになる。たとえば、目標となる知識、副次的な知識、変化した証拠や信念度、排除された可能性、認識された無知、改善された方法や能力、デルセンの意味での促進、合理的に期待される促進、そして後の研究に実際に生じた寄与などを区別する。最後のものは、実際に起きたことを振り返って記録する項目で、見込みを比べる促進とは別に本書が加えた。これらの項目は互いに重なりうる。研究システムが人工であれば、各項目について、システムのうちに実現しただけか、*S* の認識的状况に入ったかを、さらに分けて記録できる（第7節）。結果が *S* の認識的状况に入ることを保有に要する説では、まだ後者だけが、*S* の認識的状况の変化であり、したがって *S* についての進歩の説が数えるものである。項目を分けておけば、評価は、知識に達したかどうかだけでなく、それ以外の認識的な変化も記録できる。しかも、記録したからといって、そのすべてを進歩とみなすことにはならない。進歩の説を一つ採る人は、この記録をもとに、その説が重く見る項目を重く数えて、一つのスコアにまとめることもできる。

9. 認識論の諸説は人工システムによる研究に何を要求するか

第2節から第8節は、人工システムによる研究にとっての帰結を、それぞれの問いが生じたところで引き出してきた。本節は、それらを要求としてまとめ、各要求がどの主張に、したがって誰に掛かるのかを示す。第7節の、誰にも読まれていない、人工システムによる発見の事例は、後者の仕事がなぜ重要かを示している。同じ記録が、*S* の実践に組み込まれた後には、その結果が *S* の知識になることに寄与することも、別の説のもとでは集合的な資源として利用可能であることも、真なる結果を含んでいながら誰の知識でもないこともありうる。したがって、構築すべきものには、結果とその可能な担い手との関係が含まれ、本章で検討した諸説は、知識の場合について、その関係が成功する異なる仕方を示す。何が要求されるかは、なされる主張が、活動が *S* のための研究であるというものか、*S* にとって新しい知識が実現し、それが誰に実現したかというものか、それが *S* の認識的状况に入ったというものかによって変わる（第1章第2節）。全体を通じて、知っているとされる主体は、研究を行った行為者とは限らない（第2.1節）。そのため、ある要求がシステムに掛かるかどうかは、システムがその主体であるかどうか、システムが証拠を集め検証を組み立てたかどうか、そして採用する保有の説によって決まる。

要求は条件つきである。各要求は、その説のもとで知識が何を求めるかを述べ、それと異なる場合には、評価者が帰属を支えるために何を確立する必要があるかも述べる。第5節で述べたように、評価者が知識を帰属させる証拠を十分にもたないことは、それだけでは主体が知識を欠くことを示さない。異なる説の要求は、足し合わせて、あらゆる人工研究者が満たさなければならない一つの覧にするものではなく（第4.2節）、本節はそれらのあいだで選択しない。以下の条件文は知識の主張についてのものであり、理解の主張についての対応する帰結は第3章第9節で扱う。

知っているとされる者が誰であれ、第4節と第6節で比べた諸説は次のことを課す。アクセス可能な理由が要求されるならば、関連する主体がそれらの理由を推論、異議、修正において使用できない限り、事後的な説明は不十分である。証拠への適合が要求されるならば、知っているとされる主体の現に働いている確信度あるいは受容は、その時点でその主体がもつ証拠に適合していなければならない。適合はある時点について評価されるので、証拠が変わっても改められない態度は、適合しなくなりうる。さらに、フェルドマンとコニーのいう十分に基礎づけられた態度（第4.3節）のように、態度が、それが適合する証拠に基づいて保持されることまで求めるならば、現に働いている基礎づけの関係も要求に含まれる。責任ある証拠獲得もまた要求されるならば、主体の証拠を集めるさいになされた探索、測定、省略、停止の決定は、最終的な推論とは別に評価されなければならない。信頼性が要求されるならば、主体の信念を形成または維持する過程は、関連する環境において信頼できるものでなければならない。信頼性を根拠に評価者が帰属を支えるには、関連する環境の全体にわたってそのパフォーマンスが確立されねばならない。安全性が要求されるならば、主体の信念は、あまりにもたやすく偽でありえたものであつてはならず、帰属を支える評価者は、その信念が形成された経路における汚染と近傍の失敗可能性を検討しなければならない。感受性が要求されるならば、主張が偽であったとしても主体がなおその信念を抱いていたかどうかを検討しなければならない。徳信頼性主義的な有能さ（competence）が要求されるならば、成功は、関連するエージェントまたは組織されたシステムの安定した能力に帰属可能でなければならない。責任主義的な徳が要求されるならば、報告だけでなく、主体の探究の遂行そのものが、知的性格と動機づけとして機能する何かを表出しなければならない。無効化への感応性が要求されるならば、反駁の情報と掘り崩し的情報への応答は、報告される結論だけでなく、影響を受けた主張・器具・仮定・妥当性確認者に依拠する主体のコミットメントにまで及ばなければならない。これらのあいだの無効化可能な支持関係を保存する来歴は、その応答を追跡できるようにする一つの仕方である。社会的批判が構成的であるならば、ロンジーノの共同体の水準の意味での批判の受容を欠いた形式的な出版は不十分である。集団の機能性が十分であるならば、人間と人工物からなるシステムであれ、人工物だけからなるものであれ、その集団の境界と組織が、認識的主張の一部となる。この最後の二つは、根拠づけだけでなく保有にもかかわり、第9.4節がそこに戻る。

第4節から第7節を合わせると、さらに次のことが帰結する。人工システムによる研究における知識の主張は、自らの単位を明示しなければならない。単位とは、信頼性が主張される過程の類型と、他の者による選別がそのなかに含まれるかどうか、有能さがどこに座を占めるか（第4.2節）、システムがもつ証拠の境界と、検索の蓄えがそのなかに含まれるかどうか（第4.4節）、知っているとされる集団の境界（第6節）、参照社会 *S*、そして保有の説（第7節）である。人間の研究者であれば、信念をもつ者、証拠の範囲、能力の座は、一人の人であることによっておおむね定まっている。人工システムでは、それらがあらかじめ定まっているとは限らない。たとえばインスタンス、外部の蓄え、ツール、他のシステムがどう結びついているかによって、何通りにも引けることがある。それを引くのが作り手であれ、自らを組み替えるシステム自身であれ、どこに引くかで判定が変わる（行為者の境界については第7章第2節）。ある説が依拠する単位が明示されるまでは、主張をその説に照らして確かめることはできない。以下の小節は、要求をそれが掛かる先ごとに取り上げる。誰が知るにせよ研究の実行者としてのシステムに掛かるもの（第9.1節）、知っているとされる主体としてのシステムに掛かるもの（第9.2節）、システムを通じて知っているとされる他の者に掛かるもの（第9.3節）、結果が *S* の認識的状况に入ることに掛かるもの（第9.4節）、知識に届かない変化に掛かるもの（第9.5節）である。本節の終わりの表2.3は、その結果を単純化した形で示す。

9.1 研究の実行者としてのシステムに掛かるもの

要求のなかには、知っているとされるのが誰であれ、システムが研究を行ったことによってシステムに掛かるものがある。そのどれも、第一の判断、すなわち活動が *S* のための研究であるかどうかの判断の条件ではない。この判断は、結果が、採用する保有の説の求める条件を満たして *S* の認識的状况に入ったかどうかには依存しない（第7節）。実際、*S* にとって新しい知識へ向けて組織された活動は、誰もその結果を知らないうちにも、また失敗したときに

も、研究でありうる（第7節、第8.1節）。逆に、真なる出力があるからといって、その実行が研究になるわけではない。達成への因果的な寄与は、それだけではその達成への試みではないからである（第4.5節）。実行者に掛かるものは、試みがどれだけ適切に遂行されたか、すなわち第8.1節の評価の二つ目と、他の者がそれを通じて知るに至るために何を必要とするかにかかわる。

一つ目は、証拠の収集にかかわる。システムが収集を行ったのであれば、その探索、測定、省略、停止の決定は、最終的な推論とは別に評価される。その評価が認知的なものか実践的なものかは、第5節で示した論争である。カウビネンの見方では、それは実践的なものになるが、システムが証拠を集めるさいに依拠する情報源の信頼性についての支持のない想定は、なお認知的な批判に開かれている。その批判は、システム自身がそうした規範の主体であればシステムに、そうでなければその想定を組み込んだ者たちに向かう。したがってこの評価は、システムが知っているとされない場合にも、収集者としてのシステムに当てはまる。

二つ目は、検証にかかわる。誰が検証を組み立てるかは、誰が知るかは別の軸である（第5節）。システムが他の者の設計したチェックを実行するのであれば、その報告は、何がどの補助仮定のもとで検査されたかを示さなければならない。そうしてはじめて、それを読む者は、結果がどこにかかわるのかを特定できる。システムが自らチェックを組み立てたのであれば、システムを通じて知るに至る者のうち、結果の診断を自ら引き受ける者は、システムがそれをどう行ったかを報告から読み取れる必要がある（第5節）。読み取れる必要があるのは、いくつの候補が生成され、そのうちいくつが偽である見込みが高かったか、どの選択がデータを見た後になされたか、主張を支えるデータがその主張の構成に用いられたかどうか、何がそれに反する証拠に数えられたはずか（第5.2節）、そして形式的に正しい手続きが問いとどう関係づけられたか（第5.3節）である。アクセス可能な理由を求める説のもとでは、受け手には自分で使える理由が要るが、それは結果の確かめ方についての理由でも、当の用途でのシステムの信頼性についての理由でもよく（第5節、第11章第5.2節）、システムの報告はその源の一つになる。過程の信頼性で足りるとする説のもとでは、受け手はこれらを読めなくても知りうるが、その場合にも、帰属を支える評価者には、システムの検証の仕方が信頼できることについての証拠が要る。システムがさらに、統御的な基準そのものを組み立て、改めるかどうかは、自律性の問いであって（第7章第4.3節）、誰が知るかは別の軸であり、知ることの条件には含まれない（第5.3節、第8章第1.3節）。人間の *S* のための研究であれば、継承された人間の基準が役立ちうる。

三つ目は、チェックの独立性にかかわる。批判者が誤りの検出に寄与するのは、それが産出者の誤るところで誤らない限りにおいてである。そのため、チェックが産出者と同じ基盤システムのインスタンスや、同じところで誤りうる別のシステムによって行われる場合には、報告は、それらがどこまで独立であるかを示す必要がある（第5節、第11章第3節）。このことは、*S* が人間であれ人工であれ成り立つ。また、そうしたインスタンスのあいだの一致は、結果を真にしない（第6節）。

四つ目は、システムに帰される主張の単位にかかわる。信頼性の主張は、その過程の類型と、それが成り立つ分布とを述べなければならない。そして、人間による選別のようなシステムの外でなされた選択を、システムについての主張のなかに隠してはならない。そうした選択は、出力を安全なものにしながら、その成功をモデルに帰属可能なものにはしないからである（第4.2節）。パイプライン全体に依拠する受け手にとっては、人間であれ人工であれ、モデル単独では信頼できなくても全体は信頼できることがある。要求されるのは、選別がないのではなく、単位が明示されることである。

答責、つまり出力について問われたときに答える責任はさらに別の事柄であり、これも誰が知るかに従うとは限らない。システムの出力について答えるべき者は、知る者のなかにいるとは限らない。システムの報告が主張ではなく器具の読み取りとして扱われるなら、第11章第5.1節で紹介するゴールドバーグの説を当てはめると、依拠は設計者や作り手の認知的な仕事へと向かう。ただし、本質的に不透明なシステムについて、その設計者や作り手が、働きを理解したうえで責任を引き受けられるかどうかは、コスキネンが論じるように争われている（[Koskinen 2024a, 464–465](#); 第11章第5.3節）。

9.2 システム自身が知っているとされる場合

システム自身が知っているとされる場合、上の条件文はシステム自身の状態に掛かる。表2.1の偽陽性はこの場合を例にとっており、そのいくつかは、第2.1節の論点と一つの形を共有する。システムが表に示すものと、その出力と修正を実際に左右しているものとが食い違う、という形である。保存された文は、推論、行為、修正への安定した統合に比べれば、信念の弱い証拠にすぎない（第2.1節）。事後的な理由づけは、実際に作動した経路と無関係でありうる。報告のなかの留保は、探究の遂行を変えないままでありうる。そして、訂正された報告は、作動しているコミットメントをもとのままにしておきうる（表2.1、第4.3節）。これらの事例を合わせると、この主張が必要とするものの多くは、システムが示すものについての証拠ではなく、システムを実際に左右しているものについての証拠だと考えられる。

本章は、いくつかの関係について、その証拠が何であるかを述べてきた。態度については、システムの状態が採用する説の信念の条件を満たすか、その形の信念を要求しない説が擁護されなければならない（第2.1節）。また、ある状態は信念でも、受容でも、信念度でも、想定でもありうるので、信念の条件を当てはめる前に、それがどれであるかが定まっていなければならない（第2節）。基礎づけについては、そうした条件を課す説のもとでは、システムの作動しているコミットメントが、それが適合する証拠に基づいていなければならない。結論とその根拠が同じ場所にあることはそれを確立せず、どの項目がコミットメントを左右したかを記録する来歴は、基礎づけの因果説に最も直接に適合する仕組みである。ただし来歴は帰属の証拠を残す仕組みであり、記録がないことは、それだけでは基礎づけがないことを示さない（第4.3節）。証拠については、信念の機能的な説のもとでは、システムが自らの蓄えの内容を知っていると認めるには、それが検索可能であるだけでなく、推論と修正に統合されていることの証拠が要る（第4.4節）。無効化と検証については、報告だけでなく、システムの作動しているコミットメントが、反駁の情報と掘り崩し的情報に適切に応じ、検証が促す修正に応じて変わらなければならない（第4.3節、第5節）。ウィリアムソンの議論に立てば、証拠を加えることしかできない更新は、証拠を失うことを表せない（第5節）。シモンの知識第一の説では、システムが訓練されたものであれば、それが知識を生成する機能をもつかどうかは、現在の正確さだけでなく、その訓練と選択が何に報いてきたかにも部分的にかかる（第4.1節）。そして、証明、観察、失敗した予測、批判によってコミットメントが改められうるシステムは、文章としてのものともらしさや社会的な承認だけに向けて最適化されたシステムとは、対象に対して異なる仕方に関係している（第3節）。

システムが *S* の成員でなければ（成員である場合は第9.4節で扱う）、*S* が人間であれ人工であれ、知っているシステムは *S* の外の担い手である。その知ることは、*S* にとって新しい知識をどこかで実現する。第二の判断はこれで成り立つが、それだけでは第三の判断、すなわち結果が *S* の認知的状況に入った

という判断は成り立たない。そしてその主張は、その達成が S に対してどのように新しく、 S のために求められたものであるかについての説明を必要とする（第8章第1節）。したがって、システムが知っているという主張は、システムが S のための研究を行ったという主張とも、結果が S の認識的状況に入ったという主張とも別のさらなる主張であり、後者と違って本小節の負担を負う（第8章第1.1節）。

これらの要求を当てはめるにあたっては、さらに二つの点がかかわる。第一に、条件どうしが食い違う判定を返しうる。ウィリアムソンは、安全性と感受性が分かれると論じる。安全性は感受性を含意せず、知識は感受性を要求しない。第4.1節で見た、水面から1メートルの高さにある印の例が示すように、ごくわずかな系統的な偏りでも、偶然的な命題についての主体の信念のほとんどすべてを、知られたまま感受的でないものにしうる（Williamson 2000, 147, 159–161, 307–310）。評価にとってこれは、わずかな系統的な偏りをもつシステムに感受性の検査を当てると、安全性にもとづく説ならシステムの知識に反するとは数えない失敗が報告されうる、ということである。

第二に、条件文が挙げる選択肢は、人工システムによる研究が人間の研究者の用いる内的経路を再現することを要求しない。人間とは異なる仕組みをもつシステムも、まったく異なる経路を通じて、信頼でき、安全で、理由に応答的な、あるいは社会的に統合された関係を実現しうる。方法が違うというだけでは、どんな認識的基準も適用されないとは推論できない。立証責任は、人間との類似から、代替的な実現が関連する説を満たしているという証拠へと移る。説を名指さなければならない理由の一つは、ここにある。

9.3 他の者がシステムを通じて知っていると考えられる場合

人間、あるいは産出者以外の人工システムが、システムを通じて知っていると考えられる場合には、システムが自らの出力を信じている必要はない。たとえばAIは、自らの出力を文字どおりには信じることなくし、人間の知識に貢献しうる。そのとき条件文は、知っていると考えられる者の信念に、その者が人工であればその作動しているコミットメントに適用され、システムの出力は、証拠として、あるいはそれらを形成した過程の一部として、それらにかかわる（第4節、第8章第1.1節）。そのため、同じ出力が、ある受け手には知識の根拠となり、別の受け手にはならないことがある。AIが提案した仮説を、その仮説が偽であれば失敗したであろう独立の検査を経てはじめて受け入れる人間の研究者は、AIが仮説の真偽にかかわらずそれを提案したであろうとしても、感受的な信念を抱きうる。逆に、主張が偽であったとしても出力がその主張を述べていたのであれば、言い回しが違っていたかどうかにかかわらず、その出力だけに基づく信念は感受的でない。したがって、システムを通じて知る者に知識を帰属させることは、システム単独についての帰属ではなく、その者の信念と、出力をその一部とするその信念の拠り所とについての帰属である。第4.3節がシステムの側に見いだした基礎づけの隔たりは、受け手の側にも現れる。説得力のある報告を読むことは理由を与えるが、信念がその理由に基づくのか、システムの威信に基づくのかは、さらに別の事柄である。

この場合にシステムが何を供給する必要があるかは、受け手について採る説によって異なる。アクセス可能な理由を求める説では、受け手が推論や異議に使える理由が要る。その理由が結果の確かめ方についてのものであれば、報告が見えるようにするものが中心になる。一方、第9.1節で見たように、理由が当の用途でのシステムの信頼性についてのものであれば、報告はその源の一つになる。過程の信頼性で足りるとする説や、第4.2節で見たソーサの説を当てはめた場合には、受け手は報告の全体を読めなくても知りうる。そのときシステムに掛かるのは、報告が読めることより、その過程が実際に信頼できることであり、評価者が帰属を支えるには、それについての証拠が要る。第9.1節の、報告が示す必要のあるものについての要求が受け手の側から来るのは、主に、受け手が結果の診断を自ら引き受ける場合と、受け手の理由が結果の確かめ方についてのものでなければならない場合であり、それ以外は、評価者が帰属を支えるために必要とするものから来ている。どの場合も、これらの要求は、システムが知っているからではなく、システムが研究を行い、チェックを実行あるいは組み立てたから、システムに掛かる。

受け手が人工であれば、第9.2節の要求は受け手へと移る。受け手が信じているかどうか、その作動しているコミットメントが出力に基づき、無効化のもとで改められるかどうか、である（第4.3節）。こうした受け手に移すと、感受性の例は、第11章第3節が検討する独立性の問いにつながる。出力のチェックを、産出者自身の基盤システムのインスタンスや、産出者と同じところで誤る別のシステムで行い、出力が真であろうとなかろうとそれを受け入れたであろう受け手は、どれほど多くのインスタンスが一致しても、感受的でないコミットメントを抱いている。

他の者が知る者である場合にも、別の関係が、システム自身が知ることを問題にしうる。システムの報告が主張として受け取られるなら、主張の説は話し手としてのシステムに条件を課す。ただし、現在の出力が主張であるかどうか、主張の規範が知識を要求するのか、合理的な信念を要求するのか、文脈によって変わる支持を要求するのかは、争われている（第11章第5.1節）。また、求められる達成が技能のようにシステムのうちに実現するものであれば、システムがその最初の担い手である（第6節）。

9.4 結果が S の認識的状況に入ったと考えられる場合

第5節の検証の図式は、知っていると考えられる者のところで終わる。 S 自身が知っていると考えられる主体でない限り、 S の認識的状況に入ることは、その図式に含まれないさらなる一步であり、それが何を要求するかは、 S と、採用する保有の説によって決まる（第6節、第7節）。構築しなければならないのは、結果が妥当性確認、推論、行為、訂正に入っていける経路であり、それは研究を行ったシステムの内部にあるとは限らない（第7節）。そうした経路で足りるかどうかは、説によって異なる。誰も読んだことのない記録に広げたバードの説では、足りるかもしれない。ラッキーの説では、実働成員をもたない社会全体は集団的信念をもたないので、結果は S の成員に届き、信じられなければならない。集団が知ることを通じて入ったと主張するなら、そのとき信じなければならないのは、 S のうちの集団の実働成員である。ロンジーノの説では、その条件のもとで批判を経て、成員に受け入れられなければならない。そして、知る主体なき知識のギアリーによる翻案では、信頼でき受け入れられた手続きによって認定され、公に受け入れられなければならない（第6節、第7節）。したがって、結果が S の知識になったという主張は、自らの依拠する説明を十分に明らかにしているときにはじめて評価できる（第6節）。休眠した記録と生きた認識的能力の区別は、記録を生きた能力にするものを名指すが、それが知識の境界にあたるのは一部の説においてだけである。また、ラッキーの行方不明の子どもの事例をめぐる論争では、集団がある結果を知っているかどうかと、それを知っているべきだったかどうかとが切り離され、記録を照合するシステムをもつ集団は、まだ知らない結果について答えを求められうる（第7節）。

S が人工であれば、これらのどれよりも先に生じる問いがある。 S の二つの役割はどちらも、 S の認識的状況が求められる知識を含みうることを前提にしており（第1章第2節）、ここではその状況は人工の成員、その組織、その記録からなる。そのため、そうした成員が知る者でありうるかという問いは、どの成員もまだ結果を知らない段階でも、また A がその一つであるかどうかにかかわらず生じる（第6節）。そのうえで諸説は、成員については、第9.2節がシステムに問うことを問う。結果を信じているか、あるいはその共同受容に加わりうるか、である。成員の組織については、それが適切に組織された認知システムを実現しているか、バードの説ではさらに、その組織が何のために選択され、あるいは設計されたのか（第7節）、その相互作用がロンジーノの説の求める多様性と公共的な批判の基準を伴っているか、人工の成員による認定が、成員がひとつの基盤システムのインスタンスである場合も含めて、認定に数えられるか、を問う（第6節、第8章第1.3節、第11章第3節）。作業的特徴づけには、成員が人間であるという条件は組み込まれていないが、人間以外の主体性は、協調や複数性だけによって確立されるわけではない。第9.2節がシステムについて述べたように、それらの関係は、人間の共同体におけるのと同じ仕方で実現される必要はない。示さなければならないのは、その実現が説を満たしていることである。

産出したシステム自身が、人工または混合の S の成員であることもある。産出したシステム自身が知る場合のうち、その知ることが第三の判断の役を果たしうるのは、この場合だけである。個人保有説では、システムが知っていれば S の成員がすでに知っていることになり、結果は産出された時点で S の認識的状況に入っていることがありうる。ラッキーの説では、このことが集団について成り立つのは、システムがその実働成員の一つであり、実働成員のかなりの割合が結果を信じている場合に限られる。ロンジーノの説では、システムが受け入れることはたかだか一つの成員の受け入れである。バードの説では、判定はなお、組織されたシステム全体にかかる（第7節）。混合の S については、入ったという主張は、結果が人間と人工のどの成員に届いたか、保有、記憶、訂正がそのあいだにどう分配されているかを述べなければならない（第8章第1.2節）。

9.5 知識に届かない変化

知識の説からではなく、第8.2節がヴィルホルトの説のもとで述べた進歩についての主張から来る要求がある。無知を明示することが認識的進歩に数えられるならば、評価はまた、答えの不在と、表象され行動に結びつけられる未知とを区別し、その未知が誰にとって行動に結びつくのか、つまりシステム自身のその後の探究にとってなのか、その仕事を引き継ぐ S の成員（人間であれ人工であれ）にとってなのかもしなければならない。そのとき、答えの量と、問いを閉じた割合は、根拠づけられた判断保留、現に生きている代替案、あるいは問いを解決するためになお必要な証拠を消し去る場合には、研究の質と逆相関しうる。ウィリアムソンの議論が成り立ち、システムの取り出しにかかる時間に上限がなく、欠けていることとまだ届いていないことを区別できないなら、それはここで求めうるものを限る。システムが知らないことについての記録が完全であることは、システム自身のチェックによっては保証できないので（第8.2節）、この要求がかかわるのは、何が誰にとって表象されているかであって、完全さではない。

進歩の説を選ばない評価は、第8.3節の記録を保つ。変化を種類ごとに分け、研究システムが人工であれば、各変化がシステムのうちに実現しただけか、 S の認識的状況に入ったかでも分けた記録である。一つのスコアが進歩そのものを表すと主張することは、それだけで進歩の説のどれかを選ぶことになる（第8.3節）。

表2.3は、本節の要求がどこに掛かるかを示す。その行は、第1章第2節と第8.1節で区別した主張であり、システムが知っているという主張と、他の者がシステムを通じて知っているという主張とを分けてある。要求がいちばん大きく異なるのは、この二つのあいだからである。最後の列は、各主張が要求しないものを述べる。これは表2.1と表2.2が暗黙のままにしていることである。この表は単純化した見通しであり、各欄は説によって異なる要求を圧縮している。また、システムがたとえば S の成員でありうるように、行は互いに排他的ではない。表2.1と同じく、第二列の記載は該当する説のもとでだけ成り立つものであり、あらゆるシステムが満たさなければならない一つの一覧にまとめるものではない（第4.2節）。

表2.3 要求はどこに掛かるか（単純化した見通し）

主張	成り立たなければならないこと（主張が説に依存する場合は、採用する説のもとで）	評価者が見て取れる必要があること	その主張が要求しないもの
活動が S のための研究である	S にとって新しく、 S が持ちうる達成へ向けて組織された方向づけ	試みの組織と参照社会 S	知識に達したこと。システムが結果を知っている、あるいは信じていること
誰が知るにせよ、システムがそれを適切に遂行した	責任ある証拠獲得。実行あるいは組み立てたチェックに見合う報告	証拠がどう集められたか。何がどの仮定のもとで検査されたか。いくつの候補が生成されたか。チェックがどれだけ独立であったか。システムに帰される主張の単位	システムが統御的な基準を組み立てたり改めたりすること。これは自律性の問いである（第7章第4.3節）
システム自身が結果を知っている	その説が信念の条件を課すならその状態がそれを満たし、作動しているコミットメントがその説のさらなる条件を満たすこと	システムが表に示すものだけでなく、その出力と修正を実際に左右しているもの	人間の経路をたどること。結果が S の認識的状況に入っていること
他の者がシステムを通じて知っている	その者の信念が、その者が人工であれば作動しているコミットメントが、出力を証拠あるいは過程の一部として、採用する説を満たすこと	その者の信念が何に依拠しているか。システムから独立なチェックがあったかどうかを含む	システムが出力を信じていること（ただし、報告が主張として受け取られるなら、主張の説が話し手としてのシステムに条件を課し、達成がシステム自身の技能であれば、システムがその最初の担い手である）
結果が S の認識的状況に入った	妥当性確認、推論、行為、訂正への経路と、採用する保有の説がそれに加えるもの	どの説に依拠しているか。結果がどの成員や組織に届いたか	経路がシステムの内部にあること。システムが知る者であること。ただし、システムが S の成員であり、その知ることが数えられると主張される場合を除く
S が人工または混合であれば、	S の成員や組織がその達成を持ちうること。それらに当てはめた保有の説	成員のあいだの独立性と多様性（成員がひとつの基盤システムのインスタンスである場合を含む）。	それらの関係が人間の共同体と同じ仕方で実現されること

主張	成り立たなければならないこと（主張が説に依存する場合は、採用する説のもとで）	評価者が見て取れる必要があること	その主張が要求しないもの
さらに		集団の組織。混合の S では、結果がどの成員に届いたか	
知識に届かない変化	種類ごとに記録された、表象された限界やその他の変化	それが誰にとって表象され、行動に結びつくか。 S の認識的状况に入ったか	その変化が知識であること。一つの進歩スコア

表2.3の要求を説ごとに見比べると、比較した説のあいだに、共通する帰結も、分かれる帰結もあることが分かる。共通するものの一つは、どの説に照らすにせよ、人工の構成についての知識の主張は、その説が依拠する単位をどこに引くかを定めてはじめて確かめられることである。ただし、単位が必要になる理由は説ごとに異なる。たとえば証拠主義では、何がシステムのもつ証拠かが、信頼性主義では、どの過程の種類の信頼性が問われるかが、有能さを求める徳の説では、有能さがどこに座を占めるかが、単位の引き方で決まる（第4.2節、第4.4節）。分かれるものの一つは、システムを通じて知る受け手が自分で使える理由をもたなければならないかどうか、もたなければならないとすれば、その理由が検証そのものについてのものでなければならないかどうかである。アクセス可能な理由を求める説では、受け手は推論や異議に使える理由をもたなければならない。しかしその理由は検証についてのものに限られず、システムの信頼性についてのものでもよく、報告はその源の一つである（第9.1節）。これに対して、過程の信頼性で足りるとする信頼性主義のもとでは、またソーサの説を当てはめた場合にも、受け手は報告を読めなくても知りうる。受け手が検証の組み立て方を報告から読み取る必要があるのは、その理由が検証そのものについてのものでなければならない場合と、受け手が結果の診断を自ら引き受ける場合である。そのため、報告が何を示すかについての同じ要求も、そうした場合には受け手のために、それ以外では評価者が過程の信頼性を確かめるために課されることになる（第9.3節）。

知識は研究の達成を汲み尽くさないが、要求度の高い目標である。比較した説のもとでは、新しさ、真理、合理的な受容、信頼できる成功、理解、社会的保有、研究能力は異なる主張に属し、その要求は異なる当事者に掛かる。そのため、採る説にもとづく論証なしに、そのどれかから別のどれかを推論することはできない。人工の構成を通すと、これらの区別は具体的な帰結をもつ。証拠主義のもとでは、検索された文書だけをシステムのもつ証拠に数えるなら、証拠の集め方が拙くても結論は正当化されうる。他方、蓄え全体を数えるように境界を広げれば、そのなかの反対の証拠もシステムのもつ証拠になるため、同じ結論が正当化を失いうる（第4.4節）。基礎づけを求める説のもとでは、よい理由が報告に添えられていることと、結論がその理由によって保たれていることは別であり、訂正された報告は、作動しているコミットメントを変えないままでありうる（第4.3節）。偶然の生成器は、知らず、研究も遂行しないまま、他の主体の知識に寄与しうる（第4.5節）。候補を大量に生成するシステムでは、チェックへの合格が主張への支持を上げても、主張を確からしくするとは限らない（第5.2節）。そして、人間の誰も読んでいない結果が S の知識であるかどうかは、バード、ラッキー、ロンジーノの説で分かれるだけでなく、機械が知る成員に数えられるかどうかによっても変わる（第7節）。記憶、更新、成員の配置は、知識を運ぶ外側の仕組みであるだけでなく、採る説によっては、知識の成立そのものにかかわる。見かけ上新しい結論を数えるだけでは、こうした違いはどれも見えない。

ここで一つの問いが避けられなくなる。人工システムが、共同体が使える、信頼できる仕方で妥当性確認された真なる予測を産出する一方、予測された関係がなぜ成り立つかを共同体のどの人間の成員も説明できず、システムもその背後の依存関係についての使用可能な表象を何も提供しないとしよう。本章では、知識がもたらされたかを問うてきた。知識がもたらされたとしても、それで意図された目標に達するかは、探究が何を求めたかによる。次章は、理解はいつ、誰において達成されるのかを問う。すなわち、何が理解に数えられるのか、理解は知識とどう関係するのか、システム自身も含めて誰が理解をもちうるのか、そして両者が切り離されるとき何が帰結するのか、である。

第3章 理解はいつ、誰において達成されるのか

知識、説明、そして人工システムによる不透明な探究

1. 研究が求めうる達成としての知識と理解

作業的特徴づけは、研究を認識的達成へと向けるものであり、本書はその達成を主に知識と理解を通じて検討する（第1章第1節）。第2章は知識を取り上げた。本章は理解を取り上げる。何が理解に数えられるのか、理解は知識とどう関係するのか、誰が理解をもちうるのか、そして、求める達成が理解であるとき、知識が理解なしに得られたとき、あるいは、説によっては、理解が新しい知識なしに得られたときに、人工システムによる研究にとって何が帰結するのか、である。本章は二つを、研究が別々にも、あわせても求めうる達成として扱い、両者がどう異なり、いつ切り離されるのかを問う。

前章は、両者が切り離されうる、意図的に強めた境界事例で締めくくられた。本章ではこれを、強化した予測の事例と呼ぶ。ある人工システムが、これまで知られていなかった規則性を同定する。独立した証拠が、特定された領域の内部でその主張を支持する。関連する共同体はその結果を使用し、その不確実性を追跡し、証拠が変化すればその結果への依拠を修正することができる。それでもなお、その規則性がなぜ成り立つのかを共同体のどの人間の成員も説明できず、システムはその規則性の原因となっている依存関係についての使用可能な表象を何も提供しない。

この事例は、結果が正確であると繰り返すだけでは答えられない問いを生み出す。正確さはそれ自体では知識を確立しない。しかし、前章で採用した説のさらなる条件が満たされていると仮定しよう。そのとき、理解はどこかに実現しているのだろうか。システムのうちに（その状態が理解にあたるかどうかは定まっていない。第6節）、ある人間のうちに、それとも共同体のうちに。試みが、その規則性がなぜ成り立つのかの理解を求めているのなら、規則性が成り立つという知識は得られたにしても、試みは目標に達したのだろうか。この種の知識は理解を要し、ここでは両者が切り離されえないのだろうか。そして、共同体がより多くを予測できるようになったがより多くを理解してはいない場合、共同体は進歩したと言えるのだろうか。活動が共同体のための研究であるかどうかは、これらの答えに左右されない。研究は、成功によってではなく試みによって特徴づけられるからである（第1章第3節）。

これらはひとつの問いの異なる定式化ではない。それらは異なる評価対象に関わっている（第2章第8.1節で区別した評価を参照）。

1. ある活動が、作業的特徴づけの意味で S のための研究であるかどうか。
2. ある出力が知識の地位を達成したかどうか、誰がそれを知っているのか（誰も知らないこともありうる）、そしてそれが、そのために求められる社会 S の認識的状況に入ったかどうか（第1章第2節、第2章第6節）。
3. 人間、人工システム、集合体のいずれかの主体が理解しているかどうか、そしてその理解が S の認識的状況に入ったか、 S にとって利用可能になったかどうか（第6節）。
4. より広いプログラムが進歩したかどうか。
5. 継続している研究システム（人工システム単体であれ、探究を時間にわたって担う人々、制度、人工の構成要素の全体であれ）が、批判、転移（結果を別の条件や問題へ持ち込むこと）、診断、訂正に必要な能力を保持しているかどうか。
6. その結果が、関連する目的と影響を受ける当事者に相対的に、重要であるか価値をもつかどうか。

失敗した調査は、求めた達成を産出することなしに、研究でありうる。共同体は、ある規則性が成り立つことを、なぜ成り立つのかを理解することなしに知りうる。対象のいくつかの特徴を誤って表象するモデルが、それにもかかわらず理解を改善することもありうる。局所的に根拠づけられた結果が、ますます均質化し、自らの誤りを検出する能力を失いつつある研究生態系に入り込むこともありうる。これらの判断は相互に作用するが、いずれも他の判断の代わりにはならない。

これらの区別は、認識的達成への一般的な方向づけが残す仕事を特定する。理解が知識と異なるなら、どちらを求めるかは目標を特定することの一部である。ある知識が理解を要するなら、そのことは両者の分離を制約する。作業的特徴づけのもとでは、研究そのものにも、個々の研究行為にも、知識であれ理解であれ、何らかの認識的達成をもたらすことは要求されない（第1章第3節）。ただし、試みの目的が理解であるとき、知識が得られたことだけでは、その試みが目標に達したことは示されない。第4節と第7節はこれらの選択肢を検討し、第9節はそれらが人工システムによる研究に何を要求するかを問う。

本章は、次の順で進む。第2節は予測と説明を分ける。第3節は主要な説が何を理解に数えるかを並べ、第4節はそれらを強化した予測の事例に当てはめて、そこで知識と理解が切り離されるかを問う。第5節は、何が、どの役割で、誰にとって、どれだけの利害のもとで不透明なのかを特定することで、システムの不透明性が、その結果の根拠づけを損なうか、理解を締め出すかを問う。第6節は、誰が理解するのか、すなわちシステムか、それを使う者か、集合体かを問う。第7節は、理解が何のために価値をもち、時間を通じて何を可能にしようかを問い、第8節は、局所的な達成が継続する探究の能力を決めない理由を示す。第9節は、理解の諸説が人工システムによる研究に何を要求するかを問い、その答えが、主張される達成、その担い手、その用途に応じてどう変わるかを示し、最後に本章の問いに立ち返る。

2. 予測と説明は関連しているが同一ではない達成である

この問題へ入る単純な道筋の一つは、探究が答えうる問いを区別することである。予測は、何が起こるか、どの値が予期されるか、あるいはどの結果がある事例に属するかを特定する。予測は未来に関わる必要はない。過去予測（retrodiction）、分類、観測されていない性質の予測も、同じ機能的形式をもち

うる。説明は、ある現象がなぜ起こるのか、あるいはそれがどのように、メカニズム、原因、制約、統一的パターン、理由、その他説明的に関連すると判断される関係に依存しているのかを扱う。

一方における成功は、他方における成功を論理的に保証しない。モデルは、安定した代理指標を利用することで対象を予測しつつ、研究者が理解したいと望む依存関係については何も明らかにしないことがありうる。説明的モデルは、初期条件が知られていないために、メカニズムを同定しつつ貧弱な点推定しか産出しないことがありうる。高度に理想化された表象は、予測の精度を犠牲にしながらパターンを理解可能にすることがありうる。したがって予測と説明は、科学的成功のただひとつのスカラー尺度に付けられた二つの名前ではない。

また、それらは対立する種類の人工物でもない。説明は新しい予測を生み出しうる。予測の失敗は説明のあいだを識別しうる。条件が違っていたら何が起こるかという多様な問いに正しく答えること（豊かな反事実的パフォーマンス）は、モデルが関連する依存関係を捉えていることの証拠を提供しうる。いくつかの領域では、同じモデルが両方に寄与する。ここで揺るがない主張は、予測的な問いに答えることが、その事実だけによって説明的な問いに答えることにはならない、ということだけである。

二つの達成がどう関係するかについての立場は、これよりも大きく分かれる。理解についての説を第3.1節で検討するデ・レフト (de Regt) は、理解なしの予測は科学にとって不十分であるだけでなく、実際には不可能だと論じる。デ・レフトの説では、理論からの予測が成功するには、その理論が使い手にとって理解可能 (intelligible) でなければならず、理論の定性的な帰結を認識する能力は、逆にその理論が理解可能であることを示す (de Regt 2017, 91–92, 107)。ここで求められているのは理論の理解であり、デ・レフトはこれを理論の理解可能性と呼ぶ。行動主義心理学や量子物理学の事例で示すように、理論を具体的な現象に適用するにはこの理解可能性が要る、と論じる。この説では、これは現象の科学的理解とは区別され、後者はさらに説明を要する (第3.1節)。のちの共著の章で、フロリアン・ボーゲ (Florian Boge) とデ・レフトは、深層ニューラルネットワークを理論駆動ではなくデータ駆動のものとし、一見したところ、それは予測力には秀でうるものの、理解可能な理論を欠くために理解をほとんどもたらさない、と書いている (Boge and de Regt 2025, 419–420)。この共著の章は、理解なしの予測は実際には不可能だという、デ・レフトの以前の主張には触れていない。ただし、両者は衝突するとは限らない。以前の主張は理論からの予測の成功についてのものであり、深層ニューラルネットワークは、理解可能な理論を欠くものとして描かれているからである。ポトチュニク (Potochnik) は、科学の目的は互いに緊張関係にあると論じる。ひとつの目的に適した道具は他の目的にはあまり適さず、それはこの見方では、現象の複雑さと、人間の認知と行為の限界から生じる。たとえば、競合する仮定をもつ気候モデル群にわたる頑健性分析は予測を生み出すために用いられるが、どのモデルも、気候への因果的影響を正確に表象することも、気候変動を説明することも期待されていない。ポトチュニクは、予測の成功が説明を促進することも、その逆も、期待すべきではないと結論する (Potochnik 2015, 16–21)。この説では、これらの目的は、人間がこの複雑な世界を把握し制御することを進める点で価値をもち、その追求は人間の限界に相対的である (19–20)。ポトチュニクは人工の探究者を論じていないが、人工の探究者が行う研究に当てはめれば、現象の複雑さは残る。一方、その緊張がどの限界に応じるかは、研究が誰の把握を進めるためのものか、つまりシステム自身のものか、人間の *S* のものかによる。

深層ニューラルネットワーク（相互に結合した多数の中間層によって入力を入力へ写す機械学習モデル）について、クラウス・バイスパルト (Claus Beisbart) は、訓練済みのネットワークから予測を得た科学者が、どのような条件のもとで、予測されたことがなぜ成り立つのかを理解したことになるのかを問う。バイスパルトは、この理解を説明的理解と呼ぶ。そして、その条件を、知識を正当化された真なる信念とみる伝統的な分析に並行させ、複数の説明の理論がそれぞれの仕方でも身を埋められるように広く述べる。この説では、科学者はまず説明をもっていなければならない。ここでの説明とは、法則を含む、原因やメカニズムを挙げる、統一をもたらしといった追加の条件を満たし、しかも無関係な情報を含まない論証である。さらに、その説明は正しくなければならない、科学者はそれを受け入れ、その受け入れについて正当化をもち、その説明を用いて推論できなければならない (Beisbart 2025, 5–8)。これらの条件は、バイスパルトが「単純な設定」と呼ぶ場面に当てはめられる。この場面では、科学者はネットワークの訓練の仕方を知っており、入力と出力を解釈できるが、ネットワークがデータに何をしているかを知っている必要はない (8)。バイスパルトの中心的な議論は、第一の条件に関わる。ネットワークの計算は、入力と、ネットワークが実現する入出力関数が対象となる系について成り立つという仮定とから、出力へ至る論証として再構成できる。しかし、ネットワークについてそれ以上のことが知られていない限り、この論証が法則に依拠しているのか、原因やメカニズムを特定しているのか、統一をもたらしのかは、何によっても示されない。これに対して、流体の運動を記述するナビエ–ストークス方程式を前提として用いる計算機シミュレーションは、自然法則を含むことが知られている (10–11)。そのため、科学者は、ネットワークが説明を含むかどうかを判断できない (13)。この議論によれば、ほかの条件もこれに伴って満たされない。成功した予測の実績は、ネットワークの入出力関数が正しいと受け入れることを正当化しうる。しかし、科学者は、知らない説明を正当に受け入れることはできず、述べることでできない説明を用いて、理解に必要な推論を行うこともできない (14, 16)。こうしてバイスパルトは、この設定では、科学者は予測されたことがなぜ成り立つのかを通常は理解していないと結論する (24)。

この議論は、強化した予測の事例に密接に関わるが、その事例を立証するものではない。たしかに、バイスパルトの議論は、予測器への正当化された依拠が説明的理解の欠如と両立しうることを示し、その場合にどの条件が満たされないのかも示す。しかし、その問いに関わるのは、規則性ではなく、個々の予測された出来事や状態である。また、その結論は、そうした理解が不可能だということではなく、通常は欠けているということである。さらに、その議論はネットワークを人間が用いる道具として扱い、ネットワーク自体が理解するかどうかは問わない (4, 9, 18)。バイスパルトの説でネットワークがなお理解に寄与しうる道筋には、第4節で立ち戻る。予測と説明のより一般的な区別は、以下で検討する説明と理解の諸理論に委ねられる。

予測的な問いに答えても、それだけで説明的な問いに答えたことにはならないという論点は、普遍的な序列づけではなく、問い相対的なものである。探究が、どの化合物が標的に結合するかを問うのであれば、メカニズムが未知のままであっても、領域を限定された予測的な答えがその問いを満たしうる。探究が、なぜその化合物が結合するかを問うのであれば、同じ出力は不完全である。探究が、証明、解釈、構成、あるいは介入を求めるのであれば、予測の正確さは、求められている達成ではないにせよ有用でありうる。研究の作業的特徴づけがこれらの探究を含むのに十分なほど広いのは、まさにそれが研究を予測と同一視しないからである。

予測そのものも、いくつかの別個の主張に分かれる。ひとつの出力が正しいという証拠は、ある規則が参照クラスにおいて信頼できるという証拠に自動的になるわけではない。ひとつの分布における信頼性は、分布シフトのもとでの転移を確立しない。観察的予測は、それだけでは介入の結果を確立しない。

モデルと対象の相関は、モデルが説明的依存関係を表象していることを確立しない。信頼性はまた、開発、展開、その後の社会技術的適応のあいだで変化しうる (Grote, Genin, and Sullivan 2024)。

ウッドワード (Woodward) の因果的説明の理論は、信頼性と転移とのあいだの隔たり、そして観察的予測と介入とのあいだの隔たりを、関係の異なる使い方がそれぞれ何を要求するかに帰する。ウッドワードは、ある変数によって別の変数がより予測しやすくなっても、前者に介入すれば後者が変わるとは限らないと論じる (Woodward 2014, §7)。またケネス・ケンドラー (Kenneth Kendler) との共著では、予測の関係が他の集団や状況でもどこまでよく予測し続けるかは、因果的でない関係についても問える問いだと論じる (Woodward and Kendler 2023, 14)。ウッドワードの説では、二つの表象が、これまでに観察された相関については一致しながら、介入のもとで何が起こるかについては食い違うことがある。そのため、受動的な観察によって二つの変数が相関していると知っても、それだけでは、一方を変えれば他方が変わるかどうかは分からない。相関が共通原因から生じている場合には、一方の変数への介入が、その相関を生んでいた因果的なつながりを断つので、相関はもはや成り立たない (Woodward 2023, §5)。

大塚 (Otsuka) は、観察的予測と介入とのあいだの隔たりに、統計的な形を与えている。大塚の説では、推測統計は、確率モデルによって記述されるデータの背後の母集団が、観察された事例と観察されていない事例にわたって変わらないと想定するのに対して、介入は母集団の構造そのものを変えてしまう。そのため因果的な効果を扱うには、さらなる存在論的な前提が必要になる。構造的因果モデルの方法では、母集団を生み出す因果構造は確率分布だけでは決定できない。そこで、マルコフ条件や忠実性（データに現れる独立性はすべて因果構造から来ており、偶然の打ち消し合いによるものではないという仮定）のような仮定が、因果構造と確率分布をつなぐ橋として想定され、それによって介入の結果が評価され、統計的なデータから因果構造が推測される (Otsuka 2018, 188–189)。

したがって第1節の強化した予測の事例は、何が知られているのかを特定してはじめて評価できる。上で引いた区別は、その候補を分ける。共同体は、ある規則性が条件 C の内部で成り立つことを知っているのかもしれないし、ある予測器が参照クラス R において誤差率 e をもつことを知っているのかもしれないし、ある介入が管理されたプロトコルのもとで一定の結果をもたらすことを知っているのかもしれない。ひとつのベンチマークが正確だというだけで、共同体が説明されていない普遍法則を獲得するわけではない。どの命題がどの範囲で知られているのかを特定することは、どの対象とどの種類の理解が主張され、あるいは欠けているのかを名指すこと (第4節) とあわせて、知識と理解の関係をそもそも検証できるようになる前に必要とされる。どれだけの理解が欠けているかは、何が知られているかによって変わるからである。

3. 何が理解とみなされるのか

理解 (understanding) は理論中立的な成功のラベルではない。現在の哲学文献は、理解の対象、それらの対象に対して理解が主体に求める心理的關係、そして見かけ上の理解を成功した理解にする規範的關係について論争している (Grimm 2021)。これら三つの次元を明示してはじめて、システムが何かを理解しているという主張や、理解をもたらすという目標を、評価したり、それを目指して構築したりできるほど明確に述べることができる。

対象は、命題、説明、現象、主題領域、依存構造、モデル、実践、あるいは別の視点でありうる。関連する**心理的關係**は、信念、受容、表象、整合的な組織化、推論能力、操作、実践的技能、あるいは視点取得を含みうる。**規範的關係**は、真理、経験的十全性、理由や証拠による適切な基礎づけ、非偶然性、説明的正しさ、あるいは知識に要求される正当化よりも弱い何かを要求しうる。

さらにひとつの論争が、これらの次元を横断している。理解は知識の一種なのかどうか、という論争である。グリム (Grimm)、カリファ (Khalifa)、グレコ (Greco) は、理解の規範的な特徴を本質的に知識のそれとして扱う。これに対してクヴァンヴィグ (Kvanvig)、プリチャード (Pritchard)、ヒルズ (Hills)、デルセン (Dellsén) は、知識が許容しない種類の運や無効化を理解は許容すると論じる。どのような運が問題になるかは、プリチャードの事例に見ることができる。ある人が自宅の火事の原因を本物の消防士に尋ね、配線の不具合だと正しく教えられたとする。ただし近くには消防士の扮装をした人々がいて、その誰かに尋ねていれば、たやすく誤った答えを得ていた。プリチャードの説では、この人は容易に誤りえたので知識はもたないが、なぜ家が燃えたのかを理解している。配線の不具合がどのように火事を起こしうるかをこの人が把握している限り、周囲の環境にあるだけの運は理解と両立するからである (第7.1節を参照)。これに対して、尋ねた相手そのものが扮装した人で、答えがたまたま正しかっただけなら、知識も理解もない。無効化についてはデルセンが同様に論じる。エンジンが教えられたとおりに実際に動いているなら、教えた人が信頼できないという誤解を招く証拠があっても、エンジンの仕組みについての理解は失われない (Grimm 2021, §2.3; Pritchard 2008, §§9–10)。理解を知識の一種とする第一の見解では、理解に向けられた試みは、何かが成り立つという知識だけでなく、説明や依存関係についての知識に向けられている。したがって、その知識を誰が担い、それが S の認知的状況に入ったかどうかは、第2章が知識について問うたのと同じ問いになる (第2章第6・7節)。これに対して、理解は知識が許容しない運や無効化を許容するとする第二の見解では、知識がなくても理解がありえ、知識があっても理解が欠けうる。そのため、理解が実現したかどうか、誰のうちに実現したか、そして S の認知的状況に入ったかどうかには、それ自体の説が必要になる (第6節)。

この見取り図はすでに、いくつかの短絡を防ぐ。説明の形をした文を産出しても、誰かがそれを理解していることは示されない。首尾よく行為しても、それだけでは理解されている対象は特定されない。偽なるモデルが自動的に役に立たないわけではないが、有用性だけで理解が確立されるわけでもない。同じ人工システムの振る舞いが、ある説のもとでは強い証拠とみなされ、別の説のもとでは弱い証拠とみなされうる。本書は以下で、それぞれの説がどの次元を強調するかによって説を系統に分ける。ただし、いくつかの説は複数の次元を組み合わせているため、系統は互いに排他的ではない。

数学の実践は、簡潔な例解を与えてくれる。二つの証明は、同じ定理を確立しながら、説明の深さ、露わにする依存関係、伝達する戦略、扱いやすくする新しい問題において異なりうる。第1章第4節で紹介したラヴ (Rav) は、証明が概念のあいだにつける道筋は確信させるだけでなく、数学的な観念を照らし出す相互のつながりを打ち立てると考え、サーストン (Thurston) は概念と思考法の成長を強調している (Rav 1989, 67, 73 n. 30; Thurston 1994)。定理が真であると知っていることを固定したまま、候補となる説明や理解という達成が変化しうる。たとえば、1991年にエルゴード理論の方法で初めて証明された密度ヘイルズ–ジュエット定理 (density Hales–Jewett theorem) に新しい証明を与えた、共同で書かれた論文は、そうする理由を二つ挙げた。その初等的な論証は、以前の証明が原理的にさえ与えられなかった明示的な上界を与えること、そしてそれが、著者たちがおそらく最も簡単な既知の証明と呼んだ、セメレディの定理 (Szemerédi's theorem) の証明を与えることである。著者たちはまた、自分たちの論証がなぜ最初の場合を超えてこれほど

容易に一般化したのかを完全には理解していないと書いた (Polymath 2012, 3–4)。定理を固定したまま、新しい証明が付け加えたものには、以前には知られていなかった命題、すなわちその上界と、セメレディの定理へのより簡単な道筋が含まれていた。著者たちが欠いていると述べたのは、論証がなぜこれほど容易に一般化したのかについての十分な理解である。ラヴの説では、新しい証明そのものが、それが付け加えた命題以上のものをもたらす。証明が概念のあいだにつける道筋は、定理を確信させるだけでなく、数学的な観念を照らし出す相互のつながりを打ち立てるからである (Rav 1989, 73 n. 30)。証明どうしの対比も、この共同で書かれた論文も、理解のどの理論が正しいかを決定するものではない。どちらも、その達成を、定理が真であるという知識とは別に評価する必要があることを示している。その達成を、説明や依存関係についてのさらなる知識に数えるか、知識とは別の理解に数えるかは、採る説による。

3.1 説明に基づく説

説明に基づく説によれば、ある現象を理解するには、その現象の適切な説明を把握することが必要である。ストレヴンス (Strevens) は、説明なしに科学的理解はないという強いテーゼを述べている (Strevens 2013)。カリファは、意図的に還元的な説を擁護する。何かがなぜ成り立つのかの理解は、理想的な場合にはその説明についての科学的知識であり、それに及ばない理解はそうした知識に近づいたものである (Khalifa 2022, 1–5; Grimm 2021, §4.1 を参照)。デ・レフトの文脈的理論が扱うのは、科学的な説明から得られる、現象がなぜ生じるのかの理解である。その基準によれば、ある現象がこの意味で科学的に理解されるのは、それを用いる科学者にとって理解可能な理論に基づき、経験的十全性と内的整合性という価値にかなう、その現象の説明があるとき、かつそのときに限る。この理論が文脈的と呼ばれるのは、デ・レフトの説では、理解可能性 (intelligibility) が理論に内在する性質ではなく、理論の使用を容易にする諸性質に科学者が認める価値であり、そのため理論が理解可能かどうかは、科学者の技能や背景知識といった文脈的な要因に依存するからである (Boge and de Regt 2025, 417. de Regt 2017, 40 に従う)。デ・レフトが提案する理解可能性のひとつの十分な基準は、それらの科学者が正確な計算を行わずに理論の定性的に特徴的な帰結を認識できるかどうかである (de Regt 2019, 336 and n. 11)。したがってこの説は、能力に基づく説でもある。これに対してストレヴンスは、説明を構成する能力を、なぜそうなのかを理解していること (understanding why) の条件ではなく、その強い証拠として扱う (Strevens 2013)。

これらの説は、説明されていない不透明な予測についての判定を比較的明確にする。そうした予測は予測的知識を支えるかもしれないが、それによって現象の科学的理解を産出するわけではない。これらの説のもとでは、説明の形をした出力もまた不十分である。これらの説が求めるのは、適切に正しく、支持され、その説が要求する関係を実現する形式で利用可能な説明である。

説明に基づく説のあいだでも、どの関係が説明を理解の源にするのかについては見解が分かれる。デ・レフトとデニス・ディークス (Dennis Dieks) は、説明的理解の主要な構想のひとつとして統一説を検討している。統一説によれば、科学は、多くの現象が少数の型や前提から導けることを示すことによって理解を増す (de Regt and Dieks 2005, §3.2)。シュルツ (Schurz) とランバート (Lambert) は、この考えを明示的に相対的な形で述べる。ハイスベルス (Gijssbers) によるまとめでは、二人の説で理解は、個人であれ共同体であれ、認識の主体がもつ認知的のコーパスに相対的である。「なぜ P か」という問いへの答えがコーパスに P の理解をもたらしうるのは、P がコーパスにどうはまるかを示すとき、すなわち P をコーパスの他の要素と結びつけ、しかも答えによって変わったコーパスが以前より統一されているときに限られる (Gijssbers 2011, 17)。ハイスベルスは、統一の理論としてのシュルツとランバートの説には異論がないとしながら、説明の理論としては、統一は説明の十分条件でも必要条件でもないと論じる。たとえばあるクラブの会員は全員はげているという偶然的な一般化も、事例が十分に多ければ統一を増し、それらの事例を説明することになってしまう。他方で、正当な説明が、説明されるものに劣らず説明を要する前提に立つことがあり、その前提がコーパスの統一を減らすことさえある (Gijssbers 2007, 22–30)。デ・レフトとディークスは、科学者の理解可能性の基準が歴史的に変わってきたことをもとに、統一は理解を得るための有効な道具ではあるが複数の道具のひとつであって、その普遍的な基準ではないと結論する (de Regt and Dieks 2005, §3.2)。この相対的な形を人工システムによる研究に当てはめると、そうした研究が鋭く提起する二つの問いが明示される。ひとつは、ある答えが理解をもたらすというとき、それは誰のコーパスに相対的なことなのか、人工システムのか、研究チームのか、科学共同体のか、という問いである。もうひとつは、ある人工システムがシュルツとランバートの意味でのコーパス、すなわち、信じている文の集まりと、演繹・帰納・確率的な推論を含む論証のうち習得しているものの集まりを、そもそも持つのかどうか、という問いである (Gijssbers 2011, 17)。コーパスによって定式化された説は、この問いに答えていない。誰が理解するのかについては、第6節で立ち戻る。

3.2 対象的理解説と依存関係説

すべての理論が説明の保有を必要条件とするわけではない。デルセンの依存関係モデリング説は、おおまかに言えば、主体がある現象を理解するのは、依存関係のネットワークにおけるその現象の位置についての十分に正確かつ包括的なモデルを把握することによってである、と述べる (Dellsén 2020)。ここで正確さとは、モデルがそれらの関係を誤って表していないかにかかわり、包括性とは、現象が何に依存しないかも含めて、それらの関係を表しているかにかかわる。この二つの基準には第7.3節で立ち戻る。これは、ひとつの対象についての説明を知ることを超えて届くように設計された、対象的理解、つまり現象や主題領域の全体の理解の説である。説明に基づく説も対象的理解を扱う。争点は、対象的理解が説明を超えうるかかどうかである。デルセンはまた、理解は、知識が一般にそうするような仕方では正当化や信念を要求する必要がない、ただし近似的真理の把握はなお要求する、というより急進的な主張も擁護している (Dellsén 2017)。それ以前に、ストレヴンスとデルセンが論じる例において、リプトン (Lipton) は、なぜそうなのかの理解でさえ説明なしに得られうると論じた。その結論を退けるストレヴンスの紹介によれば、リプトンの説では、理解は説明的でない事実を把握することによっても得られる。たとえば太陽系儀 (orrery) のような物理的模型を扱ったあとでは、惑星がなぜときに逆行するように見えるのかを理解していながら、何を学んだのかを言葉にできないことがある。リプトンの説では、明示できないものは説明ではありえないので、このとき理解は説明を経由していない。また、物体は質量によらず同じ加速度で落ちなければならないことを背理法によって示すガリレオの思考実験は、リプトンの説では、その科学的な説明を把握させないまま、なぜそうなのかを理解させる (Strevens 2013)。説明に基づく説の論者は、そうした事例はおそらく暗黙のうちに説明の一部を把握することを含むと応じ、デルセンはそのいくつかを対象的理解としてのみ認める (Strevens 2013; Dellsén 2020)。第3.1節で見たデ・レフトの基準は説明が生む理解にかかわるが、デ・レフト自身は、説明なしになぜそうなのかの理解が得られるかをめぐる論争に触れたうえでその問いを脇に置き、なぜそうなのかの理

解は説明に還元できない、それ自体としての認知的な達成だとする (de Regt 2020, 924)。ハイスベルスは、第3.1節で見たように統一を説明の理論としては批判したが、うまくいった分類はそれ自体として一種の理解をもたらすとも論じている。ハイスベルスの説では、信頼できる帰納を支える解剖学的特徴を選び出し、説明を目指さず分類だけを行う生物学も、新たに見つかった種の性質を予測させ、類似の関係によって生物どうしがどう結びついているかを見せる。それでいて何も説明しない。ハイスベルスはこの理解を、説明が決定の関係によって与える理解から区別し、その達成を帰納そのものではなく、帰納を支える特徴を苦勞して選り出すことに見る (Gijssbers 2011, 178–181)。この見方を採るなら、第4節の強化した予測の事例は、さらに一つの問いにかかってくる。妥当性確認済みの予測器の分類がどの特徴に支えられているかが、把握できる形で利用可能かどうか、という問いである。

こうした見解のもとでは、慣習的な「なぜ」の説明が欠けていることは事例を決着させない。機械が生成した表象も、それが関連する依存関係についての正確で十分に包括的なモデルを把握できるように利用可能にするのであれば、理解を支える。同じ理由で、正しい点予測を大規模に集めても、それが諸現象が互いにどのように依存しているかについての統合された表象を何も提供しないのであれば、対象的理解になるとは限らない。グリムが紹介するグレコの区別は、こうした表象の役割をより正確にする。理解の対象とは理解されるものであり、理解の乗り物 (vehicle) とは、それを通じて対象を理解する表象である。たとえば街のよい地図は、ふつう、通りや目印の配置を把握するための乗り物である。表象そのものが理解の対象になることもある (Grimm 2021, §2.1)。ある主題についての見解の体系も、そのようにして対象になりうる。エルギン (Elgin) の例では、ある人が占星術に通じていることはありうるが、占星術は宇宙についての理解をもたらさない (Elgin 2007, 4–5)。この区別を機械が生成したモデルに移すなら、そうしたモデルはどちらの関係にも立ちうる。乗り物として使われれば、それが表す依存関係の理解を支える。対象として捉えられれば、モデルそのものは理解されても、その理解が現象にまで及ぶとは限らない。したがって、研究者が人工システムの表象を理解しているという証拠は、それだけでは、その表象を通じて現象を理解しているという証拠にはならない。

3.3 能力説と理解可能性説

能力に基づく説は、理解が要求する把握を、主体が何をなしうるかに位置づける。これらの説は、上で見た区別を横断する。ヒルズは、 q が p の理由であるとき、なぜ p なのかの理解を支える能力を、認知的制御として特徴づける (Grimm 2021, §2.2)。エマ・ゴードン (Emma Gordon) が『インターネット哲学百科事典』で紹介するところでは、それらは、 q が p の理由だとただ信じるのではなく、 q を p の理由として扱えるようにする能力である。すなわち、他者が与えた p の説明についていく能力、 p を自分の言葉で説明する能力、 q という情報から p という結論を引き出す能力、 p が与えられたときに正しい説明 q を与える能力、そして後の二つの能力を関連して類似した事例についてももつことである (Gordon n.d., §3c)。ヒルズの説は能力に基づくが、なお説明を要求する。これに対してウィルケンフェルド (Wilkenfeld) の表象の操作可能性説は、対象の理解にかかわり、説明を要求しない。主体がある対象を理解するのは、その対象についての有効な推論や操作を支えるように修正しうる、その対象の表象をもっているときである (Wilkenfeld 2013, 1, 8–9)。上で述べたデ・レフトの理解可能性の条件もまた、能力に基づく (概観として Grimm 2021, §2.2 を参照)。こうした説は、主体が何をなしうるかのテストを示唆する。すなわち、帰結を引き出すこと、「もし〜なら」という問いに答えること、関連して類似した事例を説明すること、あるいは新しい推論を支えるように表象を修正することである。これらの考えは、アクセス不可能な把握の感覚に訴えるのではなく観察可能なテストを示唆するので、人工システムを評価するうえで魅力的である。

それらはまた、構成概念妥当性の問題をもち込む。説そのものは、近道を理解に数えるわけではない。たとえばウィルケンフェルドの説は、運のよい遂行や、暗記したアルゴリズムの単なる適用を、理解から除外している (Wilkenfeld 2013, 8–9)。困難は、ベンチマークのほうが、そうした近道と、説が要求する能力とを見分けられないことがある点にある。システムは、暗記、検索、汚染、隠されたツール、あるいは第二の不透明な手続きによってベンチマークに答えるかもしれない。選ばれた反事実に対して正しくふるまうことは、有用な能力を示しはしても、システムが主題領域について統合された理解を保有していることを示さないかもしれない。したがってテストでの遂行は、特定の説が理解を構成する、あるいは理解を示すときみなす能力についての証拠である。それらの能力を近道から見分けるためには、遂行は敵対的に選ばれた変形の全体にわたって集められる必要がある。そうしたテストの集まりは、規定によって理解を定義する理論中立的なチェックリストではない。

3.4 事実性、理想化、認知的に有用な歪曲

p であるという知識は、前章が分析の中心に置いた厳格な命題的用法においては、事実に (factive) である。理解が同じ仕方で事実にあるかどうかは論争されている。グリムとデ・レフトがまとめるところでは、エルギンは、理解は事実にではないと論じる。真ではないと知られているモデル、理想化、思考実験は、適切な虚偽 (felicitous falsehoods) でありうる。それらは、標的の特徴を例示する、つまりその特徴を際立たせ、ほかの特徴を目立たなくすることによって理解を進める。ただしエルギンは、理解はなおなんらかの仕方で事実に応答すると認める (Grimm 2021, §4.2; de Regt 2018)。ポトチュニクは、広範な理想化は科学が真理を目指すという考えに疑問を投げかけると示唆し、科学の認知的な目的は理解であって、それは科学の複数の目的のひとつだとする (Potochnik 2015, 1, 4, 16)。デ・レフトも同様に、理解を得るにはしばしば表象の正確さを手放す必要があり、理想化をはじめとする不正確な表象は、科学的理解を高めるためによく用いられる道具だと論じている (de Regt 2020, 932)。事実性を退けることは、あらゆる正しきの条件を退けることを意味しない。そして正しきの条件を保つ説のあいだでも、その条件が何に及ぶかは異なる。ポトチュニクの説では、理解はなお世界に応答する。表象のなかのそれぞれの指定が認知的に受容可能であること、すなわち真理からの隔たりが、表象のなかでのその役割と研究の目的とに照らして無視できるものであること、そして把握される因果的パターンがその現象に実際に体现されていることが求められる (Potochnik 2020, 6–8)。ストレヴンスは、理想化を含む説明を、その説明的内容、おおよそ何が違いを生み何が違いを生まないかについてのその主張が真であるとき、正しいとみなす (Strevens 2013)。これに対してポトチュニクは、研究の焦点でない要因であれば、大きな違いを生む要因であっても理想化によって置き換えてよいとする (Potochnik 2015, 10–12)。

関連する論点は、モデルの評価が何を確立しうるかに関わる。オレスケス (Oreskes)、シュレーダー＝フレシェット (Shrader-Frechette)、ベリッツ (Belitz) は、特に開放的な自然システムの数値モデルについて、検証や妥当性確認は真理の決定的な論証にはなりえないと論じている。経験的な一致は部分的な確証を与えるにすぎず、異なるモデルが利用可能な観測と両立したままでありうる (Oreskes, Shrader-Frechette, and Belitz 1994)。パーカー

(Parker) の目的適合性 (adequacy-for-purpose) 説は、異なる評価の標的を提示する。モデルは、その全体的な表象の正確さによってだけでなく、対象、使用者、方法論、状況、目的に相対的に評価される (Parker 2020, 458, 464–465)。これらの立場はひとつの合意された理論ではなく、第一の立場はあらゆる形式的モデルや科学的モデルについての主張でもない。両者はあわせて、**モデルは妥当性確認された** (the model was validated) ということが、それだけでは、そのモデルが予測、説明、介入、教育、あるいは他の用途に適合しているかどうかを語らない理由を示している。

この論争は、知識と理解の成立条件を区別するうえで重要である。文字どおりには偽であるモデルは、それ自体としては、偽なる命題についての知識を提供できない。それでもそのモデルは、領域を再編成し、依存関係を明らかにし、能力を支え、あるいは理解を増大させうる。その寄与の一部は、モデルが何を表象し何を可能にするかについての真なる命題を通じて再構成されうる。ケルプ (Kelp) の知識にもとづく理解の説は、理想化されたモデルについてそうした再構成を与える。理想気体の法則を含む理論が何を述べているか、現実の気体がさまざまな点でさまざまな程度に理想気体に近似していること、その理想化の限界がどこにあるかは、知ることができる。ケルプの説では、この知識は気体の物理についての理解の一部をなす (Kelp 2017, 18)。そうした再構成が達成を汲み尽くすかどうかは争われている。ポトチュニクも真理に居場所を与える。その説では、理想化された表象は、それを体現する現象については偽であっても、それが描く因果的パターンについては真でありうるからである。しかしポトチュニクは、理解とはそうしたパターンを把握することであり、それは信念を要求せず、したがって知識とは別のものだとする (Potochnik 2020, 6–7, 14–16; Potochnik 2015, 7–8)。したがって、そうした寄与を理解に数える説のもとでは、理解は、単に命題的知識のより冗長な形式ではない。

人工システムによる研究にとつての直接的な合意は、有用性が真理に取って代わるということではない。上の説では、理想化されたモデルが理解に寄与するのは、それが何をどの目的のために理想化しているかに相対的にだからである。そのため、モデルを扱う人工システムは、各表象が何を理想化しているのか、どの対象と目的に関わるのか、どの推論を支えるのか、そして対抗するモデルが証拠とどこで両立したままであるのかを、保持しておく必要があるかもしれない。それは、その結果を評価し用いる者のためでもあり、システム自身に理解を帰属させる場合のためでもある。ワイスバーグ (Weisberg) のいうモデルの解釈 (construal) は、これらの項目に構造を与える。解釈は、モデルの部分を対象の部分に割り当てて、どの部分を無視するかを定め、意図された射程を定め、二種類の忠実性の基準を設ける。動的な忠実性の基準は、モデルの出力が現象にどれだけ近くなければならないかを定め、表象上の忠実性の基準は、モデルが正しい理由で正しい予測をしているかを問う (Weisberg 2007b, 14–17)。ワイスバーグの説では、解釈はモデルの使い手の意図に依存する。そのため、人工のモデル作成者について、誰の意図が解釈を定めるのかは、さらに別の問いになる (第8章第2.2節)。

こうした項目を保持する必要は、人工システムに特有のものではない。たとえばクレランド (Cleland) は、生命が原始のスープで始まったという仮説の検査として提示されたミラー＝ユリーの実験を、生命の基本的な構成要素であるアミノ酸が特定の気体の混合物のなかでの放電によって作られるかどうかの検査として読み、実験室で実際に検査された仮説と生命の起源についての仮説との論理的な関係は、初期の地球の条件から生化学的な可能性に至る多数のきわめて思弁的な想定を経由している、とする (Cleland 2002, 484)。同じように、より詳細なモデルや予測的により正確なモデルが、あらゆる説明目的や介入目的にとってより良いとは限らないし、有用な理想化も、うまく機能するというだけの理由で文字どおり真なる記述になるわけではない。エルギンはたとえば、いくらでも精密にできるビリアル方程式があっても、理想気体の法則を使い続けることは正当化されると論じる。モデルに求められるのは現象を複製する表象ではなく、現象の理解をもたらす表象であり、同じ標的の異なるモデルは異なる特徴を例示するからである (Elgin 2022, 18–20)。

ワイスバーグによる理想化の分類は、モデルを作るのが機械になったときに何が変わるのかにかかわる。この分類は理想化の種類を、それを正当化する理由によって区別するからである。ガリレオの理想化は、理論を計算可能にするために歪曲を持ち込む。その正当化は実用的で、扱いやすさに結びついているので、ワイスバーグの説では、計算能力と数学的技法が進歩すれば、理想化した者は理想化を解いていくはずである。最小限の理想化は、現象に違いを生む因果的要因だけを残す。その正当化は説明にあり、ワイスバーグによれば、科学が進歩しても弱まると期待すべきではない。多重モデル理想化は、互いに関連しながら両立しない複数のモデルを、単一の最善のモデルを期待せずに作る。一般性、正確さ、精度、単純さといった望ましがトレードオフの関係にあるところでは、その正当化は恒久的だとワイスバーグは考える。そうしたトレードオフは論理と数学の帰結だからである (Weisberg 2007a, §I)。マッシュューソン (Matthewson) とワイスバーグは、これらのトレードオフの一部は論理と表象だけから生じ、科学が進歩しても消えないと論じる (Matthewson and Weisberg 2009)。またワイスバーグは、三つの種類のいずれも科学についての實在論と両立するとみなしており、その分類は、事実性を否定するエルギンの側に立つものではない (Weisberg 2007a, §IV)。この分類を、人間の対応者よりはるかに大きな計算能力をもつシステムに移すと、そのシステムに理想化を解く理由を与えるのは、扱いやすさによって正当化された理想化だけだということになる。ワイスバーグが引くレヴィンズ (Levins) は、モデルが複数になる理由の一部を、少数の変数しか同時に扱えない心に求めた。これは人間の認知についての前提である。これに対して、論理と表象だけから生じるトレードオフはその前提に依存しない。そのため、一つの対象について複数のモデルを保持する理由のうち、そうしたトレードオフに基づくものは、モデルを作る者が替わっても残る。したがって、設計がこれらの立場から一つに定まるわけではないが、研究システムは、あらゆる探究を、ますます文字どおりになっていく単一の世界モデルへと押し込めるのではなく、目的相対的な表象の族を保持することもできるだろう。

これらの系統の説は、異なる可能な標的と異なる証拠の負担を露わにする。それぞれが、人工研究者が何を表象し、何をなし、何を伝達し、あるいは他の主体に何を利用可能にしなければならないかを定めるのである。

4. 説明的理解なき知識

第3節は、何が理解に数えられるかを説ごとに並べた。本節は、それらの説を強化した予測の事例に当てはめ、そこで知識と説明的理解が切り離されるか、それがどの説を選ぶかにかかわらず成り立つか、そして切り離されるとき、試みの目標と進歩について何が帰結するかを問う。

強化した予測の事例に戻ろう。新しい規則性は真である。主張は明示的な領域に限定されている。独立した測定経路と独立したグループがそれを支持している。予測器の誤差特性は知られており、検査間の依存関係は記録されており、無効化要因があれば共同体はその依拠を修正するであろう。共同体のどの人間の成員も、対象となる規則性の十全な説明を、暗黙のうちにさえ把握しておらず、与えることもできない。システム自身がそれを理解しているかどうかは、ここでは定めない (第6節)。

第3節で検討した理解の説は、前章の議論と合わせて、条件つきではあるが実質的な次の一点を支える。

ある規則性が成り立つという知識と、それがなぜ成り立つのかについての説明的理解とは、概念的に分離可能である。知識についてのいくつかの重要な説のもとでは、強化した予測の事例において前者は実現しうる。そして保有についての適切な説のもとでは、共同体の認識的状況に入りうる。その一方で、なぜそうなのかの理解に説明を要求する説のもとでは、後者は共同体のどの人間の成員のうちにも実現していない。この特定の結果が知識であるかどうか、そしてそれが共同体の認識的状況に入ったかどうかは、前章で比べた条件のうちどれを採るかに相対的なままである。

あることが成り立つという知識と、なぜそうなのかの理解との区別だけでは、想定した事例で後者が欠けているかどうかは決まらない。説明に基づく説、なぜそうなのかの理解についてのヒルズのような能力説、そしてなぜそうなのかの理解を説明についての知識として扱う還元的な説は、いずれも、ある規則性が成り立つことを知っていることと、なぜそれが成り立つのかを理解していることを区別する。説明なしの対象的理解を認めるデルセンも、なぜそうなのかの理解は説明から切り離されえないとする (Strevens 2013; Dellsén 2020; Grimm 2021, §4.1)。これらの説は、説明が利用可能でないときにも現象についての何らかの理解が残るかどうかで分かれる。デルセンのような依存関係説と、ウィルケンフェルドの表象の操作可能性説 (第3.2節、第3.3節) は、説明なしの対象的理解を認め、リプトンは、説明なしにもなぜそうなのかの理解がありうると論じた。そのためリプトンの説では、人間の成員の誰も説明を把握しておらず、与えることもできないという規定だけでは、規則性がなぜ成り立つのかを誰も理解していないことまでは決まらない。

この分離は、自動化された科学をめぐる議論でも擁護され、またその帰結に抵抗されてきた。ハンフリーズ (Humphreys) は、人間を科学の中心から遠ざけることで世界についての理解が失われかねないという懸念を検討し、知識を拡張して捉えるなら、「人間の現在の理解を超えた」科学的知識を認めるべきだとする。その根拠として挙げるのは、アスピリン、コレラ、多くの種類の癌のように、かわる仕組みが理解されないまま効果的な予測と操作がなされた事例や、測定を満足に説明できないまま数十年にわたって予測に成功してきた量子力学である。ただしハンフリーズは、人間が科学を理解していなければ誤りが気づかれなくなるという別の懸念を、科学の目的についての二つの見方、すなわち科学は主に世界についての知識と理解をめざすという見方と、主に予測と操作をめざすという見方の、どちらのもとも保持する。自動化された方法は、その安定性の条件が満たされないときに突然失敗しうるからである (Humphreys 2020, 14–15)。同じ論集でブーン (Boon) は、科学的知識をその使用に結びつける。ブーンの説では、不透明で機械のなかに近づけないまま収まっているために、人間が取り出して自ら選んだ認識的課題に用いることのできないデータモデルは、従来の科学的実践が生み出すような科学的知識ではない。そうした知識は、公共的で移転可能な認識的道具として表された科学者の構想からなるからである。ただしブーンは、知識を人間にとって扱えるものにする実用的な基準を機械学習が原理的に満たしうるかどうかは、答えずにおく (Boon 2020, 14–15, 16–17 n. 11)。ブーンの基準は、上に掲げた主張を否定しない。強化した予測の事例では、妥当性が確かめられた規則性は共同体によって述べられ、用いられており、その基準が科学的知識の地位を否定するのは、予測器のなかの不透明なモデルのほうだからである。ブーンはこの基準を人間の実践について述べている。ブーン自身は論じていないが、人工社会である *S* に当てはめれば、この基準は、あるモデルがその社会の成員にとって認識的道具として役立ちうるかどうかを問うことになる。

バイスパルトの深層ニューラルネットワークについての議論 (第2節) は、この分離を一種の予測器に当てはめるものであり、あらゆる予測が理解を欠くことを示すものではない。バイスパルトの説では、科学者は、ネットワークの実績によって、それが入力を入力へ正しく写すことと受け入れることを正当化されうる一方で、それが予測することについての説明的理解を欠いている (Beisbart 2025, 14, 18)。この議論は、上の諸説が分かれる問いにも関わる。たとえばネットワークを何度も走らせることで、入力の変化に応じて出力がどう変わるかを予測できるようになった科学者は、理解に必要な推論の一部を行うことができる。バイスパルトは、これを予測される現象についてのある程度の対象的理解に数えるが、説明的理解には数えない。というのも、その推論は、科学者のもつ説明に基づいていないからである (16–17)。また、科学者がネットワーク自体を走らせてそれらの推論を行うことを認める可能性も検討し、これを退ける。バイスパルトの論じるところでは、その場合に厳密な意味で理解をもつのは人間とコンピュータからなる結合したシステムであり、それは、理解が問われている、補助を受けない人間とは別の主体だからである (17–18。誰が理解するのかは第6節で扱う)。ただし、この議論は、ネットワークが説明的理解に寄与する道筋を閉ざしてはいない。ネットワークが説明に関わる特徴に反応していることを、たとえばどの変数が出力に影響するかを検出する説明可能なAIの手法によって科学者が知っており、さらにネットワークが含む説明を正当に受け入れ、それを用いて推論できるなら、科学者は予測されたことがなぜ成り立つのかを理解しうる (15, 18)。また、単純な設定の外では、ネットワークを多様な入力で繰り返し走らせることで、部分的な原因のような、説明に関わる要因を同定できる。バイスパルトはこれを、ネットワークが理解を高める一つの意味に数える。しかし、それには説明についての事前の知識が必要であり、得られるものは完全な説明には遠く及ばない (21–23)。さらにバイスパルトは、きわめて広い範囲の事例での成功が、ネットワークが法則や原因を捉えていると考える理由を与えうる可能性を残しつつも、適用範囲が通常は狭い現在のネットワークについては、そうした理由はないとする (15–16)。そのうえで、ネットワークをどこまで理解の提供へ近づけられるかは、人間がそれをどう用いるかに依存すると結論する (25)。これらの道筋に照らすと、強化した予測の事例は、ネットワークが何に反応しているかを知るという道筋を、規定によって除いている。というのも、この事例では、どの人間もその規則性を説明できず、システムはその規則性の原因となっている依存関係についての使用可能な表象を何も提供しないからである。これに対して、ネットワークを繰り返し走らせるという道筋は、説明には至らないが説明に関わる要因を得る源として残されている。

チリムータ (Chirumuuta) は、神経科学の一分野について、理解の不足が方法の改良を超えて残りうると論じる。運動皮質の活動の解読器と、視覚ニューロンのモデルという二つの事例で、理解可能だが単純な実験室条件の外ではうまく予測できなかった、単純でおおむね線形のモデルから、はるかによく予測するものの、それを作った者がモデルの応答を定性的に予期することを許さない人工ニューラルネットワークへの移行をたどる。以前のモデルにあつてネットワークには欠けている理解可能性の源を四つ挙げる。回路図として視覚化できること、神経符号についての明示された理論的仮定、線形性、そして機能的透明性、すなわち訓練されたモデルが計算する関数を明示的に知りうることである。ネットワークの正確さは複雑な非線形の関数を近似することから来るので、チリムータの論じるところでは、理想的なまでに逆解析され、入出力の関数が書き下されたネットワークでさえ、手作業で組まれた先行のモデルよりも理解しにくいままであり、予測と理解のトレードオフは残る。さらに、対象が複雑な非線形の依存関係を含み、統制された条件の外での予測が求められるところでは、このトレードオフがこれらの事例を超えて現れると予想する。この説では、そうしたネットワークも何種類かの説明は与えるが、「何が予期されるかは教えても、なぜかは教えない」 (Chirumuuta 2021, 15–23)。バイスパルトが、ネットワークが理解にどれだけ寄与するかを人間

による用い方に依存させたのに対して、チリムータの議論は、この種の対象については、そうした方法がどこまで進んでも、ネットワークはより単純な競合モデルより理解しにくいままであると考える理由を与える。その限界は、ネットワークを正確にしている複雑さそのものから来るからである。この論文は理解可能性を「モデルの、人間に相対的な長所」と呼んでいる（26）ので、残ると予想されるトレードオフは、予測と、人間にとっての理解しやすさとのあいだのものと読める。チリムータは人工の理解者を論じていないが、それに当てはめれば、ネットワークを正確にしている複雑さは残る。しかし、それがその理解者の把握を限るかどうかは、何がその理解者にとって理解可能かによる。これは、第6節で取り上げる、誰が理解するのかという問いである。

デ・レフトの説は、次に見るように、強化した予測の事例となじみにくい。デ・レフトの見方では、理論からの予測の成功は、その理論がそれを用いる者にとって理解可能であることを要する（第2節）。予測器がこの意味での理論であったなら、その理論は使い手にとって理解可能でなければならないので、この事例はまったく理解を欠くものではありえない。ただし、規則性がなぜ成り立つかの理解には、この説ではなお、その理論に基づく説明が要る。ボーゲとデ・レフトが記述する種類の、よく予測しながらほとんど理解をもたらさないかもしれないデータ駆動のニューラルネットワークについては、この分離は述べたとおりに当てはまる。

強化した予測の事例において知識が実現しようという主張は、正確さを知識として扱うことに依存していない。この事例は、関連する説が要求するさらなる証拠、信頼性、基礎づけ、運を排除する条件、あるいは社会的条件を規定している。内在主義が自動的にこれを退けるわけではない。共同体は、規則性がなぜ成り立つのかについての説明を保有することなしに、その規則性が成り立つと信じるためのアクセス可能な理由をもちうる。信頼性主義は、適切に確立された真理促進的な経路を通じてこの事例を認めうる。有能さに基づく説や様相説は、成功が能力の発現であるかどうか、そしてあまりにも容易に失敗しえたかどうかを問うであろう。社会的な説は、誰がその態度を保有しているのか、そして評価、受け止め、訂正がどのように組織されているのかを問うであろう。

この分離はまた、二種類以上の理解を許容する。研究者は、なぜ証拠がその予測器への依拠を根拠づけるのかを理解しつつ、なぜ対象となる規則性が成立するのかを理解していないことがありうる。彼らは、予測器の内部計算を理解することなしに、その作動限界を理解しているかもしれない。彼らは、いくつかの説が部分的理解と呼ぶような、薄い依存関係の表象を獲得するかもしれない。したがって **理解なき知識**（knowledge without understanding）という言い方は、欠けている理解の対象と種類を名指してはじめて中身をもつ。ひとつの達成のなかにも、複数の種類の理解がありえ、そのあるものは得られ、あるものは欠けていることがある。たとえば、リキッド・テンソル実験（Liquid Tensor Experiment、第1章第3節）の中心的な定理が形式的に検証されたのち、ペーター・ショルツェ（Peter Scholze）は、形式化によって何が実際にその証明を機能させているのかが分かったと書いた。ショルツェはまた、証明を検証するときには通常、論証に小さな変形を加えてそれが壊れるかどうかを試すのに対して、自分はその証明の経路がうまくいくという安心を得たものの、その周りの地形についてはなお何の感覚ももっていないと書いた（Scholze 2021）。この例では、その経路がなぜうまくいくのかの理解は得られたが、経路を少し変えたらどうなるのかについての感覚は得られなかった。

本章の主要な事例である強化した予測の事例は、ある規則性が成り立つという知識を、それがなぜ成り立つのかについての説明的理解なしにもつことに関わる。この分離は人工の予測器に特有のものではない。たとえば、モリス・タネンバウム（Morris Tanenbaum）は、2004年のインタビューでベル研究所でのシリコン・トランジスタの初期の仕事を振り返り、加熱して急冷すると伝導型が変わってしまう結晶から、焼きなましの技法を使ってトランジスタを得なければならなかったこと、そしてその変化を引き起こす不純物が酸素であることを同僚が示したのはその後だったことを語っている。結晶を育てるときに気づいていた関連する規則性、すなわち融液をかき混ぜないとこの効果がずっと弱くなることについては、「なぜなのかは分からなかった」と述べている（Tanenbaum 2004, 42–43）。

知識が説明的理解なしに得られうるということは、理解を知識に従属させるものではない。理解が目標であるところでは、規則性が成り立つという知識を得ても、その目標には達しないからである。進歩の説は、これと関連する対比を、進歩とは知識の蓄積であるとするパードの認識説（第2章第8.3節）に対して立てる。デルセンのノエティック（noetic）説は、科学的進歩は蓄積された知識としてよりも増大した理解としてよりよく理解されると論じており、その論拠は両者が乖離する事例にある。一方で、アインシュタインが1905年にブラウン運動を説明したことは、デルセンの説では進歩に数えられる。ところが同じ説によれば、当時のアインシュタインは、分子運動論も説明されるその現象も、知っていると言えるほどの正当化をもっておらず、知識は加わっていなかった。他方で、ランダムな実験結果、コウノトリの数と出生数のあいだの相関のような見かけだけの相関、どんな説明も予測も支えないほど貧弱な観察記録についての知識は、この説では、何の進歩ももたらさずに蓄積されうる（Dellsén 2016, §§3–4; Dellsén 2021）。ポトチュニクも同様に、科学的目的の多元性を強調しつつ、理解を科学の認識的な目的として動機づけている（Potochnik 2015）。デ・レフトは、理解を予測的成功へと還元することなく、成功した科学の中心に据えている（de Regt 2020, 932）。

デ・レフトとディークスがこの見方をとる理由は、強化した予測の事例に直接に当てはまる。その理由は、成功した科学理論を、お告げがつねに正しいと分かる仮想的な神託と対比することにある。そうした神託は経験的十全性を確保するが、その完璧な予測がどのようなもたらされたのかについての理解がないので、大きな科学的成功とはみなされない、と二人は論じる（de Regt and Dieks 2005, §2）。同様に二人は、支配方程式を数値的に解いて正しい予報を出すことだけに従事する気象学者は天気を理解していないことになる（de Regt and Dieks 2005, §4.1）。デ・レフトの基準のもとでは、理解可能な理論に基づく説明を何も与えない妥当性確認済みの予測器は、それが知識としてどのような身分をもつにせよ、その規則性を、その基準が特徴づける意味では科学的に理解されないままにしておくことになる。

デルセンの理解は、狭い意味で説明的なものではなく対象的なものである。改めて述べられた説では、理解の増大とは依存関係についてのより正確な、あるいはより包括的なモデルであり、それは説明だけでなく予測と操作をも基礎づける。そして進歩は、社会の関連する成員が自らの理解を増大させることを可能にする、公的に利用可能な情報にある（Dellsén 2021）。この読みでは、強化した予測の事例がノエティックな進歩に数えられるかどうかは、単に説明が利用可能かどうかではなく、妥当性確認された規則性がそのようなモデルを改善するかどうかにかかっている。デルセンの以前の定式は、予測についてより直接的であった。そこでは、ある対象のある側面を多様な状況にわたって正しく予測する仕方を把握することは、説明がなくても部分的な理解に数えられた。たとえば、気圧計のさまざまな読みから天気の変化を予測できる人は、天気を部分的に理解している。ただし、一度の正しい予測や、その手

順を暗記しているだけでは足りない (Dellsén 2016, §2)。この定式では、説明なしの予測の把握はそれ自体で部分的な理解でありうる。こうした把握は、理解可能な理論に基づく説明を求めるデ・レフトの基準からは外れる (de Regt 2019, 334)。ただし第3.2節で見たように、デ・レフトは、説明なしになぜそうなのかの理解が得られるかどうかを脇に置いている。改めて述べられた説では、予測はむしろ、理解がもたらす利点の一つに数えられる (Dellsén 2021, 11256)。第2章で導入した真理近似性の説と問題解決の説は、判定を理解にそれほど直接にはかからせない。

第2節で導入したウッドワードの介入主義的な説は、さらに別の判定を与える。この説では、説明は「もし事態が違っていたら」という問いに答える。すなわち、挙げた要因が変わったなら被説明項がどう違っていったかを、それらの要因への少なくともいくつかの介入のもとで成り立ち続ける一般化を用いて告げる (Ross and Woodward 2023, §3)。そうした問いにまったく答えない候補は、そもそも説明にならない。その閾値を超えると、説明は、そうした問いに答えるのに関わる情報をどれだけ多く与えるかに応じて、深さが異なる (Ross and Woodward 2023, §4)。このように解すると、ウッドワードの説では、強化した予測の事例は、妥当性確認された予測器が介入のもとで何が起こるかについての正しい答えも支えるのでない限り、説明を欠く。そうした答えを支える予測器は、その情報がそれを使う人々にとってアクセス可能である限りで、その程度において説明する。この条件には第6節で立ち返る。

検討した立場は、共同体が規則性を知っていることを否定する必要はない。知識が求める達成であること、あるいはその全部であることや、知識の増加が重要な進歩であることを否定しうるのである。作業的特徴づけは、知識と理解のどちらが求められるかを定めないの、この相違を許すが、解決はしない。理想化、説明、対象的な把握を通じた領域の再編には、なお成功の説明が必要である。また、知識や理解の局所的な増加だけでは、より広いプログラムがどう進歩したかは決まらない。

したがって、強化した予測の事例は異なる判定を支えうる。規則性についての知識と、それがなぜ成り立つかの理解を分ける説では、活動は知識の目標に達しても、理解に向けた探究は未完でありうる。理論が使用者にとって理解可能であることを求めるデ・レフトの要件は、予測の事例を別の仕方で制限する (第2節)。予測器がその意味での理論であるなら、その使用の成功は、デ・レフトの説では、使い手がその理論を理解していることを前提とする。ただし、規則性がなぜ成り立つかを使い手が理解していることまでは前提としない。ポトチュニク、エルギン、デルセンのように理解を科学やその進歩の中心に置く説 (第7.1節) は、作業的特徴づけが確立も排除もしない優先性の主張を加える。特徴づけ自体は、研究の個々の操作にも、研究活動そのものにも、知識であれ理解であれ、何らかの達成をもたらすことを要求しない (第1章第3節)。成功に理解が必要かどうかは、試みが何に向けられているかによる。そして、組織された試みの方向、その最終的な成功、特定の操作の寄与は別である。

研究プログラム全体と個々の操作という規模の区別は重要である。研究プログラムは、その構成要素であるひとつの実験が測定しか産出しない場合でも、理解を目指しうる。分野は、それ単独ではどれも説明的でない結果から組み立てられた説明的総合を通じて進歩しうる。逆に、プロジェクトは、参加者が求めた理解を前進させることなしに、多くの真なる局所的主張を産出しうる。科学の目的や進歩についての理論は、論証なしに、研究と呼ばれる各活動が研究であるための条件へと転換されることはできない。

いくつかの説のもとでは、この分離は逆向きにも成り立つ。デルセンは、理解は、知識が一般にそうするような仕方では正当化や信念を要求する必要がないとする (第3.2節)。エルギンは、適切な虚偽が理解を進めるとし、ポトチュニクは、理解とは因果的パターンを把握することであり、それは信念を要求せず、したがって知識とは別のものだとする (第3.4節)。また、デルセンが挙げる事例では、弦理論を方法論上の留保から信じてはいないが、受容して説明に用いている物理学者が、デルセンの説では、その理論が記述するものを理解している (第6節)。こうした説のもとでは、理解に向けられた試みは、たとえば理想化されたモデルや、信じられてはいないが受容された理論を通じて、新しい知識をもたらすことなしに、*S* にとって新しい理解をもたらしうる。そして、知識や妥当性確認済みの結果だけを数える評価は、その成功をとらえない。これに対してカリファは、理想的な場合のなぜそうなのかの理解を、説明についての科学的知識として分析する。ケルプは、最大の理解を、最大限に包括的で、最大限によくつながった知識として分析する (第3.1節、第7.1節)。ただし、これらの分析だけからは、理解が増すときには知識も加わらなければならないことにはならない。ケルプの説は、このことを一方向について明示している。ケルプのリトマス紙の事例では、ある出来事についての完全な説明の各項目を別々の情報源から知っているが、それらの信念のあいだの基づけの関係が、項目のあいだに成り立つと知っている説明や支持の関係を反映していない行為者は、その出来事をほとんど理解していない。ケルプによれば、さらに知識を加えることがこれを解決するかは明らかでなく、この説では、知識がよくつながるようになるにつれて理解の度合いが上がる (Kelp 2015, 16–18)。したがってこの説では、新しい事実が知られることなしに理解が深まりうる。これを *S* にとって新しい理解に当てはめるには、そうしたつなぎ直しが *S* の認識的状況のなかで起きることが前提になる。ケルプの説はまた、理解が、知識に達しない信念と共存することも認める。理解を担うのは、その周囲にある知識だからである。コマンチ族についての真なる信念がゲティア化されている歴史家というクヴァンヴィグの事例に対して、ケルプの第一の応答は、その信念は知識でありうるというものである。しかしケルプは、自らの説では、探究者の信念の多くがゲティア化されていても理解を帰属させようと付け加える。残る知識、すなわち当該の理論が何を述べているかの知識と、その理論が非常にもつともらしいという知識が、文脈が定める基準を満たすのに足りうるからである (Kelp 2017, 16–17)。なぜそうなのかの理解についても同じことが成り立つ。複数の偽の消防士のなかにいた一人の本物の消防士から、家が焼けた理由を聞いた男というプリチャードの事例では、ケルプの第一の応答は、この男が知識を欠くことを否定するものである。しかしケルプは、その直観を認めるなら、男はなぜそうなのかを知ることなしに理解しているとする説も示す。男は、よく基礎づけられた説明、すなわち最終的には知っている情報源に根ざした説明を与えるのに足るだけのことを知っているからである (Kelp 2014, 11, 13–15)。どちらの場合も、ケルプの説で理解を担うのは、探究者が実際にもっている知識である。

5. 不透明性には所在がある

強化した予測の事例のシステムは、自らがとらえている依存関係について使用可能な表象を何も提供しない。つまり不透明である。そこから二つの反論が思い浮かぶ。その不透明性は、第4節が認めた知識の根拠づけを損なうのではない、という反論と、その不透明性によって、誰もその規則性を理解していないことが決まるのではない、という反論である。一つ目が成り立つなら、理解の欠如だけを取り出すために事例が前提にした知識そのものが失われる。本節は二つとも取り上げる。第5.1～5.3節は、不透明性がどこにあり、根拠づけられた依拠に対して何にかかわるのかを問い、第5.4節は、計算が不

透明なままでも、再構成や別のモデルを通じて理解がもたらされうるかを問う。誰が理解するのかは、不透明性を主体に相対化したうえで、第6節が取り上げる。

ブラックボックス (black box) と呼ぶだけでは、人工システムによる結果を評価するのに必要な区別はまだ引かれていない。不透明性は、何が不透明なのか、不透明な要素がどの役割を果たすのか、誰にとって不透明なのか、そしてそれに依存することの利害がどれほどかに指標づけられてはじめて中身をもつ。コンピュータ・シミュレーションと計算科学について論じるなかで、ハンフリーズは、ウィンズバーグとアルバラードがコンピュータ・シミュレーションについての百科事典の項目で引く定義において、認識的不透明性を主体と時点に相対化している。ある過程がある時点のある認識的主体にとって認識的に不透明であるのは、その主体がその時点で、その過程の認識的に関連する要素のすべてを知っているわけではない場合であり、かつその場合に限る。さらにハンフリーズは、主体の本性からして、その主体がそれらの要素のすべてを知ることが不可能な場合に、その過程はその主体にとって本質的に認識的に不透明であると言う (Winsberg and Alvarado 2026, §2.5)。本質的な不透明性は主体の本性に相対的なので、人間の主体にとって本質的に不透明な過程が、人工の主体にとってもそうであるとは限らない。そのため、結果の評価や使用に人工の主体が加わる場所では、不透明性をどの主体に相対化するかが効いてくる (第5.2節)。第5.3節で取り上げるデュラン (Durán) とフォルマネク (Formanek) は、本質的な不透明性の概念をこの定義から得ている。

ハンフリーズは、シミュレーションの速さと複雑さのために、計算の細部を人間が調べることはできないと論じる (Humphreys 2009a, §3.1)。デュラン (Durán) は、それ以前のハンフリーズの議論を引用している。計算的に還元不可能な過程であっても、人間が各段階を調べられるところまで遅くすれば、段階を見通せるようになるという議論である。その議論では、この領域の不透明性の根は、計算を速く行わなければならないという実際上の必要にある。その制約を人間は回避できないため、計算科学については透明性を要求として手放さなければならないとされる (Durán 2025b, §1)。この議論は、主体が替わったときに変わりうるものと変わらないものとを分ける。ハンフリーズは人工の検査者を論じていないが、必要な速さで各段階を追える検査者に当てはめると、段階を調べることはできても、結果への近道がないという還元不可能性は残ることになる。ただし、段階を調べられることだけでは、検査者が過程全体を理解しているとは言えない。

人間より有能な主体ならどれほどの不透明性を取り除けるかについては、後の説で見方が分かれる。アルバラード (Alvarado) はプレプリントで、計算の方法における不透明性のなかには、主体のどんな変化にも応じないものがあると論じる。すべての認識的主体が共有する特徴、たとえば有限であることから生じる不透明性を、主体中立的 (agent-neutral) と呼ぶ。これは、チャーニアク (Cherniak) がどの有限な主体にも帰した種類の限界である (第5章第6.5節)。アルバラードが挙げる例は、サイモンズ (Symons) とホーナー (Horner) から取った見積もりで、現実的な規模のプログラムのあらゆる経路を確かめるには宇宙の年齢の何倍もの時間がかかるので、何百万年もの時間を費やせる主体にとってもその不透明性は残る、というものである。またアルバラードは、主体の資源に応じて変わることのない、過程そのものの特徴から生じる不透明性を、主体独立的 (agent-independent) と呼ぶ。その例には、確率的で極度に過剰決定された過程、結合されたシミュレーションの構成要素のあいだで束ねられた丸め誤差、部分的に無作為な初期値から自動化された訓練によって定まる深層ネットワークの重みなどがある (Alvarado 2021, 10–15)。これに対してボーゲとグリュンケ (Grünke) は、同じくプレプリントで、シミュレーションも機械学習の方法もアルゴリズムとしては透明なので、ラプラスの悪魔のような理想化された主体ならその段階をたどれると論じる。そのため、本質的な不透明性も含めて、計算の不透明性は現実の主体の認識的な条件に依存する偶然的なものだとされる。ボーゲとグリュンケが根本的でありうると推測するのは、二人がモデル不透明性と呼ぶものである。すなわち、訓練されたネットワークが行うことから、その出力の背後にある物理的な仕組みのモデルへと導くアルゴリズムはないように見える。二人はこの困難を、まだ思いつかれていない対立仮説の問題と密接に関係するものとみる (Boge and Grünke 2019, 11–14, 26–27, 30)。ボーゲとグリュンケが考えているのは人間の主体だけであるが、人間より有能な探究者に当てはめると、二人の説では、その探究者は、結果が計算された経路の不透明性を取り除きながら、なお対象についての不透明性に直面しうる。これは、説明が、探究の説明しようとする現象を十全に表しているかという、第5.3節で立てる第三の十全性の問いにかかわる事柄である。アルバラードの説では、経路そのものの不透明性の一部も、どの有限な探究者にも残る。

第5.1節は、不透明な要素の役割を取り上げ、候補の生成と、それを受け入れる根拠づけとを分ける。第5.2節は、不透明性をほかの次元とあわせて見取り図にし、第5.3節は、根拠づけられた依拠にとって透明性が必要か十分かを問い、第5.4節は、後からの再構成が、後付けの言い訳という意味での単なる合理化とは限らないことを示す。

5.1 候補生成と最終的な根拠づけ

不透明な要素が果たしうる役割のうち、候補となる答えを生み出す役割と、それを受け入れる根拠を与える役割とは、分けることができる。ライヘンバッハ (Reichenbach) が導入した発見の文脈と正当化の文脈の区別 (第5章第4.1節) と、以下で見るデュエデ (Duede) によるその深層学習への適用のもとでは、生成における不透明性について、限定的な許容が導かれる。

候補となる答えを生成する経路は、それを受け入れることを根拠づける経路である必要はない。

推測は、類推、美的判断、偶然の観察、夢、ランダム探索、あるいは不透明なモデルから生じうる。これらの説では、その起源がその根拠づけを提供する必要はない。デュエデは、モデルが発見の文脈で作動し、結果が別個の手段を通じて正当化されるならば、深層学習の不透明性は科学的発見と両立しうると直接に論じている (Duede 2023)。あらゆる生成上の動きについて公共的な合理的再構成を求める要件は、機械による発見と同じく、通常の人間による発見にも当てはまるだろう。他の説は、この許容を狭める。信頼性主義の説 (第2章第4.1節) や、第5.3節で見る徳認識論的な読みでは、源泉の信頼性や、認められた能力が、それ自体で根拠づけを与えうる。それらの説では、候補がどのように生成されたかが、それを根拠づけるものの一部でありうるので、生成器を評価することが、結果を評価することの一部になる。

デュエデとデイヴィー (Davey) は、より鋭い数学的事例を展開している。不透明なシステムは、人間が見渡すには大きすぎる形式的に表象された証明を産出するかもしれないが、透明な証明検査器が保存された導出を検証する。二人の説では、人間が生成過程を把握することも証明を一步步ずつ点検すること

もできないにもかかわらず、その検査器はア・プリオリな数学的知識を支える (Duede and Davey 2025)。この結果は意図的に領域特定のである。それは形式的な証明対象と、数学的役割が透明な検査器とに依存している。それは、最初の不透明なモデルの後に別の不透明なモデルを置くだけで経験的主張が根拠づけられうることを示すものではない。それでもこの結果は、論理的分離を鮮明にする。すなわち、候補生成における不透明性は、結論を根拠づける経路における不透明性である必要はない。たとえば、リキッド・テンソル実験 (第1章第3節) は、そのような経路のうち人間の担う部分がどこにありうるかを示している。ショルツェは、形式的な定義と定理の言明は驚くほど読みやすいと感じたが、証明はまったく読めなかったと書いた。そして形式化の担い手たちは、Leanの実数が唯一の完備順序体をなすことなど、関わる対象についての形式的に証明された結果を、検査された言明が意図された対象にかかわっていることの証拠として付け加えた (Scholze 2021; Buzzard 2022)。検査器が人間の判断に残したのは、導出の各段階が妥当かどうかではなく、検査された言明が意図された定理であるかどうかであった。

ペルク (Pelc) は、証明の背後にある形式的な導出が定理への確信に寄与するという形式主義の見方に反対して、形式的な検査が数学の信頼性に何を加えるかについて、デュエデとデイヴィーよりも懐疑的な見方をとった。ペルクは次のように論じた。十分に短い導出を実際に行えるなら、導出は確信を高めるだろう。しかし深い定理についてそれが可能であることは、現在の知識の状態では知られていない。そのため現在のところ、証明と導出のあいだのどんな結びつきも確信を高めることはできない。定理への確信はむしろ社会的な過程を通じて築かれ、結果は、その証明が他の人々によって検証し直され使われるにつれて、より信じるに値するものになる (Pelc 2014, 5–7, 25–26, 29)。ペルクはまた、ラヴの二つの異議を引いている。証明は、そこに現れる用語の意味に基づく議論によって固められており、その意味は形式的な計算体系では捉えられない。また、非形式的な証明を導出として書き写すには論理学者の技能が要り、その書き写し自体は機械的に検査できない。そのため、証明検査器が妥当という判定を出しても、それが安心させるのは、形式化と検査が正しく行われたことだけかもしれない (Pelc 2014, 6–7)。ペルクは二つめの異議の一部を退ける。妥当という判定がもし得られるなら、それが高めるのは非形式的な証明への確信ではなく定理そのものの確信だ、とペルクは考える。ペルク自身の異議はむしろ、多くの定理については十分に短い導出がないので、そうした判定はそもそも得られないという点にある (Pelc 2014, 6–8)。リキッド・テンソル実験では、形式的な導出が実際に行われた一方で、検査された言明が意図された定理を表しているかどうかの判断は、人間の参加者に残された。ペルクが引く書き写しへの異議をこの事例に当てはめると、問題になるのは、形式的な言明と意図された意味との関係である。どちらの観察も、定理への確信は社会的な過程から来るというペルクのテーゼを決着させはしない。

オチガメ (Ochigame) は、証明の歴史と社会学の観点から自動化された数学について書き、証明の基準そのものを、数学者たちが改めてきたものとして扱う。オチガメはハッキング (Hacking) の対比を借りる。一つは、証明を理解し、それを一度に心に取めることができるというデカルト的な証明の理想であり、もう一つは、すべてのステップを機械的に確かめられるというライブニッツ的な理想である。そのうえで、両者から派生した実践上の基準をたどる。そのうち、計算機に符号化された証明が推し進めてきた基準は、すべてのステップがプログラムによって確かめられ、その公理的基礎から導けること、そして「証明された言明の符号化を理解するか信頼する」ことである。そのためオチガメは、こうした証明も、あるコードが主張された言明に対応しているという社会的な合意を必要とし、審査のための社会的制度をなくしはしないと論じる。結果の重要性和価値を判断するには、そうした制度がなお必要である。この説では、数学者たちはいま二つの決定に直面している。一つは、計算機に符号化された証明を必須とするかどうかである。もう一つは、構成し検査したプログラムを信頼することで受け入れることになる、まったく見通せない、深く自動化された証明を受け入れるかどうかである。オチガメはどちらについても立場を擁護しない (Ochigame 2024, §§2–4)。オチガメの説では、リキッド・テンソル実験が人間の参加者に残した判断は、共同体がなお改めうる証明の基準に属する。そして第二の決定は、どの成員もその大筋さえ理解しない結果を S が受け入れるかどうかを問うている。

生成と根拠づけを切り離してよいという許容は、生成を認知的に無関係にするものではない。その経路は、データ漏洩、事前の露出、選択効果、開示されていない人間による選別、重複した作業、あるいは評価者への依存を明らかにしうる。また、独立しているとされる検査も、生成器の訓練データ、測定誤差、あるいは背景仮定を単に再生産するだけなら、独立の検査にはならない。この区別は認知的役割のあいだの区別であり、必ずしも厳格な時間的順序ではなく、諸役割が因果的に独立していることの保証でもない。

5.2 不透明性の多次元的な見取り図

生成と根拠づけという区別は、関連するすべての不透明性を分類するには粗すぎる。クリール (Creel) は、複雑な計算システムにおける機能的透明性、構造的透明性、実行の透明性を区別している。おおよそ、プログラムが実現するアルゴリズムについての知識、コードがそのアルゴリズムをどう実現しているかについての知識、特定のハードウェアと入力データでの特定の実行についての知識である。クリールは、三つは解離可能であり、システムはこれらのうち一つの仕方で透明でありながら、他の仕方では透明でないことがありうると論じる (Creel 2020, §§3a–3d)。その分析が扱うのは、そうしたシステムの熟練し知識のある作り手と使い手にとって残る不透明性だけであり、隠蔽によって生じる不透明性や、使い手が関連する専門知識を欠くことによって生じる不透明性は扱わない (Creel 2020, §3)。

サリヴァン (Sullivan) は、困難の別の源をモデルの外に見いだす。その説では、不透明なモデルが与える理解を主に制限するのは、モデルの複雑さや実装の不透明さではなく、サリヴァンが結びつきの不確実性 (link uncertainty) と呼ぶもの、すなわちモデルとその対象である現象とを結ぶ結びつきを支える科学的・経験的な証拠の欠如である。その理由は比較による。診療記録から多数の疾患を予測するモデル、ほくろの画像からメラノーマを分類するモデル、顔画像から性的指向を分類するモデルという三つの深層学習モデルは、不透明さの程度が同じでありながら、与える理解が異なる。そしてサリヴァンの分析では、その違いは、モデルの外からの証拠が各モデルをどれほど対象に結びつけているかに従っている。メラノーマの場合には皮膚科学の豊富な知識がその結びつきを支えるのに対し、顔画像の場合には既存の証拠がむしろ反対の方向を示しており、モデルは身だしなみのような文化的な手がかりを追っている。ただし、この主張には限定がある。それが成り立つのは、入力と出力しか知られていないという最上位の水準でモデルが不透明ではない限りにおいてであり、深層ニューラルネットワークはその水準では不透明ではないとサリヴァンは論じる。その全体の構造と、それを生み出すモデル化の過程は知られているからである (Sullivan 2022b, §§3.2, 4.3, 5.2–5.3, 6)。

この説を踏まえた論文で、サリヴァンは帰納的リスクを、結びつきとモデルの内部の両方にかかわらせる。帰納的リスクとは、証拠が仮説を確実にしえないもつて仮説を受け入れたり退けたり、モデルを用いたりするときに残る誤りのリスクである。帰納的リスクからの論証によれば、どれほどの支持があれば

ば十分かは、ありうる誤りがどれほど悪いかに依存し、その悪さには非認知的な帰結、つまり真理以外の事柄にかかわる帰結、たとえば人々への害も含まれる（第11章第4節）。サリヴァンの説では、モデルがある現象を説明できると受け入れることには、こうした判断が含まれる。というのも、モデルのデータが適切であること、そしてモデルが捉える反事実的依存、すなわちある特徴が違っていたら対象の何が違っていたかが対象について成り立つことを、ともに受け入れなければならないからである。誤ってそれを受け入れることが害をもたらす場合には、より多くの証拠が要求される。顔画像の分類器について、サリヴァンは、そうした依存関係が科学的に論争的であること、骨相学や人相学が人種差別、性差別、優生学の正当化に用いられてきたこと、そして同性愛者が受けている差別を挙げる（Sullivan 2022a, 1067–1070）。そして、ここから二つの帰結を導く。外的な側では、帰納的リスクが大きいほど、結びつきの不確実性を減らすのに必要な独立の証拠も強くなる。そうした証拠には、たとえばモデルを超えたさらなる経験的研究があり、対象が統計的相関である場合には複数のモデルにわたる頑健性のチェックがある。内的な側では、データから学習されたモデルは依存関係を明示的な形では含んでいないので、依存関係は訓練の後に取り出さなければならない。そのための手段には、たとえば分類に最も寄与する画像の部分を強調表示するサリエンシーマップ（saliency map）がある。そして帰納的リスクが大きいほど、取り出されたものが単なるアーティファクトではなくモデルの働き方に忠実であることを確かめる基準も高くなる。サリヴァンが念頭に置くのは主にクリールのいう機能的透明性であるが、利害によっては構造的透明性や実行の透明性も求められるとしている。さらに、研究者の問いと対象を固定しても、これらの要求は利害とともに高まると論じ、その主張を、モデルの実地での使用ではなく説明と理解に限定している（Sullivan 2022a, 1070–1074）。それゆえサリヴァンの説では、解釈可能性の手法の現在の限界は、利害の小さい現象の理解を妨げるとは限らないが、モデルが重大な社会的帰結を伴う事柄にかかわる場合には理解を妨げうる（Sullivan 2022a, 1072–1073）。バイスパルトは、説明的理解を妨げるものを、このどちらの概念からも区別する。その説では、高度に不透明な計算機シミュレーションでも、関連する説明的な法則を含むことが知られている場合は多いので、ハンフリーズの意味での認知的不透明性が妨げになっているのではない。また、ネットワークが対象における相関を捉えているという強い証拠があれば、結びつきの不確実性は小さくなるが、それは説明の助けにはならない。ただし、ネットワークが含む説明が実際の説明であるかどうかは、この不確実性の一部に数えうるとバイスパルトは認めている。その論じるところでは、欠けているのは、ネットワークが何をしているかを対象となる系の変数や過程によって記述する、情報量のある記述である（Beisbart 2025, 24）。

以下の診断のためのグリッドは、クリールとサリヴァンの区別をもとにここで組み立てたものであり、クリールが扱わない不透明性の形も含むように広げである。「対象」に挙げた対象との結びつきと、「利害」に挙げた誤りのコストは、サリヴァンの二つの分析に由来する。これに対して、同じく「利害」に挙げた分布シフトは第2節から引き継いだものであり、可逆性、新しさ、時間的視野は、サリヴァンが論じる考慮事項には含まれていない。「主体」の座標には、ハンフリーズの定義を機械学習に合わせて用いたゼドニク（Zednik）の議論という先例がある。ゼドニクの説では、不透明性は主体に相対的である。機械学習システムの利害関係者は、運用者、執行者、決定の対象者、データの主体、作り手、審査者などのように役割によって異なる。そして、システムのどの要素が認知的に関連するか、すなわち主体がシステムのふるまいを説明するのに引き合いに出せる要素はどれかも、主体とその役割によって異なる（Zednik 2021, §§2.2, 3.1–3.2）。ゼドニクは人工の利害関係者を考えていないが、「別のAI」という値に当てはめると、主体が定めるのは、システムが誰にとって不透明かだけでなく、そのどの要素が関連するとみなされるかでもある。

各行は、一つの座標と、その座標を定める問いを示す。ある不透明性の所在は、四つの問いすべてに答えることで特定される。各問いに挙げた項目は、座標がとりうる値のよくある例であり、ほかの対象、役割、主体、利害を加えることもできる。

表3.1 研究システムにおける不透明性の所在を特定するための座標

座標	問い
対象	不透明なのは、生成器か、コードか、データか、学習された関係か、証拠か、対象との結びつきか、制度か。
役割	それは探索、測定、予測、根拠づけ、説明、意思決定、訂正のいずれを果たすのか。
主体	それは開発者、運用者、科学者、査読者、影響を受ける公衆、別のAI、集合体のいずれにとって不透明なのか。
利害	誤りのコスト、可逆性、新しさ、分布シフト、時間的視野はどうか。

このグリッドは、四つの必須モジュールのリストではない。それは、ある場所での不透明性が、あらゆる場所での不透明性として扱われるのを防ぐ。あるモデルは、領域の科学者にとって計算的に不透明でありながら、その領域限定的な信頼性は公共的にテスト可能でありうる。透明なアルゴリズムが、不透明なデータセットや妥当でない測定モデルに依拠していることもありうる。人間が読める根拠説明が、追跡不可能な証拠の経路を隠していることもありうる。分散した協働は、すべての参加者にとって局所的に不透明でありながら、信頼できる集合的検査を維持しうる。

5.3 透明性は必要か、十分か

本節で比べる説は、依拠の理由をどこに置くかで分かれる。信頼できる過程か能力を認められたモデルか（本節の分析によれば、どちらもシステムの実績によって示されうる）、システムの構築のされ方とその根拠か、それともシステムの働きへの接近か、である。いくつかの説では、内的な透明性は、根拠づけられた依拠にとって普遍的に必要なわけではない。デュランとフォルマネクは計算的信頼性主義の説を展開しており、ここでは、シミュレーションが信頼できる過程であるとき、研究者は本質的に不透明なシミュレーションの結果を信じることを正当化されうる。二人はその信頼性の源として四つを挙げる。検証と妥当性確認、頑健性分析、理論・モデル・構築技法の（不）成功した実装の履歴、そして専門家の知識であり、専門家の知識は単独ではかなり弱い源だとみなす（Durán and Formanek 2018, §§3.1, 4）。2018年の定式化では、二人の説は純粋な外在主義ではない。第二の段階で、これらの源がそれ自体として信頼できると示されることを求め、ある検証手続きを信頼できると考える研究者は、そう考える理由を挙げなければならないとする（Durán and Formanek 2018, §3.2）。これは実質的な立場であって、理論中立的な解決ではない。それはなお、シミュレーションと同一視された信頼できる過程をどう類型化するか（第2章第4.2節）、信頼性の源の信頼性をさらに別の源で示すという連鎖がどこで終わるか（第11章第5.2節）、信頼性が確立された環境が安定したままかについて、問いを提起する。デュランはのちに、計算的信頼性主義をアルゴリズムの外在主義的な認識論として示している。正当化が、アルゴリズムの内部の働きを示すことではなく、アルゴリズムの外にある信頼性の指標に基づく、という意味においてである。そしてデュランはそれ

を、アルゴリズムの出力を信じる人間の主体について述べている (Durán 2025a, §2)。この意味での外在主義はアルゴリズムにかかわるものであり、信じる者が理由をもたなければならないかどうかにかかわるものではない。

グロテ (Grote)、ジェニン (Genin)、サリヴァンは、これらの問いの最後のもの、すなわち信頼性が確立された環境が安定したままかという問いを、機械学習について押し進める。三人の読みでは、この説のうち、実装の履歴と、関連する専門家の信頼にかかわる安全装置は、モデルが認知的によく秩序づけられた環境に埋め込まれていることを前提している。ところが刑事司法、臨床医学、科学のような領域に実装される過程で、機械学習モデルはその領域の認識規範を攪乱しうる。こうした困難は、この説が与えるはずのものを前提している可能性を示す、と三人は論じる (Grote, Genin, and Sullivan 2024, 6)。

大塚は、深層学習モデルについて別の診断を示す。大塚の説では、従来の統計的な手法は、仮説が証拠によってどこまで支持されるかについて明示的な基準と理論的な理解を与えるが、深層学習には、それに並ぶ原理的な評価の方法がない。正当化はむしろ経験的なもので、ベンチマーク上で個々の事例ごとに確かめられ、そのため信頼性は名前をもつ個々のモデルに結びつけられるようになる。大塚は、このようなモデルないし行為者にもとづく正当化は、正当化の根拠を行為者の認知的な能力、すなわち徳に求める徳認識論 (virtue epistemology) とよく合うと示唆し、現在のAI研究の共同体は、ベンチマークの得点で測られる認知的な徳にもとづいてモデルの性能を正当化しているように見えるとする。そのうえで大塚は、そうした徳とは何か、機械がそもそもそれをもつのか、もつとすればそれをどう確かめられるのかを問う (Otsuka 2026, "Understanding the Machine Virtues")。デュランとフォルマネクの説と大塚の説は、異なるシステムを扱っているが、依拠を基礎づける信頼性の所在を異なる仕方で定める。上に挙げた四つの源に信頼性が支えられるシミュレーションの過程に置くのか、能力を認められた個々のモデルに置くのか、である。

大塚はさらに、そうしたモデルの判断を正当化する二つの方法を、ソーサ (Sosa) による動物的知識 (animal knowledge) と反省的知識 (reflective knowledge) の区別に対応させて区別する。動物的知識は、動物や人間が信念を形成する過程の適切な働きによって得る知識であり、反省的知識は、それらの過程の信頼性についての高階の自覚を伴う。一つの方法は、モデルが動物的知識をもつかどうかを問う。生理学的な調査が、カエルの知覚の仕組みがどのようにしてハエを確実に捕らえさせているかを示しうると同様である。モデルのアーキテクチャや訓練についての研究は、機械論的解釈可能性 (mechanistic interpretability) の研究を含めて、モデルの帰納的な性能の基礎にある機構を探る。もう一つの方法は、モデルが反省的知識をもつかどうかを問い、モデルが結論の拠りどころとする内的な理由に注目する。大塚はソーサに従って、反省的知識は、一階の動物的知識が適切に働く条件を評価するものであり、機械学習ではそれがモデルの頑健性や訓練領域を超えた外挿の問題にあたる。そして、説明可能なAIの研究を、深層学習モデルの反省的知識に向けられた研究の流れの一つに数える (Otsuka 2026, "Understanding the Machine Virtues")。大塚は、より早い著作では、この二つの問いについて一歩踏み込んでいる。そこでは、深層学習モデルがすでに何らかの形の動物的知識をもつことは否定しがたいが、反省的知識をもつか、あるいは今後獲得するかは明らかでない、とされる (Otsuka 2023, 136–139)。そうしたモデルの内部表象には唯一の正しい解釈がないかもしれない、という大塚のさらなる議論には、第5.4節で立ち戻る。

認識的技術としてのAIについての論文のプレプリント版で、アルバラードは、AIをふさわしい仕方信頼するには、その働きへの接近が必要だと論じる。その説では、AIは認識的技術であり、知識の源として依拠される。そうした技術にふさわしい信頼は認識的信頼だけであり、それは認識的な理由に基づかなければならない。そしてその理由を評価するには、その技術がどう働くかのうち認識的に関連する要素に接近できることが要るので、説明可能性がきわめて重要な役割を果たす。金槌は仕組みを知らずに受け入れてよいが、そうした接近がなければ、認識的技術に向ける認識的信頼を正当化するのはいまますます難しくなり、不透明な認識的技術の受容を正当化するのはいまますます難しくなる、とアルバラードは論じる。慎重な考慮に訴えるのでは足りない、とも言う。そしてAIが本質的に不透明であるなら、私たちはAIを信頼することができない、と結論する (Alvarado 2023, §3)。アルバラードは、そうした技術は健全な認識的手段によって評価されなければならないという要求の典拠として、デュランとフォルマネクを引く。しかし、二人が検証、妥当性確認、実装の履歴といった源によって不透明なシミュレーションへの依拠が基礎づけられるとみるのに対し、アルバラードは、システムがどう働くかの理解、したがって説明可能性が必要だとみる。二つの立場は、システムへの依拠や信頼を正当化する理由が、システムがどう働くかにかかわらなければならないのか、それともその実績にかかわってよいのかをめぐって異なると解釈できる。本質的な不透明性は主体の本性に相対的なもので、接近に理由を置く説の判定は、依拠する者によっても変わりうる。人間の成員にとって本質的に不透明な計算をたどれる S の人工の成員は、そうした説が求める接近をもちうるかもしれない。ただし、アルバラードの主体中立的な不透明性の説では、経路の一部は、どの有限な依拠者にとっても不透明なままである (第5節)。

より古い立場は、依拠の理由を完全に実績のうちに置く。ライヘンバッハは、系列の早い段階でその頻度の極限を言い当てられる千里眼を考察した。ライヘンバッハの説では、千里眼がどう働くかについては何も知られている必要がない。その予言を後の観察と比べ、それが確かめられれば、帰納そのものが、直接の帰納的な推定よりもその予言を優先するよう勧める。一方、そうした点検なしに千里眼を認めることはできない (Reichenbach 1938, 353–354, 358–359, 363–364)。いま述べた区分に当てはめれば、ライヘンバッハの見方はその実績の側に位置し、点検そのものを、その源の過去の予言の系列についての帰納にする。そのため、その系列をどの類型のもとで数えるかという第2章第4.2節の問いが再び生じる。

計算科学については透明性を要求として手放したハンフリーズは (第5節)、不透明なシミュレーションに依拠する理由の一部を、それがどう構築されたかに置いた。それ以前の二つの論文でも、構築をモデルの正当化の一部とし、構築の前提を、検査に先立つ正当化に結びつけている (Humphreys 1995, abstract; Humphreys 2002, abstract)。バランディアラン (Barandiaran) は、境界条件の摂動に対する安定性、スケール不変性、得られる解析解との一致といった長所が、計算モデルとその解とのつながりを不透明なものとして扱う試行錯誤によって達成されうるというハンフリーズの説明を引用している (Barandiaran 2008, 53)。アーヴァイン (Irvine) はさらに、理論はパラメータ族を正当化し、その細部は経験的データによって定められるというハンフリーズの指摘を引用している (Irvine 2016, manuscript 17)。ハンフリーズはまた、データを生む物理的・計算的な過程についての知識が、アーティファクトの検出と修正を支えうると論じる。成功した使用の記録があっても、将来の対象に十分似たものがそこに含まれず、その対象で生じるアーティファクトを見分けられないことがありうるのである (Humphreys n.d., §4)。したがって、構築に由来する根拠は、デュランとフォルマネクの挙げる源の一つである実装の履歴とは異なるが、その履歴による支持を排除するものではない。ハンフリーズが論じたのは人間が構築するモデルである。人工システムが構築し、あるいは選ぶモデルに当てはめると、 S の成員が構築に由来する理由に頼れるのは、構築の根拠が記録され、彼らがそれを評価できる場合に限られる。

サイモンズとアルバラードも、シミュレーションについて書くなかで、科学における正当な信頼は実用上の理由だけには基づかないとする。ウィンズバーグとアルバラードの報告によれば、サイモンズとアルバラードの説では、コンピュータ・シミュレーションの認識論の課題の一部は、認識的に不透明なシミュレーションを用いる科学者の立場と、神託を信じる者と神託との関係との違いを説明することにある。実用上の考慮と認識的な考慮は並び立ち、実用上の理由を最終的に説明し根拠づけるのは認識的な理由である。また同じ説では、シミュレーションは、それを構築するのに用いた方法から根拠を受け継ぐのではなく、それ自身の根拠をもたなければならない (Winsberg and Alvarado 2026, §2.3)。ハンフリーズの説で構築が正当化に寄与することは、個別に実装されたシミュレーションの根拠は何かという、サイモンズとアルバラードの問いを決着させない。構築に着目する別の立場が、ウィンズバーグの説である。ウィンズバーグとアルバラードの報告によれば、その説では、シミュレーションを作る技法は、長い使用の歴史を通じて信頼の根拠を得る (Winsberg and Alvarado 2026, §2.1.1)。この見方では、構築と蓄積された経験とがともに働く。その根拠は、実績だけにあるのでも、計算の段階を一つひとつ見渡すことにあるのでもない。

深層学習モデルについては、そうした依拠が頼る根拠が同じ形では得られない、と論じる後の説が二つある。デュエデは、科学者がそうしたモデルに依拠することを正当化されうること自体は否定しない。そのうえで、その正当化が、装置に依拠する場合の正当化にも、専門家に依拠する場合の正当化にも還元されないと論じる。その説では、テストデータ上のモデルの実績だけでは足りない。同じようによい性能を示す多くのモデルが、実際に用いられると大きく異なるふるまいをしよううえ、現実のデータはしばしば裾の重い分布に従うからである。また、モデルを世界と結ぶ因果的な過程がその出力を媒介しているわけではないので、モデルは物理的に媒介された装置ではない。さらに、手続きが数値解析の確立した原理に基づくシミュレーションとは異なり、モデルは、現在の解釈可能性の研究の水準では、学習された手続きがどんな原理を実装しているのかを明らかにしない。理論的に媒介された装置を正当化するには、それが必要である。そして、モデルが出力について理由をもつと認めたとしても、その理由は、専門家の理由のように他者が評価できるものではない (Duede 2022, §§2-4)。モデルそのものを観察と実験の方法で調べる解釈可能性の手法は、新しい種類の信頼性をなす、とデュエデは結論する (§5)。アルバラードは、システムに依拠するうえで関連する要素のうち、誤りにかかわる要素は外せないと論じる。その説では、依拠に必要なのは、システムがどう働くかを見渡すことでも、失敗の率だけでなく、その誤りの性質と源についての知識である。そして深層ネットワークでは、そうした誤りはしばしばたどれない。同じ結果に多くの経路で達しうること、自然な入力が敵対的事例と同じように誤りを引き起こしうること、誤りが多くの重みに分散していること、誤った出力が正しい出力と同じ確信度で出されることによる (Alvarado 2025, §§3.2-3.3)。したがってデュエデの説では、シミュレーションを理論的に媒介された装置として扱う根拠は、深層学習モデルについては同じ形では得られない。誤りの性質と源を知るというアルバラードの要求は、ハンフリーズのアーティファクト論で、生産過程についての知識が診断に果たす役割にかかわる。アルバラードの説では、そうしたモデルからは、その知識がしばしば得られない。ラトコフ (Rathkopf) は、どのような設計でもそうした誤りを生むとみる生成モデルについて論じ、アルバラードのジレンマは、信頼性を評価する単位をモデルとすることを前提にしていると応じる。ラトコフの事例であるAlphaFold 3と、アンサンブル予報を生成する気象予測システムSEEDSでは、出力は、対象系についての独立の知識から組み立てられた手続きによって順位づけられ、除かれ、確信度の評点を付され、あるいはアンサンブルに組み込まれる。そのため、誤りをその源までたどらなくても、ワークフローは信頼できるものになりうる。ただしラトコフは、そうしたワークフローを組み立てられるのは、対象についての十分に成熟した知識がすでにある場合に限られると付け加えている (Rathkopf 2026, 13-15, 17-18, 20-21)。この応答は、求められるものをモデルの失敗についての知識から対象についての知識へと移す。そのため、*S* のための研究がその知識をまだ生み出していないところで、依拠は確保されないまま残る。

システムがどう働くかへの接近が必要かどうかについてどの見方をとるにせよ、内的な透明性は十分ではない。実行トレースは、出力がどのように計算されたかを忠実に示しつつも、データが対象を測定していること、推論が根拠づけられていること、あるいはモデルが現象を説明していることを示さないことがありうる。特徴帰属は、分類器内部の感度を明らかにしつつも、世界における因果的あるいは安定的な関係を確立し損なうことがありうる。ウッドワードは、導出によっては、そこでの「操作」が物理的にありうる操作ではなく、数学的あるいは概念的なものだと述べる。たとえば系の次元が異なるモデルを構成してその帰結を計算すれば、次元がどのような違いを生むかは示されるが、それは系の物理的な変形を記述しているわけではない (Woodward 2018b, §3)。モデルへの入力も、同じように、世界の側では何も変えないまま変えることができる。そうすればモデルが何を出力するかは示されるが、対象への介入が何をもたらすかまで示されるのは、モデルが符号化する関係がその介入のもとで不変である場合に限られる。流暢な推論の連鎖は、判読可能であって、しかも偽でありうる。

本書は、十全性の問いを少なくとも次の三つに区別する。

1. ある説明は、出力を産出した過程を忠実に記述しているか。
2. それは、利用可能な根拠がなぜその主張の受け入れを支持するのかを十全に明確化しているか。
3. それは、探究が説明しようとする現象または依存関係を十全に表象しているか。

完全な過程トレースは、科学的には何も明らかにしないまま第一の問いを満たしうる。証拠的論証は、対象となる規則性がなぜ成り立つのかに答えることなしに第二の問いを満たしうる。単純化された説明的モデルは、発見したシステムの計算を再現することに意図的に失敗しつつ、第三の問いに寄与しうる。これらの人工物は互いを支えうるが、いずれも自動的に他の代わりになるわけではない。

5.4 事後の再構成は自動的に合理化であるわけではない

事後の説明は、多くの場合ブラックボックスに対して与えられるものであり、第5節の冒頭で見たブラックボックスという語と同じく、事後の説明 (post-hoc explanation) という語も、二つの異なるものを指しうる。どちらも、結果が得られた後に作られる説明である。ひとつは、結果の後に産出され、過程、証拠、対象によって制約されていない説得的な物語である。それは何も説明することなしに理解可能に聞こえることがある。もうひとつは正統な再構成である。すなわち、醜い探索の後に見出された優美な証明、複雑な予測の後に展開された単純な理論、あるいは「なぜ」の問いに関連する関係を切り出す理想化されたモデルである。ここでいう再構成とは、結果そのもの、あるいはそれにどう至りうるか、なぜそれが成り立つかを、実際にその結果が得られた経路よりも見通しのよい形で述べ直すことである。

フライシャー (Fleisher) は、ブラックボックスを理想化して説明したものは、そのアルゴリズムが実際にしていることをそのままぞっていなくても、理解に寄与しようと論じる。フライシャーが取り上げるのは、事後的な説明可能AI (explainable AI, XAI) である。これは、分類器がすでに用いられた後で、それに手を加えることなく、より単純な第二のモデルを作り、分類器がなぜある判定を下したのかを説明する手法である。主な例であるLIMEは、説明しようとする入力を書き加えて変えた入力に対する分類器の出力に、線形モデルを当てはめる。こうした説明は、分類器とは異なる関数を計算し、分類器の出力をいつも再現するわけでもない、たんなる合理化にすぎない、という反論がある。これに対してフライシャーは、同じことは理想化された科学的モデルにも当てはまり、しかもそうしたモデルが本物の理解をもたらすことは広く認められている、と論じる。たとえば理想気体の法則は、気体分子が互いに作用しないものとして表す (第3.4節を参照)。ここでいう理解は説明的理解、すなわち、システムがなぜその判定を下したのかの理解である。フライシャーはポトチュニク (Potochnik) にならい、それを、その判定にかかわる因果的パターンについての情報を把握すること、つまり、その情報を受け入れ、それを利用したり操作したりできることとして扱う。そして念頭に置いているのは、システムの開発者、それを運用する人々、そしてその判定の影響を受ける人々の理解である (Fleisher 2022, 535–536, 540, 542–544, 547–550)。だからといって、フライシャーはどんな代わりのモデルでもよいとするわけではない。XAIを考えるうえでとくに役立つものとして、理想化についての文献から三つの特徴を取り出す。フライシャーの説では、それらは、表象のずれがたんなる歪曲ではなく理想化であることのしるしである。すなわち、そのずれは、モデルを理解可能なほど単純にし、結果に違いを生まない特徴をそれと分かるように示し、説明を必要とする人々にとって重要な因果的パターンに注意を向けさせる。これらのしるしは、経験的十全性や予測力といった、成功したモデルにふつう見られる特徴とあわせて考慮される。またフライシャーは、説明される入力の近くでは、元のモデルの出力とかなりの程度一致することを求める。そして、説明のためのモデルは科学的モデルと同じ基準で評価されるとし、異なるXAI手法がしばしば食い違うことから、現在の多くの使用例はその基準に届いていないかもしれないと認める (Fleisher 2022, 541, 551–556)。これらの条件を、説明のためのモデルから再構成一般へ広げれば、再構成が正当であるのは、それが制約を受けている限りにおいてだと考えられる。その制約とは、それが何を表しているかと主張するのか (対象)、何のための説明なのか (目的)、そして、元のものにどれだけ忠実か、データにどれだけ合うか、肝心な関係を保っているか、といった要件である。第5.3節の十全性の問いに照らすと、フライシャーの説明のためのモデルは、分類器そのものを対象とする。そのためそれらは、出力がどのように産出されたかという問いにかかわり、ただしその問いに、文字どおりではなく理想化された形で答える。これに対して次に見るモデルは、計算と並べて、現象を表すために保持される。

デ・レフトとディークスは、二人の説では理解をもたらさない計算と並べて、科学者が理解可能なモデルを保持する実践を記述している。量子色力学では、信頼できる検証は、ブラックボックスとして働く数値的な格子法によってしかできず、それだけでは洞察は得られない。これに対してMITバグ模型は、その予測とデータとの一致がそこそこにとどまるものの、視覚化できるモデルとして、格子計算を行わずにハドロンの構造について定性的な予測を立てることを可能にするので、理解をもたらす、と二人は論じる。気象学では、渦位 (potential vorticity) を用いた推論が、数値予報と並んで定性的な像を与える (de Regt and Dieks 2005, §§4.1–4.2)。こうしたモデルは、計算が何をしているかを説明するものではない。計算と並べて、現象そのものを表すために持たれるものである。そのためこれらの事例が関わるのは、第5.3節の第一の十全性の問い、すなわち説明が出力を生んだ過程をどれだけ忠実に述べているかではなく、第三の問い、すなわち探究が説明しようとする現象をうまく表しているかである。格子の事例は、理論からの予測の成功にはその理論が使い手にとって理解可能であることが必要だという、第2節で見たデ・レフト自身の主張にも関わる。ただし、この議論では、両者は関係づけられていない。その要件が、予測を得るための個々の手続きではなく、使い手による理論の把握に関わるとすれば、両者は両立することになる。そうした把握を支えるのが、バグ模型のようなモデルである。

ハドロン模型についての事例研究のなかでバグ模型を論じているハルトマン (Hartmann) は、このモデルを保持する別の理由を挙げる。ハルトマンの説では、物理学者たちがMITバグ模型を、経験的十全性がそこそこにとどまるにもかかわらず検討の対象として採り続けているのは、使いやすく拡張しやすいからだけではなく、それに伴う物語がよいものだからである。その物語は、量子色力学から演繹することはできないが、量子色力学に照らしてもっともらしく、模型が取り出す特徴がなぜ重要で、他の特徴がなぜ無視できるのかを論じる。ハルトマンは、模型を、解釈された形式体系とそのような物語とを合わせたものとして扱う。物語は理論と矛盾してはならず、ハルトマンはこれを必要な基準とする。はるかに弱いとする形式的でない基準には、もっともらしいことと、物理的機構の理解を助けることが含まれる。そしてハルトマンは、模型とその物語がどのように理解をもたらすのかを明示するのは非常に難しいと認める (Hartmann 1999, 10–11, 17)。同じ事例研究で検討される別のハドロン模型では、模型に閉じ込めが欠けていることを論じ去るために物語が用いられた。その物語の帰結の一つ、すなわち高く励起したハドロンほど有効ポテンシャルの形に敏感なはずだという予想は、数値研究で確かめられた (Hartmann 1999, 16)。ハルトマンの事例を第三の十全性の問いと結びつけると、計算と並べて保持される模型の物語がこの問いに関わるのは、その物語が重要でないと脇に置くものが、それ自体確かめられている範囲においてだけだと考えられる。

さらに別の議論は、モデルの内部の推論に、そもそも唯一の正しい言い表し方があるのかを問う。大塚は、深層学習モデルの内部表象を評価することを、クワイン (Quine) の根元的翻訳になぞらえる。根元的翻訳とは、まったく知らない言語を話し手の振る舞いだけから訳す状況であり、クワインは、そこでは互いに両立しない複数の訳がすべての証拠に同じようによく合うと論じた (Otsuka 2021, 11–12; Otsuka 2026, "Can We Understand What Machines Think?")。モデルが「猫」とラベルづけするものは、われわれがその概念で理解している動物なのかもしれないし、大塚が引く研究が示唆するように、猫に特徴的な性質や背景の組み合わせにすぎないのかもしれない。そして大塚は、そのどれが正しいかは一意には定まらず、そもそも唯一の正解はないかもしれないと論じる。それでも大塚は、そうしたモデルの説明の探求は簡単には退けられないとする。モデルが何を存在するものとして認めているかは、その社会的な応用にとって重要だからである。自動運転に用いられる認識モデルがどのようなものを存在として認めているかがある程度分からない限り、人はそれに自らの生命を預けようとはしないだろう、と大塚は述べる (Otsuka 2021, 12)。この見方に立てば、モデルの推論を人間の概念で言い表すとき、その推論への忠実性という要件には、計算そのものは正確にたどれる場合でも、忠実さの基準となる唯一の事実がないかもしれない。第8章第3.1節は、同じようによい言い表し方が複数ありうるというこのクワイン的な懸念を、適切な言い表し方がそもそもないかもしれないというクーンの懸念と対比する。

いずれにせよ、ここで重要な問いは、ある説明がいつ産出されたかではない。それは、その説明がどの対象を説明すると主張しているのか、そして何がその説明をその対象に対して十全なものにするのか、である。

人工システムによる研究にとって、本小節の事例が示すのは、出力をどう評価するかだけではなく、正統な再構成が、それ自体、研究のもたらす達成でありうるということも示している。探索の後に見出された優美な証明、複雑な予測の後に展開された単純な理論、計算と並べて持たれる理解可能なモデルは、根拠づけられてはいるが理解されてはいない結果から、理解へと進む一歩である。強化した予測の事例でいえば、規則性がなぜ成り立つのかの理解に向けた試みは、この道をたどりうる。再構成を作るのがシステムであっても、その使い手であっても、両者の協働であってもよい。そうした再構成が理解に数えられるかどうかは、上で述べた制約による。それが表すと主張する対象と、それが何のためのものかという目的に合っていること、そしてハルトマンの事例が示唆するように、それが無視してよいとしたものが、それ自体確かめられていることである。大塚の根元的翻訳との比較が当たっているなら、そうした試みのたどる道についてさらに一つの帰結が出る。深層学習モデルの内部の推論を人間の概念で言い表したものには、忠実であるべき唯一の事実がないかもしれない。これに対して、現象のモデルが応えるのは第三の十全性の問いであり、それは計算ではなく対象にかかわる。そのため、現象の理解を求めるなら、その道は第一の十全性の問い、すなわち説明が出力を産出した過程を忠実に記述しているかという問いの決着を待つ必要がない。こうしてもたらされた理解がどこに実現するのか、すなわちシステムか、*S* の成員か、その両方かは、第6節の問いである。*S* が人間であれば、再構成は人間の成員が使える形でなければならない。受け手が、モデルの表象をそのまま扱える人工システムであれば、人間の概念への言い表しはそもそも要らないかもしれない（第8章第3.4節を参照）。

本節の冒頭で挙げた二つの反論の行方は、それぞれ異なる。不透明性が、第4節が認めた知識の根拠づけを損なうかどうかは、根拠づけられた依拠についての説による（第5.3節）。アルバラードの説のように、システムの働きへの接近を求める説では損なわれうるが、依拠の理由をシステムの実績や構築のされ方に置く説では、それぞれに残る問いはあるものの、損なわれるとは限らない。これに対して、不透明性によって誰もその規則性を理解していないことが決まる、ということにはならない。上で見たように、理解は、再構成や、計算と並べて持たれるモデルを通じて求めうるからである。

これらの区別は、そうした仕事をどう記録するかにもかかわる。出力が第5.3節の三つの十全性の問いのそれぞれについて評価されるものであるなら、人工研究システムは、人間向けのインターフェースがそれらを一つにまとめる場合でも、過程の来歴、証拠的支持、対象レベルの説明、教育的な提示を、区別できる人工物として保持する必要がある。それぞれがどの機能に役立つかは、第9節が述べる。これら四つすべてを **推論** (reasoning) と呼ぶと、明晰な語りが根拠づけと理解の両方の代わりを務めてしまいやすくなる。

6. 誰が理解するのか

人間が研究を行う場合、結果を生んだ者はたいいてい、それを理解する者のうちにも含まれるので、理解の担い手を別に名指す必要はあまりない。人工システムが結果を生む場合には、それが理解するものであるとは限らないので、担い手がそれ自体として問いになる。そのため、研究が理解をもたらしたという主張は、少なくとも次の三つの異なる主張でありうる。

1. AIが、人間であれ人工のものであれ、他のエージェントあるいは共同体が理解するのを助ける。
2. AI自体が理解する。
3. 人間とAIの集合体あるいは人工の共同体が理解する。

これらはいずれも、誰が理解するのかにかかわる。そのいずれかに実現した理解が *S* の認識的状況に入ったかどうかは、さらに別の問いであり（第1章第2節）、以下で取り上げる。

第一の主張は第二の主張を要求しない。望遠鏡は、現象を見ることなしにそれを開示しうるし、探索手続きは、定理を理解することなしに証明を可能にする。クレン (Krenn) らも、科学における AI について同じ分離をしている。クレンらが区別する、AI が科学的理解を支えうる役割、たとえばパターンを明らかにすることやインスピレーションの源泉となることは、第一の主張にあたる。これに対して、同じ論文がさらに強い可能性として提起する、理解の行為者としての AI は、第二の主張にあたる (Krenn et al. 2022)。この枠組みは、現在のシステムがこの意味で理解の行為者であることを確立するものではない。

クレンらがそれ以前に発表した、量子光学の実験を設計する計算機プログラム Theseus (テセウス) についての研究は、第一の主張の具体的な事例を提供している。この研究は、そのプログラムのグラフに基づく表象と自動化された探索が量子光学の未解決問題の解を生み出し、研究者たちがそこから新しい概念を引き出したと報告している (Krenn et al. 2021)。著者たちは、このシステムが概念的理解に寄与していると述べている。この事例が直接に示すのは、人工システムの手続きと解釈可能な表象が人間の理解を媒介しうるということである。

第二の主張、すなわちAI自体が理解するという主張は、本節の大部分を占める。以下ではまず、説ごとに何が試されるかを見る。次に、機械が理解しうるかを直接論じた立場を、デ・レフトとその共著者からウィルケンフェルドまで取り上げ、最後に、理解に信念が要るかを問う。この主張をどう評価するかは、選ばれた理論に依存する。説明に基づく説は、システムが説明を表象し適切に把握しているかどうかを問うかもしれない。なぜそうなのかの理解についてのヒルズのような能力説は、システムが説明についていき、説明を与え、類似した事例について結論を引き出し説明を与えることができるかどうかをテストするかもしれない。表象の操作可能性に基づく説は、反事実的推論、介入、転移、誤り診断を支えるようにシステムが表象を修正できるかどうかをテストするかもしれない。依存関係説は、整合的な依存モデルがシステムの推論と行為に統合されているかどうかを問うかもしれない。これらのテストは、流暢な報告生成を超えるものであってはじめて中身をもつ。理解を把握の感覚とも結びつける説（後述のポトチュニクの説など）では、これらのテストはなお別の問いを残す。機能的関係あるいは表象的關係だけを求める説では、人工システムにおける実現は、その関係を満たしつつ、人間における実現とは異なりうる。

第二の主張に直接に関わる立場がある。デ・レフトは、科学的説明を構成し評価するのに必要な技能には、規則を適切な状況でどう適用するかについての暗黙知が含まれるとする。コリンズ (Collins) に従って、これを集合的暗黙知に分類する。デ・レフトの説では、集合的暗黙知は社会的な実践に参加することによってしか身につかない。また、それを明示的な科学的記述として書き出し、その記述をもとに人間の振る舞いをまねる機械を作る、ということもできない (de Regt 2017, 27–30)。同様にデ・レフトは、カイザー (Kaiser) に従って、一部の物理学者は公開された資料だけからファインマン図

(Feynman diagrams) を使った計算の仕方を身につけたが、まったく新しい問題にそれを用いることは身につけなかったと報告している (de Regt 2017, 252–253)。この議論が退けているのは、そうした技能を明示的な規則や記述として書き下し、それによって捉えるというやり方である。そのため、明示的な記述からではなく訓練を通じて傾向性を身につけるシステムが、関連する実践に加わるようになりうるかどうかには、この議論は答えていない。また、この議論は実践に加わるのに何が要るにかかわるので、人工のエージェントのあいだの実践そのものが、そうした集合的暗黙知を保持する社会的な実践でありうるかどうかにも触れていない。デ・レフトが報告するファインマン図の事例も、公開された文献だけでは、実践者にできることの一部しか伝わらなかった、ということを示唆するにとどまる。

デ・レフトが共著者に加わった後年の仕事は、デ・レフト自身の説を二通りに読んでいる。第2節で触れたボーゲとの共著の章は、文脈的理論を、とくに人間の科学的理解に当てはまるものとして示す。理由は二つ挙げられている。一つは、この理論が、人間の科学者が実際にどのように理解に至るかを分析して作られたことである。もう一つは、その中心概念である理解可能性が、特定の技能をもつ科学者にとって、それも特定の歴史的・学問分野的・社会的な文脈のなかで成り立つものであることである (Boge and de Regt 2025, 418)。著者たちはそのうえで、深層ニューラルネットワークがこの意味で理解に至りうるかを検討する。一見すると、答えは否定的である。文脈的理論では、理解は理解可能な理論に基づく説明から得られるが、ニューラルネットワークの予測は科学的な理論に基づいておらず、現象の理解可能なモデルも与えないからである (Boge and de Regt 2025, 419)。ただし、ここでいう理論は厳しいものではない。デ・レフトは理論を、ギアリーに従って、概念を構成要素とする原理の集まりと捉えている。ニューラルネットワークがデータから特徴を抽象することを概念の獲得とみなせば、ニューラルネットワークも自らの概念から理論を作りうることになる。問題は、その概念が人間の概念と大きく異なりうることである。そうであれば、ニューラルネットワークが自らの概念に基づく理論を通じて理解を獲得したとしても、その理解は人間には近づけなかもしれない。人間の科学者は、自分がかめない概念を使って説明を組み立てることができないからである (Boge and de Regt 2025, 420, 429)。著者たちはこの帰結を、人間の科学的理解と人工の科学的理解のあいだの分断と呼び、科学的理解全体の進歩にとっての障害とみる。そして、その分断を越える道を、AI が自らの理解を人間の専門家に伝えられるかどうかを試し、どこまで伝えられたかを測る提案に求める (Boge and de Regt 2025, 427–428)。第1章第2節の言葉でいえば、これは、理解が人工システムのうちに実現しても、人間の *S* の認識的状況には入らないかもしれない場合である。第1章第6節は、テッド・チャンの短編を、これと似た形の事例として思考実験のように読んだ。人間の *S* の外の担い手において実現した結果が、どの成員もまだたどれない形で記録されている事例である。そこで見た読み方のうち二つでは、解説は、その結果をもつ者にとっては新しくなくても、*S* にとっては新しくありうる。その読み方に立てば、ボーゲとデ・レフトのいう分断を越えることも、それ自体、人間の *S* にとって新しい達成でありうる。

もう一つの読みは、この説を人間にも機械にも同じように当てはまるものとして扱う。バーマン (Barman)、キャロン (Caron)、クラーセン (Claassen)、デ・レフトは、科学的理解を能力とみなし、内部の仕組みや理解しているという感じではなく、主体に何ができるかによって評価する。そして、理解は人間とは異なる仕組みによっても実現しようと論じる。バーマンらはデ・レフトの基準から出発し、その焦点を現象から主体へと移して、程度の問題にする。主体がある現象を理解しているのは、その現象について十分に完全な表象をもち、内的に整合し経験的に十全な説明を生み出せ、その現象についての正しい反事実的推論を幅広く引き出せる程度に依拠してである。そのうえでバーマンらは、これらの能力を、「何か」「なぜか」「もし〜ならどうなるか」という問いで試すことを提案する (Barman et al. 2024, 2–6)。メレイン・ムーディ (Merijn Moody) とデ・レフトのプレプリントは、文脈的理論の理解可能性の基準、すなわち正確な計算をせずに理論の定性的に特徴的な帰結を認識できることを、科学的理解の機能的な定義として採る。二人が避けようとするのは、ジョン・サール (John Searle) の「中国語の部屋」の議論に倣う反論である。中国語を知らない人が、中国語の記号を操作する手順書に従って答えを部屋の外へ返す。外から見ると中の人は中国語を話せるように見えるが、その人は中国語を少しも理解していない。サールはここから、プログラムを実行することは、出力がどれほど適切でも、それだけでは理解を生まないと結論する (Cole 2024, §§1, 3)。これに対して、ムーディとデ・レフトの定義は理解を能力とみなし、その能力のほかには何も求めない。そのため、二人の説では、この定義はサール流の反論をはじめから回避する。そして、言語モデルが分布外の幅広い科学的な問いに正確に答えるという間接的な証拠に基づいて、現在のモデルに、脆いながらも自明でない程度の科学的理解を認める。ただし、広範な評価はまだ行われていないとも述べている (Moody and de Regt 2026, 9, 23)。

プラグマティックな理解 (pragmatic understanding)、すなわち何かがなぜそうであるかの理解ではなく、あるシステムを用いて何かをするやり方の理解と呼ばれてきたものについての説も、同じ点で分かれる。ウェスターブラッド (Westerblad) とスチュアート (Stuart) の紹介によれば、レンハルト (Lenhard) は、プラグマティックな理解を、現象を予測し制御する能力のうちにあるとみなす。レンハルトが取り上げるナノ科学の事例では、シミュレーションが理論の代わりを務め、設計規則によって、科学者は、その動態が理論的には理解されていない現象について、人工物をつくり、現象を制御できる (Westerblad 2026, 13; Stuart 2025, 4)。スチュアートは、こうした能力の見方のもとではAIシステムも理解をもちうると指摘したうえで (12)、この見方を退ける。スチュアートの説では、ある行為者があるシステムについてプラグマティックな理解をもつのは、そのシステムやその部分を、目標を達成するよう頑健かつ首尾よく操作する技能をもち、しかもその技能を身につけたことについて責任を負う、つまり称賛に値する場合であり、その場合に限られる。技能は、たんなる能力とは異なり、課題に特有であり、学ばれるものであり、ある程度の認知的な制御のもとで発揮される (6–8)。そのためスチュアートは、行為者に何ができるかを測るバーマンらのベンチマークが、理解に必要なものをとらえているかを疑う (12–13)。そして、現在と近い将来のAIシステムはプラグマティックな理解をもたないと結論する。ただし、その前提、すなわちそうしたシステムは責任を負えないという前提を論証せず、AIの哲学における多数派の見解と自ら呼ぶものに従って採っている。さらに、説明的理解や対象的理解はしばしば概念の習熟のような技能を要するので、科学の多くの場合には、システムはそれらももたないと結論する (13–14)。それでもスチュアートは、成員の一部が関係する技能をもち、それに責任を負っている研究室には、AlphaFoldを装置の一つとして用いても、プラグマティックな理解を認める。また、責任の条件を外す道筋や、人工システムがそれを満たす道筋を、採用はせずに素描している (14 n. 7, 16)。したがってスチュアートの説では、*A* がこの意味で理解しているかどうかは、どのような一連のテストでもなく、*A* がその技能について称賛に値する行為者であるかどうかにかかっている。これは、第7章第3.3節が真なる出力、信頼できる過程、有能さに帰属可能な成功と区別して保つ、功績の区分の四つ目にあたる。ウェスターブラッドは、ハンフリーズに応答して、装置は探究手続き、すなわち目的に達するためにそれを用いる、公に利用できるやり方を通じて、科学者にとって機能的に理解可能になるとする。そして、新しい装置は、すでに受け入れられている手続きに照らして較正されることでその手続きを広げる。そのためこの説では、科学が人間の能力を拡張する手段は、人間の限界を越えて進むわけではない (Westerblad 2026, 15–20)。この議論で点検し、検証し、比較する行為者は人間の科学者である。またウェスターブラッドは、自らの結論が、科学はもはや人々によるものでも人々のためのものでもないという見方にかかわることを認めつつ、その点を明示的には論じていないと断っている。

る (20)。受け入れられた手続きに照らして新しい手続きを校正する人工の A に当てはめれば、ウェスターブラッドの説のもとでは、その拡張が人間の S の把握のうちにとどまるかどうかは、誰が校正を行ったかではなく、その手続きが S の成員に利用できるかどうかにかかることになる。

ウィルケンフェルドは、人工知能を、第3.3節で紹介した自らの表象の操作可能性説をおそらく最もよく活用できる分野として提示している。この説では、主体がある対象を理解するのは、その対象の表象をもち、それを自ら少し修正して、その対象についての有効な推論や操作を可能にする新しい表象を作れるときである。ウィルケンフェルドの例では、論理学の規則を誤って覚えていたことに気づいて直せることや、理解している証明を少し異なる体系に合わせて作り変えられることがこれにあたる。ただ暗記しただけの証明は、このようには作り変えられない (Wilkenfeld 2013, 9–10)。この説には、理解の担い手が人工システムである場合にかかわる特徴が二つある。第一に、表象に求めるものが少ない。求めるのは、内容をもち、変換を受けうる計算的構造であることだけである。ウィルケンフェルドはこれを人間の心についての主張として述べ、古典的な計算主義とも、思考の言語仮説とも、ニューラルネットワークのいくつかの消去主義的でない解釈、すなわちニューラルネットワークの状態が内容をもつとみる解釈とも両立するとしている (Wilkenfeld 2013, 8–9)。そのため、この説は、表象が記号的でないという理由でシステムを理解の担い手から除くことはない。第二に、どの操作が理解に数えられるかを決める文脈的な基準は、理解を帰属させる側の特徴づけに属する。理解する者自身がその基準を表象している必要はなく、表象を修正し、推論できさえすればよい (Wilkenfeld 2013, 19)。したがって、問題の理解する者が人工システムである場合、 S が人間であれ人工のものであれ、この説のもとで問われるのは、そのシステムが対象を表す状態をもつかどうか、そしてその状態を、外から内部状態を書き換える誰かではなく、システム自身がこのように修正できるかどうかである。ウィルケンフェルド自身はこの問いを扱っていない。人工知能に目を向けるのは、人間の認知者のようにふるまうコンピュータを作るには、人間が理解するときに何をしているかを明確にしておく必要がある、という理由からである (Wilkenfeld 2013, 18–19)。

関連する問いは、人工システムが信念をもたずに理解をもちうるかどうかである。第3.2節で触れた、理解は知識のような仕方では正当化も信念も要求しないというデルセンの主張は、この問いに関わる。ただし、どう関わるかは、信念で何を意味するかによる。デルセンの中心的な事例では、弦理論が真である世界の物理学者が、方法論上の留保からその理論を信じていないが、それを受容し、説明に用いている。デルセンは、この物理学者はその理論が記述するものを理解していると論じる (Dellsén 2017, 2, 13–15)。この事例が依拠する対比は、デルセンがL・J・コーエン (L. J. Cohen) から採るものである。 p を信じるとは p が真であると感じる傾向をもつことであり、 p を受容するとは p を所与として扱う方針を採用することである (Dellsén 2017, 13–14)。この意味の信念は、第2章第2.1節で論じた信念より狭い。そこで述べた機能説・傾向説のもとでは、 p を保存したり発話したりすることよりも、 p が推論、予測、行為、修正に安定して統合されていることのほうが、信念のより強い証拠である。そのため、コーエンが受容に分類するふるまいそのものが、この広い意味での信念の証拠に数えられる。コーエンの区別に大きく依拠するエルギンは、それを、信念という常識的な概念の解明として読む。その概念には二つの成分がある。一つは p がそうだと感じることで、あるいは感じる傾向をもつことであり、これはおおむね意のままにならない。コーエンが「信念」の語をあてるのはこの成分である。もう一つが受容である (Elgin 2025, 35–36)。

ポトチュニクも、受容を信念から区別する。何かを受容するとは特定の理由からそれを前提することだ、という定義を用いて、理想化を含めて理解に寄与するものは、認知的に受容されること、すなわち認知的な理由から前提されることを要するが、信じられる必要はないとする。理想化は信じられも主張されもせず、事実と反して仮定される (Potochnik 2015, 8, 13, 16–17)。ただしポトチュニクの説は、理解を心理的な条件から解放するわけではない。この説では、理解は認知的な達成であるとともに認知状態でもある (Potochnik 2015, 5)。ラジャ (Raja) とサンシェス・デ・オリヴェイラ (Sanches de Oliveira) が報告するポトチュニクの説では、理解には、因果的パターンを実際に把握することと、把握したと感ずることの両方が要る (Raja and Sanches de Oliveira 2024, §§3.1–3.2, 10–13)。ファン・デル・メルヴェ (van der Merwe) も、真正の理解は、その感覚を引き起こすのにふさわしい種類の情報から生じるという彼女の定式化を報告している (van der Merwe 2023, 330)。

これらの見解は、理解は信念ではなく受容で足りるという点では一致する。しかし受容の定義が異なるので、理解する者が人工システムである場合に問うべきことも、 S が人間であれ人工のものであれ、見解ごとに異なる。コーエンの区別を用いるデルセンによれば、理解に要るのは、受容と、正しく説明できるほどの主題の習熟である。他方、知識にはなお信念が要るとされ、弦理論の事例の物理学者は、理解しているが知ってはいない (Dellsén 2017, 14, 17–18)。信念が p を真と感じる傾向であるなら、感じることを帰属できないシステムでも受容はしうるので、そのシステムにとっては知識より理解のほうが近いことになる。残る問いは、説明において p を前提として使うことが、方針を採ることにあたるかどうかである。デルセンは、方針を採ることを、人が決めて行うこととして述べている。エルギンは受容を、目的が認知的であるときに、 p を推論や行為に用いる意思と能力をもつことと定義し、規則、基準、方法も受容されうるとする。そして、おおむね意のままにならないコーエンの意味での信念と違って受容は自発的であるとし、その意味での信念は傍観者的であるのに対して、受容は行為者的だとする (Elgin 2025, 11, 35–36)。そのためエルギンの説では、人工システムについての問いは、感じることから、そのシステムが認知的な目的をもち、そこでいう自発的な意味で受容しているかどうかへと移る。上で報告されたポトチュニクの説では、信念を外しても、人工システムについての事情は楽にならない。理解の感じがなお必要だからである (Raja and Sanches de Oliveira 2024, §3.1, 11)。またポトチュニクは、何が理解を促すかは人間の認知の特徴に依存するとみている (Potochnik 2015, 11, 15–16)。最後に、こうしたことがどれほど効くかは、採用する信念の説、すなわち第2章第2.1節で提起した問いによる。コーエンのような狭い説のもとでは、受容は信念なしに実現しうる。そこで論じた機能説・傾向説のもとでは、デルセンの事例のように説明に用いられる受容は、それ自体が信念の証拠に数えられる。そうであれば、信念なしの理解は、人工システムにとって別の道にはならない。これらの説のもとでも信念からはっきり区別されるのは、偽と知られ、ひとつの研究目的のなかでだけ前提される措置である。それは、これらの説が信念のより強い証拠とみなす、推論、行為、修正への安定した統合を示す必要がない。

第三の主張、すなわち人間とAIの集合体あるいは人工の共同体が理解するという主張には、集団の理解が、人間であれ人工のものであれ、その成員が理解していることとどう関係するかについての説が必要である (第2章第6節の集団の知識についての説と比較)。分散したタスクの成功だけでは、集団が理解していることは示されない。異なるエージェントが協働して正しい答えを産出しながら、どの構成要素も、どの組織された全体も、それを説明する関係を表象することも、タスクを超えてそれを用いることも、異議に応答することもできない、ということがありうる。理解を組織された全体としての集団に帰属させる説では、集合的理解の主張は、部分的な表象、能力、コミュニケーション、訂正が、どのようにして帰属される達成へと合成されるのかを特定する必要がある。集団の理解を成員の理解に基づいて考える説では、その主張は、成員が何を理解しているかにかかる。

三つの主張は、いずれも誰が理解するかにかかわっていた。本節の残りでは、冒頭で別の問いとした、ある担い手のうちに実現した理解が S の認識的状況に入ったかどうかに移る。その手がかりになるのが、個人の理解と、科学に帰される理解との区別である。ボーゲ、シュスター（Schuster）、シュトル（Stoll）は、**科学者の理解**（scientists' understanding）と**科学的理解**（scientific understanding）とを区別している。前者は、個々の科学者が、ある現象の説明を把握すること、その現象を説明する整合的な情報の集まりを抱くこと、あるいはその依存関係のモデルを把握することによってもつ理解であり、暗黙的あるいは言語化不可能でありうる。後者は、三人の説では、その現象について、十分に広い科学的聴衆に伝達可能な科学的説明があるときにのみ成り立つ。その聴衆は少数の専門家の集団でもよく、説明はまだ伝達されていなくてもよい（Boge, Schuster, and Stoll 2026, 2, 5–7）。三人の事例の一つでは、ファインマン（Feynman）は自分の図がなぜうまくいくのかを理解していたが、他の物理学者がたどれる説明が与えられたのは、その後のダイソン（Dyson）の証明によってであった（Boge, Schuster, and Stoll 2026, 9–10）。著者たちは、両者の違いは一方が説明的で他方が対象的だという点にあるのではないと強調する。個々の科学者は説明にもとづいて理解することもあるが、著者たちの説では、科学全体は説明を媒介として理解する。そのうえで、科学者の理解は非常に重要な認識的達成でありうるが、まだ科学的理解ではない、と述べる（Boge, Schuster, and Stoll 2026, 4, 6）。この区別は、個人と共同体とで理解がどう実現されるかではなく、何を達成に数えるかにかかわると理解できる。それは、個人に帰されるものと科学に帰されるものととの区別である。その点でこの区別は、第1章第2節と第6節で区別し、知識について第2章第2.1節と第6節で取り上げた二つの区別に、理解の側で対応する。一つは、研究を行う行為者と、達成がそれに相対的に新しく、そのために求められる社会との区別である。もう一つは、成功を評価するところで、ある担い手のうちに実現した達成と、その社会の認識的状況に入った達成との区別である。どちらの組でも、二つは一致しうるが、一方から他方が決まるわけではない。ただし対応は部分的である。三人の説では、科学に帰される達成に必要なのは伝達可能な説明が存在することであり、共同体の成員がそれを実際に受け取っていることは必要とされない。したがって、科学に帰されるものは、成員がその達成を保有していることではなく、その達成が社会にとって利用可能であることに対応する。知識についてのこの違いを、第2章第7節は、知っていることと知る位置にあることとの違いとして論じている。対応が部分的である点はもう一つある。三人の事例の個人は結果を生み、それを理解する者でもあるが、第1章の区別では、結果を生む行為者と、それを理解する担い手とは異なりうる。この提案は、科学的理解は説明を必要とするかという争われている問いで一方の側に立つものであり、科学的理解の確定した意味ではない。

この区別が切り分ける問いは、いずれも決着していない。ある立場は、研究者が明確化に先立って暗黙的に理解することを許容しうる。説明に基づく説もこれを認める。説明を把握することは、それを述べる能力を含む必要がないからである（Strevens 2013）。依存関係説を擁護するグリムは、言語化できることは理解の必要条件ではないと論じ、幼い子どもや一部の動物が、言語化できない因果的な依存関係を把握しているという証拠を挙げる（Grimm 2017）。別の立場は、科学的理解を、コミュニケーションと批判を支える表象のために取っておくことができる。ボーゲ、シュスター、シュトルはこの両方を探り、科学的理解を伝達可能な説明に限る。三人は、エルギンの説を科学者の理解にかかわるもの、デ・レフトの説を科学的理解にかかわるものと読み、両者を対立するものではなく補い合うものとみなす（Boge, Schuster, and Stoll 2026, 3, 14）。バード（Bird）の知識の集団機能的な説（第2章第6節）を理解に広げる説があれば、それは、どの個人も全体を保有していない場合でも、組織されたシステムに理解を帰属させるだろう。これらの立場は、理解をどこに位置づけるかに応じて、人工研究者に対して異なる要求をする。最初の立場のように、理解が研究者自身の暗黙的な把握のうちにありうるとするなら、人工研究者自身が理解するという主張には、その研究者が依存関係や説明を自らつかんでいることが求められる。ただし、それを言葉にできる必要はない。ボーゲ、シュスター、シュトルの説のように、科学的理解には伝達可能な説明が要するのなら、研究者は、関係する聴衆に伝達できる説明を生み出さなければならない。その聴衆が人間であれば、説明は人間に向けたものでなければならない。バードの説を広げた説のように、組織されたシステムに理解を帰属させようとするなら、求められるのは、研究者の寄与が、組織全体として実現される集合的な認識的能力に統合されていることである。

二つ目の要求が示すように、誰が理解を担うかと、誰のために理解が求められるかは別のことである。理解の担い手と、その理解がそのために求められる社会とは、分析的に区別できる。前者は理解する者であり、後者は研究についての主張にかかわる二つの当事者の一方である（第1章第2節）。ある担い手のうちに実現した理解がその社会の認識的状況に入ったかどうかは、さらに別の問いであり、同節で区別した第三の判断の、理解の側での対応物である。そして上で見た立場は、この問いに異なる仕方で答える。集団の理解を成員の理解に基づけて考える説では、人間であれ人工のものであれ、関連する成員が理解したときに、理解は社会の認識的状況に入ったことになる。上で述べたようにバードの説を広げる説では、社会の組織が全体としてその理解を実現したときである。ボーゲ、シュスター、シュトルのいう科学的理解に必要なのは、伝達可能な説明が存在することだけである。また、デルセンの改めて述べられたノエティックな説は、進歩を、社会の関連する成員が理解することを可能にする公的な情報に位置づける（Dellsén 2021）。どちらも、成員が保有しているものではなく、社会にとって利用可能なものにかかわる。そのため、そうした利用可能性がすでに認識的状況に入ったことにあたるのか、それとも社会を理解する位置に置くだけなのかは、知識についてのバードとラッキーの論争（第2章第7節）に対応する問いになる。

これらの区別を人工システムに当てはめると、少なくとも次の可能性が分かれる。人工システムは、それ自体は理解することなしに、人間の理解を可能にするかもしれない。それは、人間も機械も説明的理解を保有しないまま、人間に根拠づけられた知識を提供するかもしれない。十分に非人間中心的な説のもとでは、それはむしろ、人間にとっては理解不可能なままでありながら、自らの推論、介入、誤り訂正を支える形式で依存関係を組織するかもしれない。そうした組織化は、そのシステム自身の推論にだけ役立つこともあれば、人間には伝達できなくても、 S の他の人工の成員には伝達できることもありうる。人間とAIの組織もまた、どの一成員も保有していない統合された能力を実現するかもしれない。

本章で検討した証拠は、人工システムが、人間には理解不可能なまま自らの推論を支える形式で依存関係を組織するという可能性を、既存システムの記述として確立するものではない。それでも、この可能性は標的の重要な選択を露わにする。関係する聴衆にとっての理解可能性は、意図された達成が S の認識的状況に入ることが目指される理解、たとえば公共的な科学的理解である場合には必要でありうるが、人工システムがもつ理解についてのあらゆる可能な説に組み込まれているわけではない。そして、その理解可能性が人間に向けたものである必要があるのは、その聴衆に人間が含まれる場合に限られる。逆に、ある表象が人工システムの内部で有用であるという証拠は、それが S にとって理解可能であること、根拠づけられていること、あるいは価値をもつことを確立しない。したがって、人工システムによる研究が広がれば、知識と理解は、異なる認識的社会にとって異なる軌道に沿って前進するかもしれない。これはインターフェースの問題以上のものである。それは、人工研究システムが誰の認識的能力を拡張するのにかに関わる。この食い違いは、社会と社会のあいだにだけ生じるとは限らない。ヴェンカテシュ（Venkatesh）は純粋数学について書き、概念は人間の記憶と時間の限界に合わせて思考の経

済を実現するものであり、機械はその働きをめぐって概念と競合すると考える。そのうえでヴェンカテシュは、機械的な推論が数学の言語と概念体系を徹底して変えることで、現在と近い将来の数学者が互いに理解し合うのに大きな努力を要するようになる可能性を示唆し、数学の過去の書き換えは、そうした相互の理解不可能性が現実の可能性であることを示していると考えられる (Venkatesh 2025, 94–99)。その説では、この食い違いは、人間の S の内部で、異なる時点の成員のあいだに、機械が研究に加わることを通じて生じうる。またヴェンカテシュは、デイヴィス (Davis) とハーシュ (Hersh) にならって、伝達可能性は数学の「定義そのものの一部」であるとするので、その見方では、そうした書き換えは、数学がどのように行われるかだけでなく、数学が何であるかをも変えることになる (84, 112)。

人工システムが人間には理解不可能なままで依存関係を組織するという可能性には、二つの主張が含まれている。人工システムによる依存関係の組織化が因果関係を捉えているという、世界の側についての主張と、それがその関係を説明し、あるいはその理解をもたらすという、担い手と使い手についての主張である。ウッドワードの因果的説明の理論は、この二つを切り分ける。この理論では、因果関係は世界の側の特徴であり、ウッドワードによる介入の特徴づけは人間の行為者性に言及しない。正しい因果的特徴をもつ過程は、人間のいない世界における自然の過程も含めて、介入に当たる (Woodward 2023, §8)。そのためこの理論のもとでは、人工システムによる依存関係の組織化は、人間にとって理解不可能なままで、因果関係を捉えうる。バッシ (Bassi) が報告するように、これに対してウッドワードは、説明を人間が行う活動として扱い、他のいくつかの動物も行いうるとしている (Bassi 2013, 123–124)。またウッドワードは、説明にその意義を与える情報が、それを使用者にとって認識的にアクセス可能であることを求める (Woodward 2000, 248 n. 11)。人工システムがその使い手に含まれるかどうかを、ウッドワードの説は決着させていない。そのような組織化がさらに説明でもあるかどうかは、この説では、その説明が誰に向けたものか、そしてその使い手が、人間であれ人工のものであれ、関連する情報にアクセスできるかどうかに依存する。

ポトチュニクの説明論は、聞き手を、何が説明するかの一部にする。その説では、説明は、それを求める者に理解を生み出すものでなければならない。そして、ある現象をもたらしている多くの依存関係のどれがそれを説明するかは、その聞き手が何を理解しようとしているかに依存する (Potochnik 2016, 9–13, 17)。ポトチュニクはこの説を人間の科学について展開しており、説明は人間に理解できるものでなければならないとし、広いパターンの寄与を人間の理解の特徴に結びつけている (Potochnik 2016, 17–18; Potochnik 2015, 11, 25)。この説で何が説明的かを形づくる聞き手の特徴として、ポトチュニクは、聞き手の認識的な状況や関心と、人間一般については認知的な特徴を挙げる (Potochnik 2016, 4–5, 17)。この説を人工システムによる依存関係の組織化に広げると、それが何かを説明するのか、そして誰にとって説明するのかは、部分的には、人間であれ人工のものであれ、その聞き手の認知的な必要についての問いになる。そして、ある聞き手にとって説明になるものが他の聞き手にとっても説明になると想定する理由は、その説からは出てこない。

まとめると、三つの主張は異なる証拠を要する。他者が理解するのを何が可能にしたか、選んだ説のもとでシステム自身がその表象で何をなしうるか、集団の部分的な表象や能力がどう合成されるか、である。そして、いずれの主張についても、理解がどこに実現したかと、それが S の認識的状况に入ったかとは、別に問わなければならない。

本章から見てきたのは、理論ごとに求められる証拠が異なること、そして、ある担い手に認められた性質は論証なしには別の担い手に移らないことである。人工システムが文字どおりに信じ、内容を表象し、あるいは理解しうるかどうかに関連する問いは、第IV部が取り上げる。規範形成と認識的自己統治は、継承された基準についての自律性の分析 (第7章第4.3節) で扱う。第8章は、知識の担い手や社会が人工のものである場合を扱い (第1節)、また、人工の知る者の検査が証拠として働き、その状態が世界についてのものとなり、そのコミットメントがその世界に応答可能であり続けることを何が可能にするのかを扱う (第2節)。

7. 理解は何のゆえに重んじられ、時間を通じて何を可能にしまうか

第4節は、試みが理解に向けられていることと、研究やその個々の操作が理解をもたらすという要件とを区別した。特徴づけは、どの達成についても、そうした要件を課さない。理解の価値と、持続的な探究のなかでの理解の位置は、実質的な説に委ねられる。ここで問う価値は、認識的な状態としての理解の価値、すなわち、理解が何のゆえに重んじられるか (それ自体としてか、継続する探究のなかで可能にしまうもののゆえにか) である。第9章第3節は、こうした状態を価値の対象の一つに数える。同節が主に扱う問い、すなわち S の問い、実践、結果の分布のうちどれが重要であり、誰にとって重要かは、第1章第1節の価値とガバナンスの水準に属し、本節では扱わない。一つの答えは、理解自体の価値や探究の目的としての位置に関わる (第7.1節)。もう一つは、継続する探究のなかで、理解がさらなる能力を構成し、その証拠となり、可能にし、促進する仕方に関わる (第7.2～7.4節)。第4節の強化した予測の事例は、理解が何を可能にするかという後者の問いを鋭い形で提起する。研究システムが知識をもたらしても、その知識を誰が保有するにせよ (誰も保有しないこともありうる)、それをシステム自身も、システムがそのために探究する者も理解しないとき、単一の妥当性が確認された結果を超えて、何が失われるのか、という問いである。小節はこれに沿って並ぶ。第7.1節は、それ自体のために、あるいは探究の目的として重んじられる理解を扱う。第7.2～7.4節は、探究がひとつの妥当性確認済みの結果を超えて続くときに理解が可能にしまうものを扱い、何か把握され、あるいは利用可能になるかによって三つに分ける。対象が依存するものが違えば対象がどう違うかの把握 (第7.2節)、ある領域がどう結びついているかの把握 (第7.3節)、そして理解が他者に利用可能になること (第7.4節) である。この分け方は、理解をめぐる文献の流れに沿って本書が立てたものであり、文献自身が与える分類ではない。また、これで尽くされると主張するものでもない。

7.1 それ自体のために、あるいは目的として重んじられる理解

何人かの認識論者は、理解の価値は理解が可能にするものによって尽くされない、と考える。クヴァンヴィグは、知識の価値は、真なる信念や主観的な正当化といった、その構成要素の一部の価値によって尽くされるのに対し、理解はそうした部分を超える価値をもつと論じる。この説では、理解には、大きな情報の集まりのなかで説明的な関係やその他の整合をなす関係を把握することが必要である。そうした関係をつかむことは、真なる信念を、あるいは正当化された真なる信念をただ付け加えるのとは違う仕方、主題についての思考を組織し、体系化する。組織された思考は、ある情報から、行為の基盤と

なりうる関連した情報へと推論することを可能にする。また、主題を全体としてつかんだという感覚を伴って、探究に満足のいく区切りをもたらす。部分を超える価値とは、この組織化の価値である (Kvanvig 2003, 185, 188, 192, 202, 205)。プリチャードは、何かがなぜそうなのかの理解は、普遍的にではないにせよ典型的には、要求の高い意味での認知的な達成、つまり重大な障害を乗り越えることか相当の認知的能力を発揮することを求める達成を伴い、この種の達成は最終的な価値、すなわち道具的でない価値をもつ、と論じる (Pritchard 2016, 9–10)。同じ論文でプリチャードは、この価値の根拠を、そうした達成において自らの行為者性が世界に働きかけることと、理解が示す知的自律とに求める。ここでの知的自律とは、専門家の助けを借りて得た場合でも自分のものである、事柄の把握である (Pritchard 2016, 9–12)。

リグズ (Riggs) は、認識的な観点から本当に価値があるのは、理解のような、より高次の認知的な達成だとする論者の一人である。グリムがまとめるところでは、われわれには、古い電話帳に載っているある人物の名前のような取るに足らない情報を得たいという欲求が生まれつきあるようには思われないうが、どのような理解も、追い求める価値があるように見える。そしてリグズは、取るに足らない、ありふれたものとみなされうる主題についての理解であっても、どのような理解も人生の認識的価値に寄与すると書いている (Grimm 2021, §3.1)。ゴードンが紹介するところでは、リグズは、ある主題を理解するには、その部分どうしがどう組み合わせり、それぞれが全体のなかでどんな役割を果たしているかを深く感得することが要するとする。またリグズは、理解から、自らが理解可能性 (intelligibility) と呼ぶものを区別する。リグズのいう理解可能性も、秩序やパターンやつながりの把握を含むが、真理との実質的なつながりは要求しないように見える (Gordon n.d., §§2b, 3)。

ザグゼブスキー (Zagzebski) は、ほかの論者とともに、知識だけに焦点を当てる認識論は、理解や知恵のようなほかの認識的な善を顧みない一面的なものと論じた (Grimm 2021, §1.2)。グリムは、この軽視の背後にある歴史を紹介して、次のように述べる。古代のエピステマー (episteme) は、プラトンやアリストテレスにとって、物事がどうつながっているかの体系的な把握や、物事の原因の把握を含んでいたもので、知識よりも理解と訳すほうがよいと論じられてきた。そして、後の認識論が孤立した命題的知識に焦点を当てるようになったことは、ヘレニズム哲学における懐疑論の台頭への応答として説明されてきた。グリムはこの説明をバーニェト (Burnyeat) とザグゼブスキーに見いだす。あるいは、パスナウ (Pasnau) が示唆するように、近世ヨーロッパの宗教戦争への応答としても説明されてきた (Grimm 2021, §1.1)。ゴードンが紹介するザグゼブスキーの見方では、すべての理解が真なる信念を要するわけではない。地球以外のあらゆるものの引力を無視して、真空中の物体は一定の割合で落下するとする信念は、たいいていの目的にとって、はるかに複雑な真理よりも多くの理解をもたらす。真理を把握することには、認知的にあまりに大きな負担がかかりうるからである (Gordon n.d., §2b)。

ヒルズは、美的な理解について論じるなかで、ある作品がなぜ価値をもつのか、あるいはもたないのかの理由に気づいていることは、のちによりよい判断を下す助けになりやすいと考える。そして、そうした理解は美的な徳の不可欠な構成要素であり、美的な判断を下す理想的な仕方は自分自身の理解を通じてであるとする。その論拠の一部は、価値を映すという考えである。真なる信念は、その内容によって世界を映す。判断を真にする理由の把握に基づく判断は、さらに形式においても世界を映す。ある信念が他の信念に依存する仕方が、世界のなかの依存と対応するからである。ある小説が稚拙で、生氣がなく、文章が下手だと認めることでその小説を悪いと判断する人の場合、その小説が悪いという信念は、それが稚拙だという信念に依存しており、その依存は、小説の価値のなさがその稚拙さに依存していることと対応する。そしてその人は、小説が別の性質をもっていたなら、別の判断を下していただろう (Hills 2022, 21, 28–29)。ヒルズは、この説明が道徳的な理解にも一般化すると示唆する (Hills 2022, 22)。

ただし、理解の価値の説明がどれも、理解が可能にするものを越えた価値に訴えるわけではない。グリムは、理解の特別な価値の多くを、世界がどう展開するかを予測し、ことによると制御できる立場を理解が与えることに求める。ただし、それが理解を支持する直観を尽くすとは限らないことを認めている (Grimm 2017)。グリムのそれ以前の概観は、二種類の価値をとともに視野に入れており、理解をそれ自体のために価値あるものとするのが、理解が知識の一種であることを否定することに依存しなくてよいことを示している。グリムは、なぜ理解を原因の知識とみる論者の一人に自らを数えたうえで、なぜ理解は知識より価値があるのかという問いを、なぜ理解は命題的知識よりよいのかという問いに置き直す。そこでは知識が類であり、その種は対象によって分かれる。そのうえでグリムは、客観的な理解は、それ自体のためにも、世界を制御できる見込みのためにも望まれるものであろうとし、また、実在をより深く映すことが認識的によりよいのであれば、理解のどの項目も、対応する命題的知識の項目より価値がある、と論じる (Grimm 2012, 4, 11, 15)。

関連する立場は、理解を目的とみなす。プリチャードは、探究の目標は知識ではなく理解かもしれない、と示唆したことがある。その説では、自宅が焼けたのは配線の欠陥のためだと信頼できる情報源から知っても、配線の欠陥がどうしてそんなことを起こしうのか見当もつかない人は、まだ探究の目標に達していない (Pritchard 2008, §§10–12)。のちの論文でプリチャードはこの主張を限定する。何かがなぜそうなのかの説明を求める探究は、理解に含まれる種類の知識によってのみ正当に閉じられる。これに対して、たとえば動物園のある動物が本物のシマウマかどうかを問う探究は、信頼できる情報提供者から得た知識によってたいいてい閉じられる。プリチャードはこの見方を、探究の複合的知識説 (complex knowledge account) と呼ぶ。よく遂行された探究を正当に閉じるのはつねに知識であり、ただし探究によってはそれが特定の種類の知識である、という説である (Pritchard 2016, 6–7)。特定の問いへの探究の目標として知識を擁護するケルプは、一般的な現象への探究の目的を理解とみなし、そうした探究はその現象を理解したときに、かつそのときに限って成功するとする。この説では、そうした探究は、現象についての真理をばらばらに知ること、たとえばそれぞれを別々の情報提供者の言葉によって知ることでは完了しない。そこではさらに、信念どうしが、それらの真理のあいだの説明関係や支持関係についての知識を反映する仕方、互いに基礎づけられていることが求められる。ケルプの説では、最大の理解は、その現象についての最大限に包括的で、最大限によくつながった知識であり、それに満たない理解は程度の問題である (Kelp 2021b, S1583–S1584, S1588; Kelp 2015, 16, 18)。並べてみると、プリチャードののちの見方とケルプの見方は、何が探究を適切に終わらせるかを探究が何に向けられているかに依存させ、孤立した知識を超えるものが要求されうることを認める点で一致する。しかし両者は線を引く場所が異なる。プリチャードは説明を求める探究とそれ以外とのあいだに、ケルプは特定の問いへの探究と現象への探究とのあいだに線を引く。後者は、たとえばナポレオンの生涯への探究であり、ナポレオンがフランス人だったという知識のように、説明的でない知識を得ることによっても進みうる (Kelp 2021b, S1583, S1589)。こうした見方のもとでは、知識を探究の向かう先とするだけでは、理解が必要かどうかは決まらない。探究に何が必要かは、それが追う問いに依存する。これは第2節で見た、問いに相対的であるという論点と合う。

科学哲学では、エルギンが、真理は最上位の認識的目標ではないと論じる。真理を中心とする認識論は、真でないと知られているモデルや理想化に依拠する科学の認識的な地位を説明できないからである (第3.4節)。エルギンはさらに、よい科学がふつうもたらす認知的な達成は、知識ではなく理解だと論じ

る。科学の説明は、個別に支持された真理の集まりではない。その主張どうしは互いに支え合っており、全体の連言は何らかの変則事例によって偽とされやすい。しかも変則的な部分だけを切り取れば、残りはまとまりを失う。知識は真理を要するが、われわれが評価する科学の説明は真にならない。これに対して理解は、この説では、事実性をもたないまま事実に応答し、厳密には偽である説明を通じて進む。そして、隠喩、虚構、例示といった仕組みは芸術の理解だけでなく科学の理解にも働いているので、芸術とその仕組みが認知的に働くことを否定しては、科学の理解を正しく扱えない、とエルギンは論じる (Elgin 2004, 1-2; Elgin 2006, 1, 3-5; Elgin 2007, 2-5; Elgin 2002, 1)。デ・レフトは、現象の理解を科学の中心的な認知的目的の一つとみなし、理解が説明という目的と結びついているとする見方、追加の目的だとする見方、伝統的な目的に取って代わるとする見方を区別している (de Regt 2020, 921, 923, 932)。第4節で見たデルセンの説は、この系統に属する。第3.4節で見たポトチュニクの説もそうである。その説では、科学の究極の認知的な目的は真理ではなく理解であり、真理と知識は理解への道として重要になる。理想化された表象は、それが描く因果的パターンを正確に表しうる。そしてポトチュニクの説では、科学は、そうしたパターンについての知識を通じて現象の理解を生み出す (Potochnik 2020, 13, 16)。そうすると、この説では、試みが知識に向けられている場合、その目標は主として理解への道として価値をもつことになる。達成の種類を定めない作業文は、この序列を採らずに受け入れることができる。

これらの立場は互いに異なり、いずれも争われている。批判者は、重要な達成をまったく含まないように見える、容易な理解の事例を挙げる (Grimm 2021, §3.1)。証言による理解の獲得を擁護するボイド (Boyd) は、これとは対照的に、専門家にとって容易な理解は、それを容易にしている専門性そのものが苦勞して得られたものであるから、なお強い達成に数えられようと論じる (Boyd 2017, 37-38)。ここで争われている要求の高い意味での達成は、第1章の認知的達成という総称より狭い。この語自体は、重大な障害の克服も相当の認知的能力の発揮も要求しない。したがって、容易な理解がブリチャードの意味での達成かどうかは、それが *S* にとって新しい認知的達成となりうるかを決めない。デポール (DePaul) とグリムは、クヴァンヴィグ自身の推論をその積極的な提案に向け返す。クヴァンヴィグは、ゲティア事例を排除する条件はどれも継ぎはぎのものにならざるをえず、それを満たすこと自体には価値がないので、知識はその構成要素の一部を超える価値をもたないと言った。そのうえで、理解はこの種の運と両立するので、そうした条件を必要としないと考えた。これに対して二人は、話をでっち上げる記者の、得点の理由についての報道がたまたま真であり、読者がそこで述べられた説明上の結びつきを把握している事例を挙げる。二人の見方では、記者が別の話をしていても読者は同じようにたやすく把握したはずなので、読者は、得点を知っていないのと同様に、得点の理由も理解していない。そうだとすれば、理解にも、把握と、把握される結びつきとの非偶然的な関係が必要になり、クヴァンヴィグ自身の推論によって、理解は主張されていた価値を失う。ただし二人は、知識についてのクヴァンヴィグの議論の前提をもとと退けているので、この帰結を、むしろその議論を疑うさらなる理由と受け取るよう示唆している (DePaul and Grimm 2007, 510-513)。デポールとグリムは人工の理解者を論じていないが、それに当てはめれば、二人の事例が支える要件は、第3.3節の近道の問題にはつきりした形を与える。評価が除かなければならないものには、別の説明が与えられていればそれを同じようにたやすく把握したであろう把握も含まれる。そのため問われるのは、システムの説明が、説明上の結びつきそのものに従っているのか、それとも与えられた説明に従っているだけなのかである (知識について運を排除する条件を扱う第2章第4節を参照)。

これらの立場のどれを採るかによって、研究システムに何が求められるかが変わる。上で比べたブリチャードののちの見方やケルプの見方のように、説明を求める探究や現象への探究を適切に閉じるのが理解であるなら、妥当性確認済みの結果を出すシステムは、知識の実績がどうであれ、求められている理解が、その探究が目指す担い手のうちに実現するまでは、そうした探究を未完のまま残す。その担い手は、システム自身であることも、システムがそのために探究する共同体の関連する成員であることも、集合的理解の説のもとではその共同体自体であることもある (第6節)。また、探究が *S* のためのものであり、その担い手が *S* の外にある場合には、その理解が、採用する説のもとで *S* の認知的状況にも入るまで、探究は未完のまま残る (第6節、第9節)。ケルプの説では、そうした探究を閉じる理解は、文脈に応じた閾値を満たせばよく、探究の最終的な目的である最大の理解にまで達している必要はない (Kelp 2021b, S1588)。作業的特徴づけのもとでは、その活動はなお研究に数えられ、第4節の強化した予測の事例の想定のもとでは、その結果は知識に数えられうる。それによって、説明的理解を具体的に求めている探究が完了するわけではない。活動が研究であるかどうかと、その具体的な問いに答えたかどうかは異なる判断である。目的を中心に据える説は、どの達成がその種の探究を適切に閉じるかについて、さらに主張する。

これらの説が主張する価値は、誰のうちに実現するのか。この問いに照らして読むと、そのいくつかは、その価値を理解する者の側に置いている。ブリチャードはそれを、理解する者自身の認知的能力に帰せられる達成に置く (Pritchard 2008, §§10-11)。ゴードンが紹介するところでは、ザグゼブスキーは理解の認知的価値をその透明性に結びつけ、自分が理解していると理解することなしに理解することはできないとする (Gordon n.d., §5a)。こうした説のもとでは、その価値が実現するのは誰かが理解するところであって、説明が利用可能であるだけのところではない。したがってその価値は個人の理解に結びついており、第6節はこれを、伝達可能な説明という意味で科学に帰される理解から区別した。これらの説は、その根拠を、人生の認知的価値 (リッグス)、まとまった思考が探究にもたらす満足のいく区切り (クヴァンヴィグ)、自分自身の認知的能力に帰せられる達成 (ブリチャード) などを通じて、人格について述べている。どの論者も人工の理解者を論じていないが、それに当てはめれば、その理解がそうした価値をもつかどうかは、それらの根拠がその理解者に成り立つかどうかによる。そして、根拠ごとに問われることは異なる。クヴァンヴィグのいう区切りは、主題を全体としてつかんだという感覚を伴うので、第6節がポトチュニクのいう理解の感じについて立てた問いに行き当たる。ブリチャードの根拠では、その理解がシステム自身の達成であるかどうか問われるが、人工システムが成功を自らの達成として得られるかどうかは争われている (第7章第3.3節)。リッグスの根拠、すなわち人生の認知的価値では、その理解が認知的価値を加える生がシステムにあるかどうか問われる。誰の理解がそうした価値をもつのかという問いには、AIの道具を用いる研究について、理解とは何かの説によってではなく、基礎科学が何のためにあるのかについての主張によって、直接に答えた論者もいる。ポトヴィニク (Botvinick) とガーシュマン (Gershman) は、いくつかの見解とそれへの応答を集めた共同論文のなかの自らの見解で、そうした道具を用いて基礎科学を行うなら『誰の』理解が重要なのかと問い、人間の理解が科学の中核的な目標であり続けることを求めて、「私たちは理解を人工システムに明け渡すことはできない」と答える。二人は、技術的な限界が何の障壁にもならない場合でさえ、この目標を、取り組む問題の選択とともに人間の科学者に残そうとする。ただし、人間の認知の限界のために、いつかAIシステムが、人間には表現できない自然の側面を表現するようになる可能性は認めている。その場合にも、二人の見方では、そうしたシステムからできるかぎり多くの人間の洞察を引き出すことが求められる (Binz et al. 2025, 6-7)。この答えは担い手を定める。基礎科学において価値が数えられる理解は、人工システムが何を表現しうるにせよ、人間の理解だというのである。第1章第2節に照らすと、この答えは、*S* が人間であることから出てこない担い手についての要件を、人間の *S* に加えるものと読める。同節で見たように、*A* も *S* も達成

の担い手を決めず、ハンフリーズの説では、科学が人間のために行われることは、それだけでは、科学が知るようになることが人間の現在の理解の内にあることを求めないからである。

続く第7.2-7.4節を読むうえで、さらに二つの点が重要である。第一に、理解はそれ自体のために価値がある、あるいは探究の目的であるという主張は、理解が可能にするものの一覧からは出てこない。また、その効果のいくつかが他の仕方でも得られると示しても、反駁されない。第二に、この第一の点から、それらの節でたどる関係は、理解の価値を尽くすものとして示されるのではないことになる。上で見た、理解に最終的な価値を認める立場では、それらは理解の価値を尽くさない。それらが記録するのは、継続する研究システムが理解なしに何を失うかを問うときに見えてくる部分である。二種類の価値は競合する必要がない。クヴァンヴィグによる理解の価値の説明は、組織された思考が行為へ向けた推論を支えることと、それが探究に満足のいく区切りをもたらすことの両方に訴えている (Kvanvig 2003, 202)。さらに、ヒルズが支持するような能力説では、理解に含まれる把握そのものが一種の能力ないしノウハウであり、ヒルズが道徳的な理解について述べるところでは、そこには説明する能力や新しい事例について結論を引き出す能力が含まれる (Hills 2020, 12)。したがって、以下で論じる能力が何を可能にするかを記述することは、理解を単なる手段として扱うことにはならない。

第一の点は、理解とそれが可能にするものとの分離を、歴史的な変化として扱う説にも見てとれる。高橋恒一 (Takahashi) は、AI駆動科学を論じるなかで、近代において科学と技術を結びつけていたのは理解であったとする。自然法則を理解し、その理解を技術に転化してきたというのである。そのうえで高橋は、理解を経由しない予測が出現しつつあるとし、AlphaFold2 によるタンパク質の構造予測、深層学習に基づく気象予測モデル、AIによる材料候補の絞り込みを挙げる。そして、理解を経由せずに予測と制御が可能になるなら、高橋のいう科学と技術の結婚の理由は消え、両者は離婚へ向かうと論じる。ただしこの説では、そこで失われるのは理解そのものではなく、手段としての理解である。どうやって結果に至るかがAIに委ねられた後にも残る理解は、それ自体が目的となる。たとえば、なぜ宇宙は始まったのかという問いは、高橋によれば、技術的には無用だが人間の知的営みとしては本質的である (Takahashi 2026, 184-186)。こうした予測が誰の理解を経由しないのかを、高橋は述べていない。ただし、材料科学の例では、物理的な説明を後から与えるのは人間である。また高橋は、AIによる予測と科学的理解の関係を、単純な分離ではなく新しい結合の形として理解する余地も認めている。人間の研究者が、予測された構造を起点に新たな仮説と理解を生み出している場合がそれにあたる (185)。理解の特別な価値の多くを、予測し制御できる立場に求めるグリムの説では、高橋の描く分離のもとで、その価値の多くが理解なしに得られることになる。これに対して、上で見た、理解に最終的な価値を認める説では、第一の点が述べたとおり、それらの説が認める価値はその分離によっても損なわれない。理解がそれ自体の目的として残るという高橋の描き方は、この線に沿って読める。ただし、この描き方はこれらの説に依拠しているわけではない。

7.2 事態がどう違っていたかを把握すること：転移、介入、診断

第3.2節の依存関係説と第3.3節の能力説は、理解が何から成るかについて意見を異にする。前者は依存関係のモデルを把握することに、後者は主体が何をできるかに、理解を置く。それでも本書は、両者が共通して、ひとつの様相的な能力を問題にしていると読む。様相的とは、現に成り立っていることだけでなく、事態が違えば何が成り立つかわかる、という意味である。その能力とは、ある対象が依存している要因が違っていたら、その対象がどう違っていたかを言えるほどに、依存の構造を把握する能力である。そう読む理由は、各説が理解の中身やしるしとして挙げるものが、いずれもこの能力にかかわることにある。依存関係説が把握を求める依存関係とは、対象が、それが依存する要因に応じてどう変わるかという関係である。また、能力説が挙げる能力、すなわち似た事例について結論を引き出すこと、「もし〜なら」の問いに答えること、表象を作り変えて新しい推論を支えることは、いずれも現に与えられた事例の外へ出る (第3.2節、第3.3節)。

以下の論者の記述は、この読みを裏づける。ローレン・ロス (Lauren Ross) とウッドワードが紹介するところでは、ウッドワードの介入主義の説における説明は「もし事態が違っていたら」という問いに答え、挙げられた要因が変わったなら被説明項がどう違っていたかを告げる (Ross and Woodward 2023, §3)。グリムは、理解が取り込む依存の地図を可動的なものとして描き、理解する心を、ある変数の値を変えると他の変数の値がどう変わるか、あるいは変わらないかを予期できる心として描く (Grimm 2017)。ヒルズの認知的制御には、関連して類似した事例について結論を引き出すことが含まれる (Hills 2020, 12)。クオリコススキ (Kuorikoski) とイリコススキ (Ylikoski) は、理解を、事態が違っていたら何が起きたか、ある事柄を変えたら何が起きるか、望む変化をどうすればもたらせるかについて正しく推論する能力から成るものとみなす (Kuorikoski and Ylikoski 2015, §2)。ル・ビアン (Le Bihan) の様相的な見方は、これをひとつの構造の外へ広げる。ル・ビアンの説では、ある現象を理解することは、その現象を生じさせる可能な依存構造の空間をどう航行すればよいかを知っていることである。これを様相的理解 (modal understanding) と呼ぶのは、それが、現象が現実の世界でどう生じているかだけでなく、ほかの可能な世界でどう生じうるかにかかわるからである。ル・ビアンのねらいは、世界を誤って表していると分かっている理論やモデルも、現象の理解をもたらしうることを説明することにある。たとえば、分子どうしが力を及ぼし合わないとする理想気体のモデルと、分子間の引力と分子の大きさを加えるファン・デル・ワールスのモデルは、気体のふるまいを生じさせる、異なる可能な依存構造を表している。ル・ビアンの説では、それぞれの構造がどのふるまいをどう再現するかを知ること、構造どうしがどう関係し、どこで異なるかを知ること、そして熱力学の法則のように、可能などの構造にも課される制約を知ることが、この理解の三つの水準をなす (Le Bihan 2017)。

この能力の位置づけについては、説のあいだで意見が分かれる。理解を構成するものとみる説もあれば、理解の現れや証拠とみる説もあり、そこで働く能力は理解に特有のものではないとする説もある (Grimm 2021, §2.2)。たとえばル・ビアンは、そうした空間をどう航行すればよいかを知っていることを様相的理解を構成するものとみなす一方で、推論し、転移し、「もし〜なら」の問いに答える能力を、それを構成するものではなく、その現れとみなす (Le Bihan 2017)。この能力が理解を構成するにせよ、理解の現れにすぎないにせよ、それを欠くシステムは、以下の場面が頼りにしうるものの一つを欠く。そのため本小節は、これらの説のどれかを先に選ばなくても、理解がなければ続いていく研究システムが何を失うかを問うことができる。

本小節では、時間を通じて続いていく研究システムがこの能力を使う三つの場面を取り上げる。学んだことを新しい文脈へ持ち込むこと (転移)、何かを意図的に変えること (介入)、そして予期が裏切られたと分かること (診断) である。新しい条件のもとでの予測は、転移と介入に含まれる。予測を別に挙げないのは、本小節の問いが、妥当性確認済みの予測に理解が何を加えるかだからである。第4節の強化した予測の事例は、対象となる規則性がなぜ成り立つのかを共同体のどの人間の成員も理解していない場合にも、そうした予測をすでに認めている。そのため、予測だけでは、理解がなければ何が失われるかを示せない。

第一の使い方は、転移と一般化である。ここでいう一般化は、いくつかの事例から得たものが、それらの事例を超えて成り立ち続けることであり、機械学習で汎化と呼ばれるものに近い。関連する依存関係の表象は、文脈が変化してもどの規則性が存続するかを示しうる。それは、ある代理指標がなぜうまく機能したのか、そしてどこで失敗するのかを明らかにしうる。理解は一般化を保証しないし、頑健な経験的方法は理解なしに転移しうる。それでも理解は、外挿を単一の成績の履歴に委ねるのではなく、提案された転移の根拠を批判に開かれたものにしうる。そのような転移がなぜ理解の証拠となりうるのか、あるいはなぜ理解から帰結しうるのかを、第3節の諸説はそれぞれ異なる仕方でも説明することになる。デルセンのような依存関係説なら、転移は、把握されたモデルが表す依存関係が新しい文脈でも成り立ち続けることに支えられる (Dellsén 2020)。デ・レフトがボーゲとともに述べ直すところでは、現象の理解をもたらす説明を構成した科学者は、同じ技能を用いて、対象となる系についての知識をさらに適用し、拡張し、精緻化しうる。転移はそうした拡張の一つである (Boge and de Regt 2025, 420)。ヒルズの能力説では、認知的制御には関連して似た事例について結論を引き出すことが含まれる。そのため転移は、理解を部分的に構成する能力の発揮である (Hills 2020, 12)。第3.1節で見たカリファの還元的な説も、理解と転移が結びつくことを否定するわけではない。この説は、その結びつきを、理解に特有の能力を持ち出さずに説明する。その説では、なぜそうなのかを理解することは、その説明を知ることにはかならず、理解にかかわる能力は、グリムが紹介するカリファの言い方では「とりわけ特別でない (especially unspecial)」ものであり (Grimm 2021, §2.2)、ありうる説明を検討して比較し、説明する際に働く能力である (Khalifa 2019, 10, 31)。またこの説では、理解は、把握された説明的な情報の量に応じて深まる。そのため、何かがなぜ成り立つかに加えて、それを説明するものがどのようにしてそれをもたらすかも把握している人は、前者だけを把握している人よりよく理解している (Khalifa 2019, 33)。

上で、理解は、結果が新しい文脈でも成り立つと見込む理由、つまり提案された転移の根拠を、これまでの成績の履歴に任せるのではなく、批判できる形で示しうると述べた。ウッドワードの説は、そうした理由を述べる一つの仕方を与える。「この依存関係は、ある範囲の介入と背景状況のもとで変わらずに成り立ち、新しい文脈もその範囲に入る」と述べる仕方である。この説からは、こうして述べた転移の主張について、少なくとも三つのことが出てくる。第一に、その主張は、元の場面を超えて成り立つことについて支えを要する。ウッドワードの説では、因果的な一般化、つまりある変数が別の変数に因果的にどう依存するかを述べる一般的な主張（ここでの一般化は、上の転移と一般化という意味とは異なる）は、介入と背景状況のある範囲のなかで成り立つ。それが他の被験者や状況でも成り立つかどうかは、実験計画の文献が外的妥当性と呼ぶ問いであり、分析上は元の実験で成り立ったかどうかとは別の問いである。ただしウッドワードが指摘するように、二つの判断がまったく別の証拠を用いる必要はない。文脈を超えた安定性の証拠が、局所的な因果の主張を支えることもある (Woodward 2026, 5–6)。第二に、その支えには、範囲の境界をすべて知ることまでは要らない。ウッドワードの論じるところでは、一般化を使ってある系のふるまいを説明するには、その系が一般化の不変性の領域、つまり依存関係が変わらずに成り立つ範囲のうちにあることを知る必要があるが、その領域の正確な境界を知る必要はない (Woodward 2000, 233–234)。第三に、転移の主張は、何を持ち込むかによって成否が分かれる。構造方程式モデルを例に、リー (Li) らはウッドワードの区別を報告している。量的な主張は、何が何を引き起こしうるかという質的な主張がより広く成り立つ場合でも、母集団に固有でありうるという区別である (Li et al. 2018, §2)。そのため、こうして述べた転移の主張には、はっきりした問いを向けることができる。新しい文脈は本当にその範囲に入るのか、持ち込めるのは質的な関係だけか、それとも量的な関係までか、という問いである。転移の根拠が批判に開かれているとは、このことである。

第二の使い方は、介入と構築である。予測は行為を導きうるが、介入には、原因、制約、メカニズム、操作可能な変数を区別することが必要となりうる。予測器は、その入力そのものが政策や実験によって変えられるとき、失敗することがある。ウッドワードが論じる例の一つは、マクロ経済学のルーカス批判 (Lucas critique) である。ウッドワードの説では、この批判の争点は、経済モデルが依拠する一般化が、経済主体にとってのインセンティブ、相対価格、利用できる情報などの変化のもとで不変であるかどうかにある (Woodward 2018a, §4)。説明的理解や因果的理解は、介入点と起こりうる副作用を明らかにしうる。工学的理解も同様に、人工物のふるまいを、何を変更し、修理し、あるいは別の仕方でも構築できるかに結びつける。

この能力の使い方として介入を挙げても、この能力を因果的な依存に限ることはない。第4節と本節のここまでで見たウッドワードの介入主義的な説は、説明的な情報を操作に結びつける。説明するとは、挙げた要因に介入して変えたら結果がどう変わるかを示すことだ、とするからである。しかしウッドワードは、この説を因果的説明に限っている。そのうえで、「もし事態が違っていたら」という問いに答えながら、その問いを介入についての問いとしては読めない導出もあり、それは何らかの非因果的な説明でありうると認める。取り上げる例の一つは、惑星の軌道の安定性が空間の次元に数学的に依存するという議論である。四次元以上の空間では、安定した惑星の軌道はありえないという。この導出は、空間の次元が変わったら安定した軌道の可能性がどう変わるかを告げるが、空間の次元を変える介入という考えは役に立たず、理解できないものでさえあるかもしれない (Woodward 2018b, §§2–3)。したがって、ウッドワードの説はあらゆる理解が因果的であることを含意せず、本小節がその説を用いることも、それを含意しない。

第三の使い方は、異常の診断と修正である。実験が予測と衝突するとき、その結果は、仮説、補助仮定、装置、前処理上の選択、環境条件の束にかかわる。結果はそれ自身の修正対象を特定しない。依存関係についての組織立った説明は、焦点となる仮説、測定モデル、境界条件、あるいはテストそのもののいずれを修正するかを研究者が決めるのを助けうる。これは、単一の結果が理解なしに根拠づけられている場合でさえ、理解が自己訂正にとって重要になりうる理由のひとつである。第2章第5節で見たデュエム (Duhem) 流の決定不全性とは、不利な結果だけでは、それらのうちどれを改めるかが決まらない、という論点である。デュエムのテーゼそれ自体は誤りの所在を突き止める手順を与えないので、それを突き止めるには、この種の能力が求められる。エルギンは、期待が外れたことが示すのは、コミットメントのまとまりのどこかが誤っているということだけだが、すべての要素が等しく誤りやすいわけではないと論じる。ほかのよく確かめられたまとまりのなかで独立に支持されている要素は、誤りの所在である見込みが低い。それゆえ背景がよく支持されているほど、誤りの位置を精確に絞り込めるとする (Elgin n.d., 17–18)。ゴードンが挙げるヒルズの認知的制御の能力 (第3.3節) のうち、二つは互いに逆向きの推論である。一つは、 q という情報から p という結論を引き出す能力であり、説明するものから説明されるものへ、予測と同じ向きに進む。もう一つは、 p を与えられたときに q をその説明として挙げる能力であり、説明されるものから説明するものへとさかのぼる (Gordon n.d., §3c)。診断がとるのは後者の向きである。診断は、外れた予期から出発して、どこが誤っていたかへとさかのぼるからである。また、ウィルケンフェルドが挙げる理解の例には、自分の表象のなかの不整合や誤りに気づいて直せることが含まれる (Wilkenfeld 2013, 10)。ヒルズのいうさかのぼる能力も、ウィルケンフェルドのいう誤りに気づいて直す能力も、選ばれた修正が正しいことを保証するものではない。

7.3 領域を結びつけること：保持、再利用、さらなる問い

第7.2節が依存の構造の把握に注目する説に依拠したのに対して、文献のもう一つの流れは、理解を、情報どうしがどう結びついているかの把握に求める。クヴァンヴィグは、第7.1節で見たように、またグリムの報告するところでも、理解には、主題にかかわるさまざまな情報のあいだの説明的・概念的なつながりを把握し、それらを一つのまとまりへと組み合わせることが求められるとする (Grimm 2021, §§2.1, 2.3)。エルギンは、理解を、事実根ざし、証拠に応答し、その主題について自明でない推論、論証、そしておそらく行為を可能にする、包括的で体系的に結びついた情報の集まりへの認知的なコミットメントとして特徴づける (Elgin 2018, 8)。デルセンの説は、なぜそうなのかの理解ではなく現象の理解についてのものであり、ある現象が置かれている依存関係をモデルとして把握することに理解を求める。ここでのモデルとは、対象を表すように解釈された情報の構造である。たとえばロトカーヴォルテラの方程式は、その変数を捕食者と被食者の個体数として読むとき、捕食のモデルになる。理解に要するモデルは、現象が何に、どのように依存し、何に依存しないかを表すものである。デルセンによれば、そうしたモデルは二つの仕方で失敗しうる。依存関係を誤って表すことと、それをそもそも表さないことである。そこでデルセンは、モデルを二つの基準で測る。誤って表していないかを問う正確さと、表すべきものを表しているかを問う包括性である。どちらの基準も、満たすか満たさないかの二択ではなく、どれだけ満たすかが問われる。そのため理解も程度の問題になる。人は、ある現象の依存関係について正確で包括的なモデルを把握している程度に応じて、その現象を理解している。また包括性には、何が何に依存しないかを表すことも含まれる。因果関係の図が、どれがどれの原因かだけでなく、どれとどれが因果的につながっていないかを示すのと同じである。そのため、関係がありそうに見えた二つの要因が実は独立だと分かることも、理解を深める。たとえばガリレオの思考実験は、自由落下する物体の加速度がその質量に依存しないことを示す (Dellsén 2020, §§3, 7)。

では、こうした結びつけは、第7.2節の能力、つまり対象が依存するものが違っていたら対象がどう違うかを言える能力と、どう関係するのか。結びつけは、その能力のもう一つの使い方でありうるし、その能力を支える別の達成でもありうる。説によって、示す方向が異なる。イリコスキとクオリコスキは、二つを結びつける理由を与える。二人によれば、より大きな理論の枠組みにうまく統合された説明は、二つの知識をばらばらに考えた場合より多くの理解をもたらす。統合によって、答えられる「もし〜なら」の問いの範囲が、二つの仕方で広がるからである。第一に、既存の理論との推論上のつながりを通じて、その理論のなかの要因と現象のさまざまな側面とのあいだの依存関係が、その現象について「もし〜なら」の問いに答えられる、それまで予見されていなかった次元を開きうる。第二に、その説明そのものが既存の理論の隙間を埋め、元の現象とは別の現象について新しい「もし〜なら」の問いに答えることを可能にしうる。そのため、二人のいう統合は、形式的に矛盾しないことや概念上まとまっていることにとどまらず、その現象について新しい関連する推論を可能にするものでなければならない (Ylikoski and Kuorikoski 2010, §4.4)。ただし、結びつけによって答えられる問いが増えることだけでは、結びつけがそのどちらなのかは決まらない。別の達成でも、同じ結果をもたらしうるからである。クヴァンヴィグのように説明的な関係以外の整合をなす関係も数える説や、ケルプのように説明的でない知識も理解に寄与するとする説 (第7.1節) では、領域を結びつけるときに把握されるものに、依存以外の関係も含まれる。そうした説では、結びつけは、依存関係を把握する能力には尽きない達成でありうる。

本小節では、こうした結びつけのもとで、続いていく研究にとって重要になる二つの関係を取り上げる。一つは、得たものを時間を通じて保ち、のちの探究で使い、広げること、すなわち保持と再利用である。もう一つは、さらなる問いを開くことである。クヴァンヴィグは、説明的な関係を把握することは、真なる信念をさらに付け加えることを超える仕方で、主題についての思考を組織し体系化すると考える。その考えによれば、そうした組織が役に立つのは、ある情報から、行為の基礎として役立つ関連する別の情報へと推論できるようにするからである。組織されていない思考は、そうした推論の足場を与えない (Kvanvig 2003, 202)。この点は再利用にかかわる。ゴードンが挙げるヒルズの認知的制御の能力には、 q' という情報から p' である (あるいはおそらく p' である) という結論を導く能力も含まれる (Gordon n.d., §3c)。この点も再利用にかかわる。まとめれば、これらの説では、組織された結果の集まりは、そこに含まれるものから推論して新しい事例へ広げることができる。これに対して、孤立した答えの寄せ集めには答えどうしを結ぶ推論のつながりがないので、新しい事例ごとに答えを得直さなければならない。保持については、『メノン』に先例がある。そこでソクラテスは、真なる意見を、つなぎ留めておかなければ逃げ出すダイダロスの像にたとえる。真なる意見は「原因という紐で縛られる」までは大した値打ちがなく、縛られれば永続する。知識が真なる意見より尊いのはそのためである (Plato, *Meno* 97e–98a)。グリムの報告によれば、プラトンやアリストテレスのエピステーメーは知識よりも理解と訳すほうがよいという主張がなされてきたので (第7.1節)、そう読めば、この箇所は、保たれることを原因の把握に結びつけている。ウィルケンフェルドは、理解とは、理解されるものについての情報を使えるように圧縮することだとする説をとり、現代の事例を挙げる。健全性の証明を丸暗記した学生は、一つの段階を忘れただけで立ち往生したであろうが、証明を理解している論理学者は、より単純な表象からそれを再構成できる (Wilkenfeld 2019, 7)。教育研究も並行する主張を述べている。ノヴァク (Joseph D. Novak) とカニャス (Alberto J. Cañas) の報告では、既存の知識とほとんど統合されない機械的な暗記で学ばれた知識は、よく復習されなければすぐに忘れられやすく、これに対して有意義な学習は、知識を長期にわたって保つことを支える (Novak and Cañas 2008, 6–7)。これらの点は保持にかかわる。すなわち、根拠との結びつきや、作り直しを可能にする組織が、それらなしでは失われるものを保ち、あるいは取り戻す助けになりうることを示唆する。

第二は、さらなる問いを開くことである。理解は、ある領域を再編成し、新たな対比、異常、介入を明らかにしうる。予測システムもまた、パターンや失敗を発見することによって新たな問いを生み出しうる。違いは、理解だけが新しさを生み出すという点にあるのではない。関係の表象は、孤立した答えにはできない仕方で、可能な問いの空間を航行可能なものにしうるという点にある。イリコスキとクオリコスキの統合の説は、この主張を一つの形で述べたものである。エルギンは、理解する者は理解していることを使えなければならない、おそらくは関連する事柄についての作業仮説の源としても使えるはずだとする (Elgin 2007, 5)。さらにエルギンは、隠喩がするように領域を組み替えることが、それなしには立てることも動機づけることもできなかった新しい問いを立てる手立てを与えると論じる (Elgin 2002, 5, 9, 12)。関係は逆向きにも働く。第7.1節で触れたように、クヴァンヴィグは、組織された理解が探究に満足のいく区切りをもたらすと述べ、プリチャードは、説明を求める探究は、求める理解に達するまで正当には閉じられないと考える (Kvanvig 2003, 202; Pritchard 2016, 6–7)。二人の説では、理解は探究を閉じるものである。そうすると、理解はさらなる問いを開くという本小節の主張と食い違うように見える。しかし二つの主張は、範囲の異なる問いにかかわると読める。理解が閉じるのは、探究のきっかけとなった問いであり、理解が開くのは、その先にある新しい問いである。たとえば、ガリレオの思考実験によって、自由落下する物体の加速度がその質量に依存しないと把握すれば、重いものほど速く落ちるのかという問いは閉じる。しかし同時に、では加速度は何に依存するのか、という問いが開かれる。

7.4 理解を他者に利用可能にすること：コミュニケーション、批判、記憶

説明は、関係を、他者が検討し、教え、異議を唱え、既存の知識に結びつけることのできる形へと圧縮しうる。言語化は、実践が言語化されないままのあいだは見えなかった曖昧さを露わにすることがある。それは誤解を招くこともある。標準化された説明は、共有された把握という錯覚を生み出しうる一方で、ある種の専門知は部分的に暗黙的なままにとどまる。理解そのものを他者に伝えられるかどうかは争われている。グリムが反論するために再構成する一つの伝統は、事柄がどう結びついているかの把握は各人が自分で行うものであり、結びつきがあると人から告げられても、それだけで自分がそれを把握したことにはならない、と考える (Grimm 2019, §II)。ヒルズは、なぜそうなのかの理解には、証言だけではふつう渡せない能力が含まれると論じる (Hills 2020, 5, 12–13)。他方で、受け手がそれを受け取るのに必要な背景をすでにもっていれば、証言は理解をもたらしうる、と論じる者もいる。これらの立場と、他人から学ぶ仕方についてのヒルズの区別は第8節で取り上げ、人工システムの出力に依拠する研究システムにとってそれらが何を意味するかもそこで論じる。デ・レフトは、科学的理解を共同体の達成とみなし、受け手は他者の説明の推論の筋をたどって再現することによって理解を得ると考える。その再現の詳しさには程度の差があり、それに応じて得られる理解の程度も異なる、とも言う。そのうえで、説明する側の理解の基礎にある理解可能性は、受け手が理解を得るための要件でもあるようだ、と結論する (de Regt 2017, 91, 99–100)。クオリコスキとイリコスキは、科学的理解を本質的に集合的なものとみなし、形式モデルのような外的表象は、仮定を明示させて評価しやすくすることによって理解を高めうると論じる (Kuorikoski and Ylikoski 2015, §§2–3)。そうだとすれば、問題になるのは、受け手が理解に要る把握を実際に得るかどうかであって、説明の形をした文書が生成されたかどうかではない。第6節で見たボーグ、シュスター、シュトルは、ここで説明を中心に置く。三人の説では、説明は科学的理解の媒体であり、定まった、伝達可能な仕方で点検することができる (Boge, Schuster, and Stoll 2026, 5–6, 11)。

批判と記憶は、この利用可能性に依存する。他者がたどることのできる表象は、それを生み出した者以外の誰かによって点検されうる。他者がたどることのできる記録は、それを生み出した者が去った後にも立ち戻ることができる。理解を他者に利用可能にすることがなぜ批判と記憶を支えるのかについて、本小節で挙げた理由のうち二つは、人間の研究の特徴に依っている。一つは、結果を生み出した者がやがて去ることであり、もう一つは、デ・レフトの説で、受け手が推論をたどり再現することで理解することである。Sが人間であれば、結果を生み出したのが何であれ、その人間の成員がたどれる表象の必要は残る。生み出したシステムが存続し、受け手がSの人工の成員である場合にも、生み出した側と同じ仕方では誤らない点検はなお必要であり、それは独立性の問題になる (第11章第2節)。ただし、利用可能性が何を求めるかは形を変える (第5.4節)。ヒルズは、道徳的判断について論じるなかで、一つの限界を付け加える。専門家の言葉をそのまま受け取る者の判断は、その判断を正しくしている理由ではなく、その言葉を受け取る理由に応答している。そして言葉を受け取るだけでは、状況が違えば自分で違う結論を引き出す能力は得られない (Hills 2020, 12–13)。ヒルズの説を採るなら、その能力をもたないまま人工システムの出力に従って行為する共同体は、システムが追跡している依存関係ではなく、システムに導かれていることになる。

こうして第7.2～7.4節からは、本節の冒頭で立てた問い、つまり誰も理解しないとき、ひとつの妥当性確認済みの結果を超えて何が失われるのかという問いへの答えが見えてくる。失われうるのは、批判に開かれた転移の根拠を述べ、介入し、失敗を診断する能力、結果を保持し再利用し、さらなる問いを開くこと、そして結果を他者に利用可能にし、批判と記憶を支えることである。ただし、それらがほかの手段で補われる場合は別である。これらは候補となる関係であって、理解の定義ではなく、本節は、あらゆる研究行為がそれらを実現することを示すものでもない。そのいくつかは、モジュール化された検証、冗長な測定、頑健な探索、専門化されたエージェント、あるいは制度的組織によって供給されうる。単純な説明それ自体が、状況の変化のもとで失敗することもある。研究システムを評価するとき、これらが提起する問いは、ある目的にどの能力が必要で、そのシステムがそれを備えているかどうかである。望ましい能力のすべてを理解と呼び替えても、この問いには答えられない。また、これらの能力を他の手段で供給しても、理解がそれ自体のために価値をもつか、探究の目的であるかという問いは決着しない (第7.1節)。

ここから次の結論が得られる。作業的特徴づけのもとでは、研究がどの認識的達成を目指すかは開かれているが、特定の探究や持続的な研究システムには、理解と結びつけられる能力が必要になりうる。

作業的特徴づけを採るなら、研究がどの認識的達成を目指すかは、知識であれ理解であれ、特徴づけによって定まらず、個々の試みに開かれている。また、その達成を得ることを、各研究行為の条件にはしない。理解が必要かどうかは、試みが求める達成による。そのため、理解をそれ自体のために重んじる立場や、理解を探究の目的とみなす立場は、目指す達成として理解を置くことで、この特徴づけのもとに収まる。そうした立場のもとでは、何が必要かはその価値や目的によって定まる。それでもなお、特定の探究や持続的な自律的研究システムは、理解を目指していない場合でも、重要な諸理論が理解と結びつける能力を必要としうる。それらの能力が何に資さなければならないかは、Sが人間か人工かで異なる (第7.4節)。

この結論は、作業的特徴づけそのものを見直す手がかりにもなりうる。必要になりうるとされた能力は、理解なしでも得られるのかもしれないし、逆に、特徴づけのほうに、研究にとって理解が占める位置をとらえきれていないのかもしれない (序論第4節)。

それらの能力が合わさって理解を構成するかどうかは、哲学的な問いとして残る。あるシステムがそれらの能力を実際に保有しているかどうかは、経験的かつ工学的な問いである。二つの問いは区別されるが、互いに情報を与え合う。

8. 局所的な達成が継続的な探究能力の問題を決着させない理由

ここまでの節は、主にひとつの結果やひとつの探究を単位として、ある結果が理解なしに根拠づけられうるか、ある探究を閉じるのに理解が要るかといった形で理解を問うてきた。本節では、この単位を局所的と呼ぶ。そのうえで、何を評価するのかと、どれだけの期間で見えるのかが変わる。評価するのは、ひとつの結果ではなく、そうした結果を繰り返し産出し利用するシステムや共同体であり、見るのは、ひとつの探究ではなく、時間を通じて続いていく探究である。この切り替えが重要なのは、二つの水準での判定が、どちらの向きにも食い違いうるからである。ある特定の不透明な結果が根拠づけられている一方で、そうした結果を繰り返し産出し利用するシステムが、重要な認識的能力をしだいに失っていくことがありうる。逆に、不透明な構成要素が、他

のエージェント、装置、記録、制度が信頼できるチェックと修復を供給する頑健なシステムに属しており、その構成要素の不透明さがシステムの問題にならないこともありうる。

システムが認識的能力を失っていくリスクの一つの形を、メッセリ (Messeri) とクロケット (Crockett) は *理解の錯覚* (illusions of understanding) として描いている。科学者が、世界について実際以上に理解していると信じてしまうことである。二人の説では、AIツールは研究をより生産的で客観的にすると期待されるため、研究者が互いの専門知識に頼るのと同じように、科学者はそれを協働者として信頼しやすい。そして他所にある理解に頼る者は、それを自分の理解と取り違えることがある。さらに、そうしたツールが与える単純で定量的な答えは理解の感覚を生みやすいが、その感覚は実際の理解と一致するとは限らない。またツールの効率性、AIに向けた問いと方法、そしてAIが表現できる立脚点が他の進め方を押しのけるモノカルチャーを育てうる。そうならば科学は誤りに対してより脆弱になる (Messeri and Crockett 2024, 49–53)。二人の分析はリスクの枠組みであって、AIに支えられたあらゆる分野がすでにこれらの効果を示していることの証拠ではない。それでもその分析は、継続的な認識的能力を局所的な出力量とは別個に評価する直接的な理由を与えている。どちらのリスクも、ひとつの結果の質を見ても現れない。一つ目は、ツールに依拠する者が自分は何を理解していると思うかにあり、二つ目は、依拠が分野に広がり時間的に続くにつれて現れる。そのため両者は合わせて、何を評価するのかと、どれだけの期間で見るのかの両方にかかわる。モノカルチャーのリスクについては、六つの自然科学分野の論文を対象にした観察研究が、このリスクと整合するパターンを報告している。ただし、それがAIの手法を使ったことの効果であるとは示していない。この研究は第11章第8節で論じる (Hao et al. 2026)。理解の過大評価のリスクについては、その研究は証拠を与えない。メッセリとクロケットが書いているのは、人間の科学者についてである。二つのリスクのうち、モノカルチャーは探究の組織にかかわるので、成員が同じ問い、同じ方法、同じ誤り方を共有するようになりうる *S* であれば、人間の社会でなくても生じうる。これに対して理解の錯覚は、二人の描き方では、同僚の専門知識に頼るのに倣ってツールを信頼することや、実際の理解と一致するとは限らない理解の感覚といった、人間の認知の特徴に依っている。そのため、*S* の人工の成員がこれに対応する錯覚に陥りうるかどうかは、その成員のなかに、その感覚の役を果たすものがあるかどうかによる (第6節を参照)。

何を評価するのかの切り替えは、現代の研究のありふれた特徴からも見てとれる。現代の研究プロジェクトは、認知的分業に依存しうる。そこでは参加者が、各自では完全には理解していない装置、ソフトウェア、手続き、あるいは専門家の主張に依拠する。第2章第6節で見たとおり、ブラウンが再構成するギアリー (Giere) の説のように、そうしたプロジェクトで認知が分散していると記述する場合でも、それだけでは、組織されたシステムもまた知っているかどうかは決まらない (M. J. Brown 2015)。また、集合的な産出それ自体は、その結果として得られる知識を誰が保有するのかを決着させない。こうした分業があるので、すべての参加者がすべてを完全に理解していることを研究に求めることはできない。しかし、そのことは、不透明さがさらに加わっても害がないことを示すわけではない。分業が問いとして残すのは、依存がどのように組織されているかであり、この問いは、ひとつの結果についてではなく、プロジェクト全体の組織について問われる。すなわち、信頼性がどこで確立されるのか、相関した失敗がどのように検出されるのか、誰が異議に応答できるのか、そして条件が崩れたときに依拠の仕方がどのように変わるのか、である。

こうして評価は二つの水準に分かれる。これまでの節が主に扱ってきた個々の貢献の水準と、本節が扱う、継続するシステムの水準である。個々の貢献の水準では、どの問いが取り組まれたのか、そして結果として得られた主張やその他の達成が、主張された意味において真または適切であるか、根拠づけられているか、説明的であるか、あるいは理解にあたるかを問う。継続するシステムの水準では、その研究の組織が、多様な探索、独立した批判、記憶、依存関係の追跡、訂正、さらなる問いの形成、*S* の成員が研究を行い評価する有能さの再生産などのための能力を保持しているかどうかを問う。局所的に根拠づけられた出力の集まりは、それ自体では、継続するシステムの水準の問いに答えない。二つの評価はどちらの向きにも食い違いうるので、一方から他方を推論することはできない。結果が根拠づけられている一方で研究の組織が能力を失っていくこともあれば、能力が保たれている一方で個々の結果が失敗することもある。たとえば、出力を一件ずつ確かめながら予測器に頼るようになった研究室は、根拠づけられた結果を積み重ねつつ、予測器なしに異常を診断する能力を失っていくかもしれない (第4節で見たハンフリーズの別の懸念が指すのは、この損失である)。そして近隣のグループが同じモデルに頼っていれば、そのすべてに共通する失敗は、各グループが自分の結果に対して行う確認をすり抜ける。前者の損失は研究室を時間を通じて見たときに、後者は評価の範囲をひとつのグループの外へ広げたときに、はじめて見えてくる。

継続するシステムの水準で最後に挙げた能力、すなわち *S* の成員が研究を行い評価する有能さの再生産のための能力は、研究者の世代をまたいで見たときにはじめて分かる仕方で見失われることがある。カステラッチ (Castellacci) らは、研究における生成AIをめぐる円卓討論から生まれたプレプリントで、これを能力の枯渇 (capacity depletion) と呼ぶ。それは、AIが引き受ける作業が、研究の能力を身につけるための作業そのものである場合に生じる。そのためカステラッチらの説では、AIは、研究の産物へのアクセスを広げても、それを生み出す能力までは広げないかもしれない。著者たちはこれを、開示されないAIの使用が評価者に生む情報の非対称性や、共有された知識基盤への負の外部性とは別の、現在の導入にとまなうリスクとして示す。また、自分たちが引く学習についての研究を、道具を使うかどうかではなく、道具とどう関わるかが効いてくることを示すものとして読んでいる (Castellacci et al. 2026, §§1, 2.4)。ベンダー (Bender)、バーグストロム (Bergstrom)、ウェスト (West) も、第7.1節で取り上げたビンツら (Binz et al.) の共同論文に収められた応答のなかで、科学が何を生み出すかについて関連する点を述べる。科学は、論文に記録される結果だけでなく、その行い手である科学者そのものも生み出すというのである (Binz et al. 2025, 8)。どちらも人間の研究者について書いている。*S* が混合社会や人工社会である場合には、対応する問いは、*S* の評価がその判断に依拠する成員が、その判断を下す有能さを備えて形成され続けるかどうかである。根拠づけられた結果を数えても、上の二つの損失と同じく、この損失は記録されない。後の議論のいくつかは、この有能さを前提にしている。ある情報源を直接に校正するには、その主張の一部について独立の意見をもつ必要がある (第8章第3.4節)。また、研究を担う人工の構成が多くの分野で *S* のどの成員よりも多くの証拠を集めて評価するようになれば、*S* は全体として、その人工の構成の判断を自ら点検できず素人として頼るほかない位置に立つようになりうる (第11章第6節)。そして、異議申立てには、それを行う有能さをもつ当事者が要る (第12章第7節)。

分業のなかで成員が他者の出力から理解を得られるかどうかは、継続するシステムの水準の問いにかかわり、この点については証言と理解をめぐる説が分かれる。ヒルズの理解の説は、上で述べた認知的分業に関わる。ヒルズは、他人の言うことから学ぶ二つの仕方を区別する。伝達 (transmission) では、聞き手は話し手の言葉を信念の理由とする。これは大規模な認知的分業を可能にするが、その結果得られるのは、理解ではなく間接的な知識である。なぜそうなのかの理解は、認知的制御、すなわち新しい事例について結論を引き出し説明を与える能力を要求し、それはたいいの能力と同じく練習から生じ

るからである。伝播 (propagation) では、言われたことが聞き手自身の探究を促し、その枠組みを与える。これは別の種類の分業であり、伝播によってなら理解は渡しうる。その議論は道徳的理解についてのものだが、ヒルズはこの区別を一般的なものとして示している (Hills 2020, 2–5, 8, 10–13, 18)。この説を採り、しかも人工システムの出力への依拠が証言に当たり (当たるかどうかは説によって分かれる。第8章第3.4節、第11章第5.1節)、伝達の形をとるとするなら、そうした出力に依拠する研究システムの成員は、理解を増やすことなしに知ることを増やしうるし、根拠づけられた結果を数えても、その違いは捉えられない。

ヒルズは、これを、あらゆる場面で理解を迫る理由とはみなさない。道徳的な証言についてのヒルズの説では、伝達よりも伝播を選ぶ強い理由が成り立つのは、伝播が可能なとき、すなわち学ぶ者が自分自身の理解を育てる時間と能力をもつときである。また、道徳的に理想的な行為者でさえ、自分が共有していない経験のように、欠かせない道徳以外の情報を与える事柄については、他人の言葉を受け入れる (Hills 2020, 14)。

他方で、ヒルズの結論が依拠する非対称性、つまり知識は証言で渡せるが理解は渡しにくいという非対称性に異を唱える者もいる。ボイドは、言われたことが受け手にとって理解しやすいとき、それが専門家にとってだけ理解しやすい場合も含めて、理解は証言から直接に得られると論じる。受け手は、新しい情報を、すでに把握している関係のなかにはめ込めるからである (Boyd 2017, 23–34)。グリムは、命題的知識の場合と同じく、受け手に適切な下地があれば証言によって理解を渡すことはたいてい容易であり、下地がなければたいてい難しく、ことによると不可能だと論じる。ただし受け手が行う仕事の種類は異なり、理解の場合、それは結びつきを把握することである (Grimm 2019, §§IV–VI)。グリムが理解についての百科事典の項目で伝えるところでは、フーは、あることが別のことを引き起こしたと知っていれば、なぜかについての初歩的な理解はすでに伴っており、知識が証言で渡せるのと同じ意味で、そうした理解も渡せると論じる (Grimm 2021, §3.2)。マルファッティ (Malfatti) は、証言は理解を伝達できないという主張を支える議論を三つに分け、いずれも退けうると論じる。第一に、証言から知識を得る場合にも、言われたことを解釈するために主題についての自分の理解を働かせる必要があることがあり、そのため要する努力の違いで両者を分けることはできない。第二に、伝達はつねに話し手を全面的に信頼して従うことを要するわけではない。第三に、把握が含むとされる能力も、ときには人から言われたことによって身につけうる (Malfatti 2020, 56–70)。

これらの立場を、根拠づけられた結果を数えることに当てはめると、数えるだけでは成員がどのような理解を得たかは分からない、という点では一致する。だが、数えることで何が見落とされるかの位置づけが異なる。ヒルズの説を採るなら、見落とされるのは、成員が人工システムの出力を信念の根拠として受け取っているのか、それとも自分で行う探究の手がかりとして受け取っているのかである。ボイドやグリムの説を採るなら、見落とされるのは、証言がそもそも理解をもたらすための背景を成員が保っているかどうかである。ボイドは、専門家の場合について、その背景を受け手自身のそれまでの努力の産物として描いている (Boyd 2017, 38–39)。フーの説を採るなら、見落とされるのは、得られた理解があるかどうかではなく、その程度である (Hu 2021, 5–7)。成員がこうして何を得て何を得不いかは、継続するシステムの水準にもかかわる。それが、システムがのちに何をできるかを左右するからである。たとえば、成員が出力なしでも失敗を診断できるか、のちの証言から理解を得るための背景を保ち続けられるか、といったことである。これらの立場は、他者が言うことや生み出すものから学ぶことにかかわる。これに対して、表象や能力を人工の成員のあいだで直接に複製あるいは共有できる場合には、受け手がそれによって理解するかどうかは、採用する説のさらに別の特徴による。能力説 (第3.3節) では、受け手が関連する能力をいままっているかどうかが問題になるかもしれない。他方で、理解はたいてい、理解する者自身の認知的能力に帰せられる達成を含むとする説、たとえばブリチャードの説 (第7.1節) では、複製された状態が受け手自身の達成であるかどうかは、なお別の問いである。

継続するシステムの水準での評価は、あらゆる研究行為に暗黙のうちに付け加えられるもうひとつの条件ではない。システムの水準の能力にほとんど貢献しないにもかかわらず価値のあるプロジェクトもあるだろうし、多くの価値ある局所的結果を産出しながら脆弱なままである制度もあるだろう。いまは研究の外の当事者も担っている、探究を取り巻く関係を、研究を担う人工の構成そのものが担うようになりうる時、この水準で生じる問題、すなわち配分、独立した批判と信頼、そして権限の問題は、第V部で扱う。本章に必要なのは、より狭い結論だけである。すなわち、理解は、各局所的結果の根拠づけに必要とされない場合でさえ、時間を通じた探究の健全性にとって重要でありうる。

9. 理解の諸説は人工システムによる研究に何を要求するか

本節では、本章の区別を、人工システムが研究を行う場面に当てはめる。そこでの帰結は一樣ではなく、四つの点に応じて変わる。どの達成が主張されるのか (予測の成績、知識、理解など)、その達成が誰のうちに実現すると主張されるのか (システム自身、その利用者、組織された全体、人工の共同体など)、その達成についてどの説を採るのか、そしてその達成が何に使われるのかである。以下では、想定した事例を用いて、順に次の点を取り上げる。予測の事例と形式的に検証された証明との対照、透明性の要求とそれが仕える働き、システムが理解しているという主張を支える証拠、長期の自律的研究が何によって評価されるか、そして参照社会 *S* と達成の担い手との区別である。ねらいは、第3～6節で見た個々の説を当てはめて、帰結が主張、説、用途ごとにどう異なるかを示すことにある。

第一に、どちらにも確かに一つの達成が認められる二つの事例を対照する。一つは、第4節の強化した予測の事例である。システムは、明示された範囲で、ある性質を信頼できる仕方でも予測する。しかし人間の利用者は、その性質を生む仕組みを説明できない。その信頼性は、システム自身の出力どうしの一一致から推したのではない証拠によって確かめられている。この場合、範囲を限った予測の成績については、システムの功績を認めてよい。また、説明的理解なしの知識を認める説では、利用者も知識を得うる (第4節)。しかし、そのどちらも、その性質がなぜ成り立つのかを誰かが理解していることを示さない。

その性質がなぜ成り立つのかを誰が理解しているのかについては、説ごとに判定が分かれる。説明に基づく説のもとでは、利用者はそれを理解していない。カリファの還元的な説では、のちに誰かがその説明を知るに至れば、その知識が理解である。そのため、理解に向けた試みは説明についての知識を得る試みになり、その知識を誰が担い、それが *S* の認識的状況に入ったかは、第2章が知識について立てたのと同じ問いになる (第3節)。デ・レフトの基準のもとでは、理解可能な理論に基づく説明がない限り、その規則性は科学的に理解されないままである。完璧な予測を出す神託についてのデ・レフトとディークスの議論に従えば、この予測の成功は大きな科学的成功にも数えられない。ただし、予測器がデ・レフトのいう理論であるなら、それをうまく使えていることは、利用者がその理論を理解していることを前提にする。その場合でも、規則性がなぜ成り立つのかの理解には、この説ではなお、その理論に基づ

く説明が要る（第4節）。ウッドワードの説では、予測器が介入のもとで何が起こるかについての正しい答えも支え、その情報が利用者にとってアクセス可能であれば、その程度において予測器は説明する（第4節）。

これに対して、デルセンの以前の定式では、多様な状況にわたって正しく予測する仕方を把握していることは、説明がなくても部分的な理解に数えられる。バイスパルトの説でも、予測器を繰り返し走らせて、入力の変化に応じて出力がどう変わるかを予測できるようになった科学者は、ある程度の対象的理解をもつ。ただし、その推論をするのに科学者が予測器そのものを走らせなければならないなら、この説では、理解をもつのは人間とコンピュータからなる結合したシステムであって、補助を受けない人間ではない（第4節）。ハイスベルスの、分類が与える理解の見方を採れば、問いはさらに、予測器の分類を支える特徴が把握できる形で利用可能かどうかになる（第3.2節）。研究を行うのが人工システムで、 S が人間の社会であるこの事例で、システムが人間の理解を可能にしたと言えるかどうかは、こうして説によって変わる。説明に基づく説のもとでは、言えない。デルセンの以前の定式やバイスパルトの説のもとでは、利用者自身が多様な状況にわたって予測器の振る舞いを予測できる程度に応じて、言える。

これと対照的なのが、生成された証明を、健全な形式的検査器、つまり証明の各ステップが規則に従っているかを機械的に確かめるプログラムが受け入れた場合である。この場合、証明がどのような経路で生成されたかは不透明なままでもよい。検査が確かめるのは形式的に述べられた言明が証明されたことであり、生成の経路が見えなくても、そのことは弱まらないからである。ただし、それでは決まらないことが二つある。一つは、その形式的な言明が、もともと問われていた問いを正しく表しているかどうかである。もう一つは、証明を手に行っている者が、その証明の説明的な依存関係、つまりなぜその定理が成り立つのかを把握しているかどうかである（第3節、第5.1節、第6章第3節）。

一つ目の点が実際に何にかかわるかは、リキッド・テンソル実験に見ることができる。ショルツェは、証明はまったく読めなかったが、形式的な定義と定理の言明は驚くほど読みやすいと感じた。そして形式化の担い手たちは、関わる対象についての形式的に証明された結果を、検査された言明が意図された対象にかかわっていることの証拠として加えた（第5.1節）。したがって、一つ目の点を判断するのに読める必要があるのは、証明ではなく、言明と、それを意図どおりに読むことを支える結果である。証明を生成するのが人工システムで、 S が人間の社会であれば、証明が人間に読めなくても、言明が人間に読めるなら、この判断は人間に残しうる（第5.1節）。

二つ目の点、つまり理解については、説ごとに判定が分かれる。ヒルズの説では、なぜ定理が成り立つのかを理解するには、その説明についていき、自分の言葉で述べ、関連して類似した事例について結論を引き出せることが要る。そのため、人間が見渡せない証明を手に行っているだけの人間の利用者は、定理を知ってはいても、この意味では理解していない。他方で、ショルツェの報告が示すように、形式化の作業に加わった人間が、証明の経路がなぜうまくいくのかを理解できるようになることはある。ただしショルツェの場合、経路の周りの地形についての感覚は得られなかった（第4節）。証明を生成したシステム自身がその証明を理解しているかについては、ウィルケンフェルドの説が具体的な試し方を与える。暗記しただけの証明とは違って、理解している証明は、少し異なる体系に合わせて作り変えられる。そのため、この説のもとでは、外から内部状態を書き換えられることなく、システムが自らその証明を作り変えられるかどうか証拠になる（第3.3節、第6節）。

このように、どちらの事例にも功績を認めてよい達成がある。予測の事例では範囲を限った予測の成績が、証明の事例では形式的な言明の証明がそれにあたる。二つの事例で違うのは、それを超える理解について、誰の何が試されるかである。予測の事例では、利用者自身が予測器の振る舞いを予測できるか、それとも予測するのに予測器を走らせ続けなければならないかが問われる。証明の事例では、人間の利用者が証明の説明についていけるか、そしてシステムがその証明を自ら作り変えられるかが問われる。

第二に、「透明性」を求めるだけでは、何を求めているのかが決まらない。透明性の要求は、何のための透明性かを特定したときにはじめて中身をもつ。働きごとに、役立つものは異なる。どの手順を実行したかを残す過程の記録は、再現と不具合の特定に役立つ。証拠の記録は、結果がどれだけ根拠づけられているかの評価に役立つ。対象のモデルは説明に役立つ。適用範囲と不確実性の表示は、どこまで頼ってよいかを判断するのに役立つ。受け手に合わせた提示は、伝えることに役立つ。そのため、システムが生成する理由の説明ひとつで、これらすべてを兼ねることはできない。

説明という働きに限っても、何が役立つかは説によって異なる。サリヴァンの説では、そこで論じられる深層ネットワークのように最上位の水準では不透明でないモデルについて、それが与える理解を主に制限するのは、モデルの内部の不透明さではなく、モデルと現象とを結ぶ証拠の欠如、つまり結びつきの不確実性である。そのため、現象の理解に役立つのは、内部の記録よりも、モデルの外からその結びつきを支える証拠である。そして誤りの害が大きいほど、その証拠にも、モデルから取り出した依存関係がモデルの働き方に忠実であることを確かめる基準にも、高い水準が求められる（第5.2節）。バイスパルトの説では、欠けているのは、ネットワークが何をしているかを、対象となる系の変数や過程によって述べる記述である。この説のもとでは、計算の実行をどれだけ詳しく記録しても、それだけではこの欠けは埋まらない（第5.2節）。フライシャーの説では、分類器とは違う関数を計算する事後的な説明のためのモデルも、そのずれが理想化のしるしを備え、説明される入力に近くで元のモデルとよく一致するなら、分類器がなぜその判定を下したのかについての、開発者、運用者、判定の影響を受ける人々の理解に寄与しうる（第5.4節）。

そのうえ、深層学習モデルの場合、その内部の推論を人間の概念で言い表すことには、別の限界もありうる。第5.4節で見たように、大塚は、深層学習モデルの内部表象を解釈することを、話し手のふるまいだけを手がかりに未知の言語を訳す根元的翻訳になぞらえる。この議論が正しければ、人間の概念による言い表しは、同じようによく当てはまる複数の解釈のひとつにすぎないかもしれない。その場合、言い表しをどれだけ詳しく記録しても、それが正しい翻訳かどうかは決まらない。唯一の正しい翻訳はないかもしれないからである。この限界がどれだけ効くかは、誰が何を理解しようとするかによる。 S が人間であれば、理解は人間の成員が使える形でもたされなければならない。ただし第5.4節で見たように、現象の理解を求めるなら、現象のモデルによる道は、内部の推論の言い表しが決着するのを待つ必要がない。受け手が、モデルの表象をそのまま扱える人工システムであれば、人間の概念への言い表しはそもそも要らないかもしれない。

記録に求められるものは、モデルが理想化されている場合や、目的に相対的に作られている場合にも変わる（第3.4節）。そうしたモデルでは、表象としての正確さと、予測、説明、介入、教育といった特定の用途への適しさが一致しない。そのため、特定の用途のための評価を支える記録は、この二つを区別している必要がある。何を区別して残すかも、説によって具体的になる（第3.4節）。ポトチュニクの説では、表象のなかのそれぞれの指定について、真理からの隔たりが、その役割と研究の目的とに照らして無視できることが求められる。そのため記録には、何を、どの目的のために理想化したのかが要

る。ストレヴンスの説では、理想化を含む説明が正しいのは、何が違いを生み何が生まないかについての主張が真であるときである。そのため記録には、違いを生まないとして省いた要因が要る。ワイスバークの区分を当てはめれば、人間よりはるかに大きな計算能力をもつシステムには、扱いやすさのためのガリレオ的理想化を解いていく理由はあるが、最小限の理想化や多重モデル理想化を解く理由は、その計算能力からは生じない。そのため、計算能力が大きいことは、あらゆる理想化を解いてよい理由にはならない。最小限の理想化を解けば、違いを生む要因だけを残すことで得られていた説明が見えなくなる。多重モデル理想化を一つのモデルにまとめれば、一般性、正確さ、単純さといった望ましきのあいだのトレードオフが見えなくなる。また、第3.4節で見た立場をここまで延ばせば、記録が、証拠がまだ何を決めているかを示すものであるためには、証拠と両立する競合モデルを、いまの課題のためにひとつを選んだ後も記録に残しておくことになるだろう。

第三に、あるシステムが理解しているという主張にどのような証拠がかかわるかは、その主張が、誰が理解するのか（担い手）と、どの理解の説に立つのかを特定してはじめて言える。たとえば次の主張は、それぞれ別の仮説であり、確かめ方も異なる。AIが人間の理解を可能にした（the AI enabled human understanding）、AIが他の人工システムの理解を可能にした（the AI enabled another artificial system to understand）、AIが理解している（the AI understands）、組織された人間-AIシステムが理解している（the organized human-AI system understands）、人工の共同体が理解している（the artificial community understands）。また、主張で名指される担い手は、その理解が利用可能になる社会や、その理解が新しいとされる社会と同じとは限らない（最後の項目を参照）。転移、介入、反事実、適用範囲の認識、異常の診断といった課題での成績は、証拠になりうる。しかし第3.3節で見たように、成績は暗記や検索のような近道によっても得られる。そのため、成績が主張された能力の証拠になるのは、評価が、主張の依拠する説にとって問題となる近道を、その能力から見分けられる範囲に限られる。

どの成績が証拠になり、どの近道が問題になるかは、説ごとに具体的に見ていける（第3.3節、第6節）。ここでは、システム自身が担い手として名指される場合を見る。人間の理解を可能にしたという主張については予測の事例で、組織された全体が理解しているという主張については長期の自律的研究の項目で、説ごとの判定を見る。ヒルズの説では、システムが、他者の与えた説明についていけるか、自分で説明を与えられるか、関連して類似した事例について結論を引き出し説明を与えられるかが試される。ウィルケンフェルドの説では、外から内部状態を書き換えられることなく、システムが自らの表象を修正して新しい有効な推論を可能にできるかが試される。運のよい遂行や、暗記した手順の適用は、この説ではもともと理解から除かれている。パーマンらは、「何か」「なぜか」「もし〜ならどうなるか」という問いにどれだけ幅広く正しく答えられるかによって、理解の程度を試すことを提案する。ムーディとデ・レフトの説では、正確な計算をせずに理論の定性的に特徴的な帰結を認識できるかが問われる。二人は、分布外の幅広い科学的な問いに正確に答えることを間接的な証拠として、現在の言語モデルに、脆いながらも自明でない程度の理解を認める。これらの説では、近道を見分けられる評価を組めば、成績が理解の主張を支える。これに対して、ポトチュニクの説では、理解には表象が引き起こす理解の感じが要るので、成績がどれだけよくても、その感じがあるかという問いが残る。理解には信念ではなく受容で足りるとする説のあいだでも、問われることは異なる。デルセンの説では、理解に要るのは受容と、正しく説明できるほどの主題の習熟であり、信念は要らない。信念を p が真であると感じる傾向と解するなら、感じることを帰属できないシステムにとっては、知識より理解のほうが近いことになる。ただし、説明のなかで p を前提として使うことが、デルセンのいう方針の採用にあたるかは、なお問われる。他方、信念についての機能説・傾向説のもとでは、説明に用いられる受容はそれ自体が信念の証拠に数えられるので、この道は信念と別の道にはならない。エルギンの説では、理解に要る受容は認知的な目的のもとの自発的なものなので、システムが認知的な目的をもち、その意味で受容しているかが問われる（第6節）。

第四に、長期にわたって自律的に研究を進めるシステムを何によって評価するかは、研究をどう特徴づけるかによって変わる。理解を研究の中心に置く特徴づけを採れば、そうした研究はどれも理解によって評価される。何がその理解にあたるのか、そして誰の理解を数えるのか、つまりシステムの理解か、その受け手の理解か、集合体の理解か（第6節）は、採用する理解の説が決める。たとえば、理解を科学やその進歩の中心に置く説（第4節、第7.1節）で研究を特徴づけたとしよう。デルセンのノエティックな説では、進歩は、社会の関連する成員が自らの理解を増すことを可能にする、公的に利用可能な情報にある（第4節、第6節）。この説で評価するなら、システムが自ら依存関係のよいモデルを得ても、それだけでは足りず、その情報が S の成員にとって利用可能になったかが問われる。 S が人間の社会であれば、人間の成員が使える形の情報でなければならない。デ・レフトの説で評価するなら、理解可能な理論に基づく説明を与えたかが問われ、完璧な予測を出すだけの研究は、大きな科学的成功には数えられない（第4節）。これに対して作業的特徴づけのもとでは、研究がどの認知的達成を目指すかは開かれている（第7節）。そのため、長期にわたる自律的研究の成功は、その試みが目指す達成によって評価される。デルセンやデ・レフトの基準が成功の基準として当てはまるのは、理解を目指す試みに限られる。

これとは別に、そうした研究は、自律性のものが求める能力によっても評価できる。探究のより多くをシステムに任せるほど、システム自身が担わなければならない能力は増える（第7章第4.4節）。たとえば、ある規則性がもはや転移しないことを認識できるか。焦点となる主張の失敗を、装置や補助仮定の失敗から区別できるか（第6章第2節、第4.2節）。代わりとなる表象を生成し、比較できるか（第5章第3節）。自らの証拠を批判にさらし、訂正を伝えていけるか（第11章第2節、第8節）。世界から切り離されることなく、結果を評価する基準を改められるか（第6章第5節、第7章第4.3節）。いくつかの説は、これらの能力の一部を理解そのものに結びつける。規則性がもはや転移しないことを認識するとは、ウッドワードの説では、新しい文脈が依存関係の不変性の領域の外にあること、あるいはその定性的な関係しか持ち越せないことを認識することである（第7.2節）。代わりとなる表象を生成することは、ウィルケンフェルドの説で理解を示すとされる、表象の修正に近い（第3.3節）。そのためこうした説は、これらの能力が統合されて実現していることを、理解と解釈しうる。ただし、これらの能力が一つのシステムに集まっているか、複数のエージェントや制度に分散しているかは、それが理解を構成するかどうかとは別の問いである。説によっては、どちらの配置も、理解を構成することなしにこれらの能力を供給しうる（第7.4節）。分散した配置については、説ごとに帰結が異なる。予測の事例で見たバイスパルトの説を延ばせば、人間がシステムを走らせながら推論するとき、理解をもつのは人間とシステムの結合した全体であって、補助を受けない人間ではない（第4節）。集団の理解を成員の理解に基づけて考える説では、その配置が理解をもたらしたと言えるのは、人間であれ人工のものであれ、どれかの成員が理解したときである。バードの集団機能的な知識の説を理解に広げる説では、どの成員も全体を保有していなくても、組織された全体が理解の担い手になりうる（第6節）。

最後に、達成が新しいとされる参照社会 S と、達成の担い手、つまり達成がそのうちに実現する主体とは区別される。両者が一致するのは、 S 自体が理解する、という主張の場合である。この区別があるので、ある成員が新しく理解したこと、集合体が理解したこと、新しい説明が利用可能になったことは、

それぞれ別の主張になる。そして、そのどれが *S* の認識的状況に入ったことにあたるかは、説によって異なる（第6節）。集団の理解を成員の理解に基づけて考える説では、関連する成員が理解したときであり、バードの説を理解に広げる説では、社会の組織が全体としてその理解を実現したときである。ボーゲ、シュスター、シュトルの説とデルセンのノエティックな説では、伝達可能な説明や公的な情報が利用可能になることが問われる。ただし、利用可能になったことがすでに認識的状況に入ったことにあたるのか、社会を理解する位置に置くだけなのかは、知識についての問い（第2章第7節）と同じく、なお残る。研究を行うのが人工システムで、*S* が人間の社会である場合には、人工システムが理解するなら、その理解は、人間の社会自体が理解していなくても、その社会にとって新しいものでありうる。ただし、それが意図された研究の記述として適切かどうかは、*S* を都合よく選ぶことでは決まらず、主張と活動の内容による。また、そうした仕事をその社会のための研究と呼ぶことは、その社会の認識的状況が、たとえば成員がたどれる説明を通じて、求められる種類の理解を含みうることを前提にする（第1章第2節）。*S* が人間の社会であれば、人工システムが把握した依存関係の組織を人間に理解可能な形にできないとき、この前提が問われる（第6節）。これに対して、*S* が混合社会や人工社会であれば、その組織を伝えられる人工の成員を通じて、この前提が満たされることもありうる。ボーゲ、シュスター、シュトルの区別を *S* が人間の場合に当てはめれば、人工システムの理解は、三人のいう科学者の理解にあたるものであって、まだ科学的理解ではない。その説では、科学的理解には、十分に広い科学的聴衆に伝達可能な説明が要るからである。三人の事例でいえば、システムは、ダイソンの証明が他の物理学者のたどれる説明を与える前のファインマンの位置にある（第6節）。ボーゲとデ・レフトが、人間の科学的理解と人工の科学的理解の分断を越える道を、AIが自らの理解を人間の専門家に伝えられるかを試すことに求めたのも、この場合にかかわる。

求められるものは用途によって変わるので、以上の対照は、人間に似たシステムほど進んでいる、という一本の梯子にはならない。たとえば、*S* が人間であり、理解が *S* の認識的状況に入ることが目指される場合には、人間に向けた説明か、説明なしの理解を認める説であれば人間が使える別の理解の乗り物（第3.2節）が欠かせない。しかし、理解がシステム自身の推論にだけ使われる場合や、*S* の人工の成員に伝えられる場合には、要らないかもしれない（第6節）。また、把握の感覚やその他の経験を理解の要件としない説、たとえば理解を能力とみなし、人間とは異なる仕組みによる実現を認めるバーマンらの説や、表象が記号的であることを求めないウィルケンフェルドの説では（第6節、第8章第1.3節）、人工システムの表象は、人間の内観に似ていなくても、本物の非人間的な理解を支えるかもしれない。そうした説のもとでこの可能性がどこまで支持されるかは、その別の実現の仕方が、採用する説の求める関係を対象とのあいだにもっていると、どこまで示せるかによる。対象が世界のなかのものであれば、その関係には、システムの状態がそもそもその対象についてのものであることが含まれる（第8章第2.2節）。そのうえで、世界とのさらなる接触を通じて誤りを正せる状態を保てるかどうかは、別の問いである。それは、第8節で見た継続するシステムの能力と、第8章第2.3節で検討する世界への応答可能性にかかわる。これに対してポトチュニクの説では、理解には理解の感じが要り、何が理解を促すかも人間の認知の特徴に依存する。そのためこの説のもとでは、この可能性は、まず、そうした感じが人工システムにありうるかどうかにかかわる（第6節）。

本章の問い、すなわち理解はいつ、誰において達成されるのかに戻れば、その答えは、どの達成が主張され、誰が担い手として名指され、どの説が採られるかによって定まる。人工システムが研究を行う場合には、このうち「誰において」が独立の問いになる。人間が研究する場合と違って、結果を生んだものがそれを理解する者だとは前提できないからである（第6節）。作業的特徴づけのもとでは、研究がどの認識的達成を目指すかは開かれている。理解を目指す試みにとっては、理解に届くかどうかはその成否にかかわる。理解を目指さない試みにとっても、続いていく研究には、理解と結びつく能力が必要になりうる（第7～8節）。そして、ここで挙げた能力の多く、つまり転移、診断、さらなる問い、批判、記憶は、ひとつの結果の性質ではなく、探究がどう組織されているかに依存する（第8節）。そこで、第4～6章からなる次の部は、達成から探究の組織へ進む。機械が問うとはどういうことか、候補や表象はどう利用可能になるのか、結果は何にかかわり、何を改めるのかを、同じ探究の三つの側から問う。そのうえで第7章は、その組織のどこに、研究の遂行や能力を帰属させてよいかを問う。

第III部 未完の探究はどのように進むのか

第1章の作業的特徴づけは、研究を、参照社会 S にとって新しい認識的達成をもたらそうとする試みとする。第II部の二つの章は、知識と理解を可能な目標として検討し、その成立条件を比べた。しかし、達成の成立条件を述べることは、そこへ至る過程を述べることではない。第III部が問うのは、答えがまだ得られておらず、失敗もありうる探究のなかで、何が行われ、何が変わりうるのかである。ここでいう探究は、研究の試みのなかで進んでいく過程を広く指す。第1章第3節で見た、個人が探究しているとはどういうことかについての説は、同じ語をより限られた意味で用いている。

第2章第8節は、問いの形成、証拠の獲得、信念度の変化、認められた無知などを、候補となる**認識的変容** (epistemic transformations) と呼んだ。特定の変化が進歩を構成するかどうかは、それが価値ある目標、正しい答え、誤りの露呈、探究能力の改善などどのような関係にあるかによるからである。ロスは、価値ある問いの定式化や関連する証拠の獲得もそれ自体で進歩でありうると論じ、ハリスは、それらは進歩を促すにとどまると反論した (第2章第8.2節)。また、データセット、方法、分類のように後の前進にかかわるものについては、前進を構成すること、促すこと、促すと合理的に期待されるだけのことが区別された (第2章第8.3節)。研究過程は、まだ知識をもたらさないうちに、問い、候補、表象、証拠的關係、方法、能力、認識された無知などを変えうる。そうした変化が探究のなかでどう起こり、互いに関係しうるかは、それが進歩を構成するのか、後の進歩を促すだけなのかとは別に問うことができる。第III部が取り組むのは前者の問いである。

第1章第7節は、作業的特徴づけから比較的直接に導かれる含意として、試みが S においてまだ決着していない点に向かうことを取り出し、それを最小限の問いの構造と呼んだが、その点がどの意味での問いなのかは決めずにおいた。しかも、答えが S においてまだ知られていなければ、提案された答えのどれが正しいのかだけでなく、何が答えになりうるのか、問題が何であるのか、どの対象を比べればよいのか、どの表象が問題を扱いやすくするのか、何が関連する証拠となり、どの操作がそれを生みうるのか、結果が何を改めるのかなども、不明確であることがある。実験系の歴史を研究したラインベルガー (Rheinberger) も、未知のものへの近づき方について、これに近いことを述べている (第4章第5節)。とはいえ、答えが新しくても、そこへ至る手続きが前もって実質的に確立されている研究もある (第5章第1節)。

何が答えになりうるかの不明確さが進行中の作業のなかでどう扱われるかは、たとえば次の短い記録に見てとれる。2015年10月、ブラジルのペルナンブコ州の保健局は、母子医療の拠点となる公立と私立の複数の病院から、小頭症をもつ新生児26例の通知を受けた。同月30日付の保健局の技術ノートは、州の出生登録に2015年1～9月の小頭症の記録が22例あることも挙げ、2005～2014年の中央値が9例 (幅は4～16例) であったことと比べて、情報は部分的でも、この先天異常の発生の型の変化をすでに示しているとした。母親の一部は妊娠中の発疹を報告していたが、ノートは、現時点の情報ではこの異常をどの疾患とも関連づけられないとし、原因は先天性の感染症でも、感染によらないほかの要因でもありうると述べた。そのうえで、通知のためにこの時点から用いる、在胎週数に応じた頭囲による症例定義を定め、さまざまな情報源から報告された90人の患者の全員について、その定義に当てはまるかを評価するとした。感染症の検査は、断層撮影で感染を示唆する所見のある症例に限って行うよう勧め、関連しうる要因についての仮説を立てるための調査が進行中であることを告げ、このプロトコルは臨床と疫学の証拠の進展に応じて更新すると述べた (*Secretaria de Saúde do Estado de Pernambuco 2015, 1-3*)。

この数頁の文書のなかに、未決の点、当面の答え、それを確かめる手段が入れ子になって現れている。発生の型は変わったのか、という点には、登録の記録と過去10年の中央値という比較の基準によって、部分的な情報にもとづく当面の答えが与えられた。ただし、そこで数えられたのは、国際疾病分類のコードで出生登録に記された例である。以後の通知で何を症例と数えるかは、それとは別に、頭囲による定義として定められた。この定義によって、報告された90人のうち誰が定義に当てはまるのか、という問いを立てられるようになった。漠然とした「多いのではないか」とは違い、これは何が答えになりうるかの定まった問いである。他方で、この異常が何によるのかは未決のまま残された。ノートは、先天性の感染症か、感染によらない要因かという大まかな答えの候補を挙げるにとどめ、要因についての仮説を立てることは調査に委ねた。それでも、候補は手段の選び方にかかわっていた。感染症の検査を、感染を示唆する所見のある症例に限るという勧めは、感染という候補に照らして、どの症例に何を調べるかを定めている。さらに、90人を評価した結果は、発生の型が変わったという当面の答えだけでなく、症例定義そのものが適切にかにもかかわりうる。プロトコルを更新するという告知は、確かめた結果が、確かめるのに使った定義のほうを改める余地を残していたと読める。

第III部は、研究の過程を記述し比べるために、一つの見取り図を用いる。見取り図は、ある時点の探究の構成を、いくつかの次元とそれらのあいだの関係によって記述する。次元には、何が問われ、何が未解決のまま残っているか、どんな内容が答えの候補とされ、それにどんな態度が向けられているか、そして表象、証拠、方法と装置、基準、実践的な能力、行為者、記録、社会的な仕組みなどがある。研究は、ある次元での変化がほかの次元で可能な変化を変え、再構成された構成がさらに再構成されていくことによって、進む。見取り図は段階の順序を定めるモデルではなく、過程を比べるための語彙である (第5章第1.1節)。それが依拠する事例は、論文、歴史研究、実験ノート、訓練のログ、リポジトリの履歴、口述史などの記録から、論点に合うものを選んだものであり (第5章第1.2節)、第4章の事例も含め、それらが示すのは、ある関係が起こりうることであって、その頻度ではない。

いま報告されているシステムの多くについて、機械が研究しているかという問いは、研究の過程を組織する仕事が機械の側にあるのか、その外にあるのか、という問いになる。仮説の生成や検証のような工程のラベルは、問いがいかにして利用可能になり、候補がいかにして思考可能なものにされ、不一致がいかにか診断されるのかなどを、説明しないまま残す (第5章第1節)。そこで第III部の三つの章はその仕事を記述し、そのうちどれが人工の行為者の内部で実現され、どれが他の行為者や仕組みにわたって分散したまま残るのかを、第7章第2節が問えるようにする。

保健局のノートは、一つの探究を三つの側から同時に記述している。一つは何が未決かの記述であり、異常の原因や、90人のうち誰が定義に当てはまるかがそれにあたる。もう一つは何を当面のものとして定めたかの記述であり、比較の基準、症例定義、大まかな原因の候補、検査を行う症例の範囲がそれにあたる。さらにもう一つは、これから得られる結果が何にかかわりうるかの記述であり、90人の評価の結果と、プロトコルの更新の告知がそれにあたる。しかも、同じ要素がいくつもの側に現れる。頭囲による症例定義は、当面定められたものであると同時に、90人について、何が答えに数えられるかの定まった問いを立てられるようにしたものであり、評価の結果によって改められうるものでもある。この三つの側は、このノートに限らず、進行中の探究を記

述するときに書き分けられる三つの関係、すなわち、探究とそれを組織している未決着の点との関係、答えの候補・表象・手段のあいだの関係、結果とそれがかかわるものとの関係に対応する。以下では、これを探究の三つの関係と呼ぶ。見取り図の次元の要素は、どれか一つの関係に割り当てられるものではなく、同じ要素が別々の関係に立ちうる。三つの関係は探究のなかの関係のすべてを尽くすものではないが、第III部の三つの章は、同じ探究をそれぞれの関係の側から見る。第4章は未決着の点との関係を、機械が問うとはどういうことかという問いとして扱い、第5章は候補、表象、手段がいかにして利用可能になるかを、第6章は結果が何にかかわり、何を改めるのかを問う。

第4章 機械が問うとはどういうことか

トウモロコシの遺伝学者マクリントック (McClintock) は、1950年の論文を、第9染色体の短腕の遺伝子の構成を明らかにするために計画した実験のなかで、トウモロコシではまれにしか起こらない、あるいはまれにしか気づかれない現象が、培養のなかで目立って高い頻度で現れはじめた、と書き起こしている。不安定な遺伝子座 (mutable loci) である。最初に現れたのは、初期の発生のなかで、染色体の切断、融合、架橋を繰り返す周期によって第9染色体の短腕に大きな構造の変化を受けた約450の植物を、自家受粉させた子孫であった。そこでは、およそ40の異なる不安定な遺伝子座が認められた。もとの系統は数年にわたって調べられていながら、これほど多くの不安定な遺伝子座を示してはいなかったので、マクリントックは、この周期の仕組みの一部分か、周期がもたらした構造の変化が、変異が一斉に現れる条件をつくったと結論している (McClintock 1950, 344–346)。1944年の年次報告も、この周期を使って新たに切断された多数の第9染色体を作り、その短腕に突然変異を得る研究を報告している (McClintock 1944, 127–132)。

何が変わったのかを分けて書いておく。当初の実験を組織していたのは、第9染色体の短腕にどのような遺伝子があるのかという問いであった。予期しない変異が現れた後に作業を組織するようになり、1950年の論文の題にもなったのは、不安定な遺伝子座がどのように生じ、どうふるまうのかという問いである。この周期の仕組みかその産物が条件をつくったという結論は、この後の問いへの答えの候補であって、問いそのものではない。予期しない観察が、何を問うかを変えたのである。1949年1月の覚書は、Ds をもつ培養から最初に分離された不安定な遺伝子座 c-m1 について、Ds と呼ばれる遺伝子座が元の位置を離れて C と呼ばれる遺伝子座のところに入り込み、C のはたらきを抑えたことで生じたものとする結論を定式化している (McClintock 1949a, 4, 9; McClintock 1950, 349–350)。これも同じ問いへの答えの候補である。しかも、不安定な遺伝子座の生じ方という問いが作業を組織するようになった後にも、そのもとで出てくるより狭い問いは未決のまま残った。同じ覚書は、穂に現れた無色の大きな区画が、Ac と呼ばれる因子が失われたことによるのか、c-m2 と呼ばれる不安定な変異そのものが安定したことによるのかを決めずにおいている。1949年4月の原稿に後から付された補遺も、分げつと主茎のあいだの Ac の違いについて、分げつが Ac を得たのか主茎が失ったのかを決めずに残している (McClintock 1949a, 49; McClintock 1949b, 33)。Ac が失われたという解釈と、c-m2 が安定したという解釈は、穂の無色の区画が何によるのかという、そのもとで出てきた狭い問いへの答えの候補であって、不安定な遺伝子座の生じ方という問いそのものでも、その問いへの答えでもない。この例から読みとれるのは、二つのことである。一つは、何を問うかが作業の途中で変わりうること、しかもその移り変わりが記録のうえで確かめられることである。もう一つは、問いが移ることと、移った問いのもとで答えの候補が決まらずに残ることとは、別のことだということである。

機械の作業の記録に、同じような移り変わりが見られたとしよう。そのとき「システムは予期しない結果を受けて問いを立て直した」と言うなら、何が主張されているのか。システムが新しい疑問文を出力したことなのか、何かを問う行為をしたことなのか、その問いに好奇心のような態度を向けたことなのか、それとも、その後の作業が新しい未決着の点に向けて組織されるようになったことなのか。人間の研究者について「問いを立て直した」と言うとき、これらは区別されずに語られることが多い。A が機械であるとき、これらは別々に成り立ったり、別々に欠けたりする。しかも、何が立て直しを促したのかは、マクリントックの場合に論文の記述が伝えているように、それとは別に記録の上で確かめることになる。

このうち最後のもの、つまりその後の作業が未決着の点に向けて組織されることは、第1章第7節が作業的特徴づけから取り出した最小限の問いの構造にあたる。同節によれば、この構造は、プロジェクトが明示的な疑問文から始まることも、一人の行為者がそれを抱えていることも、最終的な問いが当初から変わらずにあったことも求めない。また、そこで導かれたのは試みがどこに向かっているかであって、S の実際の状態ではない。そのため、この構造からは、試みが未決着の点に向かうことは導かれるが、疑問文の出力や問う行為や態度がそれに加わる必要があるのかを含めて、それがどの意味での問いなのか、その中身が当初から定まっていたのか、誰がそれを担うのか、誰にとって未知なのかは決まらない。本章が問うのは、そのうえで、機械が問うとはどういうことかである。

本章の第1節から第7節は、この構造が決めずにおく四つの点、すなわち、どの意味での問いか、中身が当初から定まっているか、誰が問いを担うか、誰にとっての未知か、に沿って分かれている。どの意味での問いかは第1節から第3節で、中身が当初から定まっているとは限らないことは、問いの形成と立て直し (第4節) と、問いが答えられないまま立たなくなることや決着すること (第5節) として扱う。誰が問うのかは第6節で、誰にとっての未知かは A の未知と S の未知を分けて第7節で取り上げる。第8節は、機械が問いや研究案を生んだとする実証研究の主張をこれらの区別に照らして読み、第9節は、問題、問い、探究を論じる後年のラベッツ、サイモン、ナーセシアン、フントヴィッツとラベッツ、ラウダン、プロムバーガー、ケルプ、リーの見方と本書の捉え方とで、機械の同じ活動の記述がどう分かれるかを見たうえで、本章全体を振り返る。

1. 「機械が問う」という主張を分ける

機械が問うと言うとき、何が主張されているのか。機械の出力は記録に残って調べられるが、それがどのような行為や態度にあたるのかは、まさに争われ、S が人間の社会であるときには、それが S の成員の追う問いとして取り上げられたかも、別に問われる。

機械が問えるかは、機械による研究をめぐる議論でも語られている。たとえば、いまの訓練が育てているのは既知の問いに答える従順な生徒であって、正しい問いを立てる力ではないと言われ (Wolf n.d.)、正しい問いを立てることは科学で最も難しく、しばらくは予想や仮説は人間の専門家から来るとも言われる (Hassabis and Manyika 2026)。機械とともに研究した物理学者からは、計算できる問題の候補は大量に出せても、何が面白いのかについての機械の判断はまだ信用できず、その判断は人間が担うという報告もある (Schwartz 2026)。こうした指摘の多くは特定の時点のシステムについてのもので、原理的にできないという主張ではない。ハサピスは、直観も極限では AI のシステムでモデル化しようと述べている。これとは別に、数学者たちの宣言には、問題を解くことは概念的な理解のための道具であり代理指標にすぎないとするもの (フィールズ賞受賞者たちによる宣言, Avila et al. 2026) や、クレジットと責任は人間に属し続けるとするもの (Alper et al. 2026) がある。機械が問いの候補を出せるか、どの問いが追う価値があるかを判断できる

か、誰が研究の目的を定め責任を負うかは、別々の主張である。以下の区別は、こうした言い方が、どの意味での問いについての、何の主張なのかを見分ける手がかりにもなる。

第1章第7節で触れたように、「問い」という語は、疑問文にも、その内容にも、それを問う行為にも用いられる。『スタンフォード哲学百科事典』で問いの哲学と意味論を概観したクロス（Cross）とルーロフセン（Roelofsen）によれば、この三つを分けることは言語哲学では通例であり、問う行為の分析は、依頼の一種とするサル（Searle）の説や、話し手が答えを知るようにせよという聞き手への命令とするオクヴィスト（Åqvist）とヒンティッカ（Hintikka）の説などに分かれる（Cross and Roelofsen 2026, §§1.1, 4.1）。疑問文の意味を、それを解決する情報によって捉える問いの意味論（inquisitive semantics）を擁護するチャルデッリ（Ciardelli）は、平叙文の内容を主張から区別するのと同じ理由から、疑問文の内容を問う行為から区別し、この内容を、決着を待つ論点という意味でイシュー（issue）と呼ぶ。チャルデッリによれば、同じイシューが、問うこと、声に出して不思議がること、「誰がやったかは脇に置こう」のように議題から外すことなど、別々の行為の内容になりうるし、何かを調べるとき調べられているのは、言語行為ではなく思考がつかむ内容である（Ciardelli 2022, 24–31）。第1章第7節で見た、問う行為についてのワトソン（Watson）の説（Watson 2015, 86–102）とハブグッド＝クート（Habgood-Coote）の説（Habgood-Coote 2026, 1, 8, 11）が、問う者に何を求め、何を求めないかは、第6節で見る。

内容と行為のほかに、「問い」の語で語られるものがある。一つは、問いに向けられた心の態度である。探究と疑問の態度を論じたジェーン・フリードマン（Jane Friedman）は、問いを疑問文の意味内容とみる標準的な考えを心の内容にまで広げ、不思議に思うことや好奇心のような疑問の態度は、問いを内容とする態度だと論じる（Friedman 2013）。二つ目は、探究を組織している未決着の点であり、第1章第7節の最小限の問いの構造はこれにあたる。それは、試みが *S* にとってまだ決着していないものとして向かう点であり、*S* において実際に未決であるとは限らず（本章の冒頭）、まだ疑問文にも、行為にも、一人の心の態度にもなっていないことがありうる。三つ目は議題、すなわち会話や集団が追うものとして取り上げた問いの並びである。これらの意味は網羅的なものではなく、以下の区別に要る範囲で挙げたものである。

これらを分けると、「機械が問うた」という主張に要る前提は、どの意味についての主張かによって違ってくる。疑問文を出力したというだけなら、それは形式についての主張である。ワトソンは、疑問文であることは問いの必要条件でも十分条件でもないと論じる（Watson 2015, 95, 102）。ワトソンは人工システムを論じていないが、この説を当てはめれば、機械が出力した疑問文が情報を引き出す行為にあたらないことも、逆に、検索語の送信やデータベースへの照会のように疑問文を含まない操作が問いにあたることもありうる。そのため、問いの生成のような工程のラベル（第5章第1節）は、それだけでは、上のどの意味を指すのかを決めていない。出力から行為へ進むには、ワトソンの説を当てはめるなら、機械の操作がどの活動のなかで情報を引き出す働きをしているかの説明が要る。その働きを設計者や利用者の目的によって定めるのか、機械自身を含む活動の組織や、それが加わる実践の目標と規範によって定めるのかは、機能についてのさらなる説明による（第1章第3節、本章第6節）。どちらの道をとっても、争点は、機械に好奇心があるかではなく、情報を引き出す働きを、どの活動のなかで、どの観点から帰属するかに移る。チャルデッリの区別を持ち込めば、出力の形だけからは、それが問う行為なのか、声に出して不思議がることに近いのか、議題を整えることなのかも決まらず、問う行為を帰属するには、たとえばその出力が *S* との会話のなかでどのようなコミットメントとして扱われるかの説明が別に求められる。態度についても同じずれがありうる。ジェーン・フリードマンの説では、自分の無知についての高階の表象、つまり自分が何かを知らないということを表すことは、疑問の態度にとって必要でも十分でもない（Friedman 2013）。これを当てはめれば、機械が「*X* はまだ分かっていない」と出力することは、それだけでは疑問の態度の証拠にならず、逆に、自分の知識状態を表していないことだけで、その機械を、態度の担い手でありうるものから外すこともできない。

探究を組織している未決着の点という意味では、事情がさらに違う。第1章第3節で見たように、本書の試みは、個人の意図だけでなく、集団の計画、人工物の組織された機能、構成要素の役割などによっても基礎づけられる。そのため、機械の活動が未決着の点に向けて組織されていることは、機械が問う行為をしたことにも、疑問の態度をもつことにも依存しない。逆に、機械が疑問文を大量に出力しても、それらがどの作業も組織していなければ、この意味での問いはない。また、問いを産み出す者と、それを受け取って追う者とは、一致する必要がない。機械が出力した疑問文を人間のチームが追うこともあれば、人間が与えた問いを機械が追うこともある。機械が出力した問いが *S* の探究を組織するには、それが *S* の成員に取り上げられ、議題に載ることが要る。*S* が機械の成員を含むなら、機械どうしのやりとりを問いの提案や情報を引き出す働きとみなせるかは、問う行為の説による（第6節）。

機械の出力は「よい問いを立てた」と評価されることもあるが（第8節）、問いの「よさ」にも別々のことが語られうる。以下の論者の区別は、問いがよいための必要条件の一覧ではなく、ある評価がどの点についてのものかを分けるためのものである。形而上学を、ある人々の思考が置いてきた絶対的前提を調べる歴史的な研究とみたコリングウッド（Collingwood）は、問いが生じること、すなわちその前提が実際に置かれていることと、その前提が真であることとを分けた（Collingwood 1940, 26, 28）。コリングウッドの意味では、偽の前提をもつ問いも生じうる（第5節）。ラベッツ（Ravetz）は2016年のインタビューで、研究の優先順位を、ある種類の探究を追わないことで残る無知と結びつけている（Ravetz 2016, second answer）。ワトソンは、問いかけの価値と問いかけのよさを分け、よさには、ともかく成功するという薄い意味と、価値ある情報を有能に引き出すという厚い意味とがあり、有能さは成功より狭くも広くもあると論じた。運よく当たった曖昧な問いかけは有能ではなく、答えが得られなくても有能に問うことはありうる（Watson 2015, 135–140）。ワトソンが評価するのは、問いかけという遂行とそれを支える問い手の技能であり、何を問うかの選択もその技能の一部として評価されるが（Watson 2015, 144）、問いの内容が意味論のうえでどう定まっているかは評価の対象ではない。答えの条件が意味論のうえでどう定まっているか（第2節）、実践のうえで扱いやすいか（第3節）、候補を追う価値や資源を向ける理由があるか（第5章第4.2節、第10章第6節）も、これらとは別の評価である。

A が機械であるとき、これらの区別は、「システムはよい問いを立てた」という評価が何を確かめたのかを問う手がかりになる。扱いやすい問いでも、*S* にとって資源を向ける理由があるとは限らない。機械が出した問いの前提が *S* において実際に置かれていても、その前提が偽であることはありうる。ワトソンの有能さと成功の区別を当てはめれば、機械が出した問いがたまたま価値ある情報に導いたことは、その機械が有能に問うたことの十分な理由にならない。これは、第1章第3節で見た、幸運な出力がその源を研究者と呼ぶ理由をほとんど与えないという論点が、問いの側に現れた形である。有能に問うことを帰属するには、能力の帰属と同じく、条件が変わっても信頼できることや、証拠の変化に応じることのような条件がかかわる（第7章第3.1節）。また、研究の優先順位を誰が決めてよいかは、機械がそれに関連する事柄を見積もれるかとは別の問いであり、第10章第6節と第12章がそれを扱う。本節の初めに見た言い方も同じように読むことができ、それぞれ別のことにかかわっている。機械とともに研究した物理学者の報告は、問題の候補を出すことと、ど

れが面白いかを判断することとを分けている。正しい問いを立てる力をめぐる言い方は、何を問うかの選択にかかわり、ワトソンはその選択を、有能な問いかけを支える技能の一部として評価する。そして、クレジットと責任は人間に属し続けるとする宣言は、目的に誰が責任を負うにかかわり、それは候補を出せるかとも、何を問うかを選べるかとも別に決まる。

ここまでの区別は、問いが何であり、どう評価されるかにかかわっていた。これとは別の水準に、作業的特徴づけそのものについての主張がある。第1章第7節で見たように、最小限の問いの構造は、問いがどこから来るのかを定めない。問いは、実験や読書のさなかにも、同僚との会話や食事の席でも浮かびうる。いま依拠している前提に立ち返って問うこともある。手順に沿って作業を進めているうちに、そもそもなぜこの手順でやっているのかと問い直すことや、改良を重ねているうちに、そもそもこのやり方は目的にかなっているのか、何を目指していたのかと問い直すことなどである。こうした問いは、それまで前提として働いていたものを問いの対象に変える。この点は、第5節で見るコリングウッドの、使っているあいだは前提として働き、点検するときには問いの答えの側に回る相対的前提とも重なる。当たり前だと思っていたことを、空の雲を見て、そもそもなぜ雲は白いのかと問うように、改めて問うこともある。また、問いは、指導者、共同研究者、資金提供者、共同体、装置、分野の歴史などから受け継がれることもある。そのため、研究者が問いを自ら発案していないことだけでは、本書の作業的特徴づけの意味で研究を遂行していることは否定されない。個人、チーム、人工システムは、別の行為者から与えられた問いや、共同体が抱える問いの一部について、研究を遂行しうる。問いと無知の論理を論じたブロムバーガー (Bromberger) も、科学的な寄与を無知の項目の解消として評価する見方のもとでは、寄与は、その無知を自覚した者がそれに応じて求めた結果である必要はない、としている (Bromberger 1992, 116–117)。

出どころを定めないことは、出どころが重要でないということではない。研究がどの社会的位置から出発するかが得られる知識にかかわるとする立脚点理論に立つなら、どの共同体の記録から問いが受け継がれるかは、研究がもたらす知識がどれほど一面的になるかにかかわる。第10章第6節は、立脚点理論と、ある社会の研究が重要性の基準を他所から受け取りうると論じたバルサフスキーの議論を手がかりに、記録がすでに立てている問いを受け継ぐ人工の構成、つまり研究が帰属される人工の単位が、その記録を生み出した者たちの立っていた場所から出発しうることを検討する。また、未解決性に気づき、問いを定式化したり選んだりし、その枠づけを修正する能力は、行為者に帰属できるものの範囲を広げ、その自律性を高めうるが (第7章第4.3節)、あらゆる研究者や研究参加者にとつての必要条件として作業的特徴づけに含まれているわけではない。問いが探究のなかで実際にどう形づくられ、立て直されるかは、問いの出どころを定めないというこの主張とは別の水準の話であり、研究の記録と探究の説が描く過程の記述として、第4節で見る。

2. 何が問われ、何が答えになるのか

第1節で問いの内容と呼んだものは、何が答えになるかと切り離せない。本節で扱うのは、この意味での問い、つまり未決着の点が、何が答えになるかの条件をもつ問いの形をとっている場合である。機械が出力する問いや、評価に使われる問いは、たいていこの形をしている。フレーゲ以来の言語哲学は命題と平叙文を重んじてきた。問いの意味論の多くもその枠組みのなかにあり、問いを、その答えにあたる命題の集合や関数と同一視してきた。問いと命題のどちらも基本としない論理を立てる試みもある (Cross and Roelofsen 2026, §§1, 2.1.1)。何を答えとするかによつて、問いを埋め込んだ文が何を述べ、そこから何が従うかも変わる。たとえば、ジョンが何を読むかを知っていれば何を読まないかも知っていることになるかは、埋め込まれた問いが真で網羅的な答えにあたる命題を指すとするグルーネンダイク (Groenendijk) とストックホフ (Stokhof) の説では肯定され、真の答えの集合を指すとするカルトゥネン (Karttunen) の説では否定される (Cross and Roelofsen 2026, §4.2)。一方、人工システムについては、「システムは問いに答えた」という主張がベンチマークや評価のなかで日常的に数えられる。そのとき何が答えに数えられているのか。

クロスとルーロフセンは、答えには前理論的に一義的な概念がなく、理論ごとに異なりうる専門用語であり、何を答えとするかを明示しない理論もある、とする。多くの論者が採る、問いにちょうど完全に答えるもの (偽であってもよい) だけを直接的な答えとする見方から、主題についての情報なら何でも答えに数える説まで、幅があり、答えが命題でなければならないかについても見方が分かれる。前提を否定する応答は訂正の答え (corrective answer) と呼ばれ、多くの論者はそれを直接的な答えに数えない。同じ疑問文にも、答えを一つ挙げれば足りる読みと、すべて挙げなければならない読みとがある (Cross and Roelofsen 2026, §§1.1, 4.3, 4.4)。同じ語が論者によって別の概念を指すこともある。推論的問い論理を展開したヴィシニェフスキ (Wiśniewski) は、表現のうえで問いにちょうど応じる文を直接的な答えとして各問いに割り当て (Wiśniewski 2022, 108 n. 4, 109)、そのうえで、部分的な答え、すなわち真の直接的な答えが直接的な答えのある真部分集合のなかにあると述べる答えと、除外的な答え、すなわち直接的な答えの少なくとも一つを除外する答えとを、別の概念とする。直接的な答えが互いに排他的とは限らず、そのどれかが真であるとも限らないとき、両者は分かれる (Wiśniewski 2013, 43–45)。これに対して、会話を協同的な探究とみなすスタルネイカー–ロバーツのモデルに依拠するハブグッド＝クートは、ロバーツ (Roberts) に従って、問いを、論理空間のうち前提を満たす部分を互いに排他的な可能な答えへと分けた分割として扱い、可能な答えの少なくとも一つを除外する答えを部分的な答えと呼ぶ。下位の問いに完全に答えれば、元の問いには部分的に答えたことになる (Habgood-Coote 2022, 1107–1111 and n. 28)。これは、ヴィシニェフスキの用語でいえば、おおむね除外的な答えにあたる。チャルデッリは、背景の情報のもとで、下位の問いの集まりが元の問いを論理的に決める関係を依存関係と呼び、それを談話についての事実ではなく意味についての事実だとする。依存関係が成り立てば、下位の問いにどう答えても、その答えは背景と合わせて元の問いに答える (Ciardelli 2022, 23–24, 29, 41–42)。

これらの区別をベンチマークの判定に持ち込むと、「出力は問いに答えた」という主張の真偽は答えのどの概念を採るかによって変わり、「答えた」と数えることは、暗に一つの答えの理論を採ることになる。前提を退ける訂正の答えが直接的な答えに数えられなくても、それによつて探究を終えられるか、問いが立たなくなるかは、論者ごとの説明による (第5節)。「答えた」かどうかを数える判定と、研究で何が決着したかとは、ここで分かれる。また、ヴィシニェフスキによれば、形式的な問いは、自然言語の問いを一つの解釈のもとで表すにすぎない (Wiśniewski 2013, 15)。そのため、同じ出力が、ある形式的な表現のもとでは直接的な答えでありながら、*S* が問うている問いの答えではないことがありうる。どの読みが問われているかが *S* の側で定まらないかぎり、出力が答えにあたるかの判定は始まらない。*S* が機械の成員を含んでいても、同じ形式言語を使うことだけで成員の解釈が揃うわけではない。さらに、ハブグッド＝クートは、利用可能な情報と合わせれば可能な答えの少なくとも一つを除外するものを、より弱い意味での部分的な答え (casual partial answer) として定義する (Habgood-Coote 2022, 1109–1110 and n. 34)。この定義を当てはめると、同じ出力が、あるデータベースに接するシ

システムにとっては部分的な答えでありながら、その情報に接しない S の成員にとってはそうでないことがありうる。答えであるかどうかは、誰に何が利用可能に相対的になるのである。

チャルデッリの依存関係を、課題を下位の課題に分けて答えを集めるシステムに持ち込むと、集めた答えが元の問いに答えるかどうかは、下位の課題を誰がどう実行したかは独立に、分け方の段階で問える。分け方が依存関係を実現していなければ、下位の課題がすべて首尾よく答えられても、元の問いが決まる保証はない。また、下位の課題を受け持つ各部分が候補を退けたことは、ヴィシニェフスキの意味で元の問いの答えを絞り込んだことと同じではない。除外が絞り込みになるのは、真の答えが挙げられた選択肢のなかにあるという前提を加えたときであり、その前提は、候補の集合が不完全でありうる（第3節）、まさに問われうるものだからである。ただし、依存関係を分け方の段階で問えるのは、可能性の空間が固定されているかぎりでのことである。研究の途中では、その空間そのものが改められることがあり（第4節）、問いが前提とのかかわりで立たなくなることもある（第5節）。

3. 問いを理解すること、問いを扱えるものにすること

第1章第7節の最小限の問いの構造は、試みが S においてまだ決着していない点に向かうことを言うが、その点が疑問文で表されていることも、何が答えに数えられるかの条件がすでに定まっていることも求めない。第2節で見た説明は、多くの場合、答えの条件がすでに定まっている問いを扱っていた。本節が取り上げるのは、この構造が許しながら第2節が扱わなかった場合、つまり、答えの条件がまだ定まっていない未決着の点と、それを答えの条件の定まった問いへと立て直す、問いの側の仕事である。また、ある点が未決着であるからといって、可能な答えの集合がすでに手元にあるとは限らない。では、可能な答えが手元にないとき、問いを理解しているとはどういうことか。機械について、この問いと立て直しの仕事が重みをもつ理由は二つある。一つは、機械にとって疑問文を出力する費用が低くありうることであり、もう一つは、機械が多くの答えの候補を生成しうることであり、そのため、候補を出せることが問いを理解していることにあたるのか、答えの条件を定める仕事が候補の生成とは別にあるのかによって、何を機械に帰属させるかが分かれる。

問いを理解することと可能な答えとの関係については、論者の立場が分かれている。問いの論理には、何が答えに数えられるかを知ることは問いを知ることだ、というハンブリン（Hamblin）に由来する考えがあり、ヴィシニェフスキはこれを、自然言語の問いを知るとはそれを表す式を知ることであり、式を知るとはその可能な答えを知ることだ、という二段に分けて保つ。ある種のなぜの問いやいかにしての問いについては前もって決まった答えの集合を語ることが意味をなさない、という異論を想定し、そうした自然言語の問いがあることは争わずに、自然言語の問いの可能な答えと、それを表す式の可能な答えとを区別して応じる（Wiśniewski 2013, 16）。チャルデッリは、標語として、問いを理解するとはそれを解決するのにどのような情報が要るかを知ることであると述べ、問いの内容を、当該のモデルにおいてそれを解決する情報状態の集まりと同一視する（Ciardelli 2022, 36）。これに対してブロムバーガーは、私たちは可能な答えを述べる立場にない多くの問いを理解しているとして、問いを可能な答えの網羅的な集合と同一視する多くの問いの論理学者の考えを退けた。ブロムバーガーによれば、問いを理解することは、可能な答えのそれぞれの真理条件を知ることではありえない。それぞれの答えがその問いに対して答えにあたる関係に立っていることも知らなければならないからである。また、偽と分かっている答えしか思いつけないのに問いは健全だと考える理由がある、という状況にあると知るには、候補の集合を広げられなくても、それが不完全だと判定できなければならない（Bromberger 1992, 4, 130–131, 133）。

これらの立場のあいだに位置する論者もいる。ニクルズ（Nickles）は、ハンブリンの洞察の「精神」には従いながら、問題を、解そのものではなく、解の範囲を定める制約によって定義する（Nickles 1981, 109–111）。制約には理論的な背景、方法上の条件、暗黙の条件が含まれ、それらは探究とともに見いだされていく（Nickles 1981, 111–112）。制約を多く、鋭く知るほど問題はよく理解され、制約は、解を見たときにそれと分かるようにするだけでなく、どこを探せばよいかも示す（Nickles 1981, 88–89, 113）。答えの条件によって問いを同定する点でニクルズはチャルデッリに近いが、その条件が暗黙の条件まで含み、探究とともに見いだされていくとする点で違う。

これらの立場は、問いをもつことと可能な答えをもつことの単純な対立には取まらない。チャルデッリとブロムバーガーを調停する一つの仕方は、チャルデッリの「知る」を、探究者がもつ能力としてではなく、問いの意味内容の特定として読むことである。この読みでは、意味論的なモデルは、探究者が答えを列挙できることも、その一つを見いだす方法をもつことも要求せずに、解決の条件を特定しうる。そうであれば、候補を作れないことだけでは問いが意味内容を欠くとは示されず、ブロムバーガーの事例が見えるようにするのは、何を問われているかを理解していても答えるための概念や手続きが利用できない、という実践上の隔たりである。ただし、ブロムバーガー自身の論点は問いの理解についての理論上の主張なので、この調停は、チャルデッリの「知る」をそう読むことを条件としてのみ成り立つ。問いの内容を表すことと、探究者がその問いについてどう前進するかを説明することは別の仕事であり、チャルデッリの標語は問いの理解を前者の側に置くものとして読み、ブロムバーガーは、答えが問いに対して答えにあたることを探究者が把握していることとして、問いの理解を後者の側に置く。

ブロムバーガーは、なぜの問いについて、さらに別の隔たりを論じた。ブロムバーガーの分析では、「何」や「どこ」の問いなら、問いを理解するだけで、答えがあるために何が成り立たねばならないかが分かる。ところが、なぜの問いの形から分かるのは、問われている事柄が成り立っていないからではないということだけである。なぜの問いに正しい答えがあるには、その事実が適切な述語や属性などのもとで捉えられなければならないが、問いの形はそれらへの手がかりを与えず、そうした概念は、それを必要とするなぜの問いが立てられたずっと後に考え出されることもある。こうして人は、答えを知らないと分かっているが、そもそも知るべきものがあるかを知るために何を見いだせばよいのかは分からない、という窮地に置かれる（Bromberger 1992, 157–168）。この窮地は、次の段落で見る p 窮地とは明示的に区別され、なぜの問いに固有のものでもない。「何であるか」の問いも、状況によっては同じ窮地を生む（Bromberger 1992, 168–169 n. 46）。この窮地への合理的な対処として、ブロムバーガーは、なぜの問いのことはいったん忘れて、信頼できる資格をもつほかの問いに向かうことを挙げ、なぜの問いへの答えはふつうそうした問いの研究から現れる、と推測する。そのうえで、この推測はなぜの問いへの答えを貶めるものではなく、なぜの問いの価値は疑いえない、と付け加えている（Bromberger 1992, 169）。本節の区別で読むなら、この窮地で欠けているのは、答えの候補というより何を見いだせばよいかの手がかりであり、答えの条件の分かるほかの問いに向かう対処は、答えの候補を出すこととは別の、問いの側の仕事である。

これと区別されるのが、第2章第8.2節で見た p 窮地である。ヴィルホルトが深い無知と同一視したもので、ある主体がある問いについて p 窮地にあるのは、その主体の見解のもとでは問いに正しい答えがあるのに、同じ見解のもとで決定的な反論のない答えを一つも思いつけないときであった (Wilholt 2020, 9–10)。 p 窮地で欠けているのは決定的な反論を免れる答えの候補であり、候補を一つも思いつけないことも、思いつく候補がすべて退けられていることもある (Bromberger 1992, 4, 29)。問いをほとんど理解していなければ候補はいくらでも作れるので、 p 窮地にはない (Bromberger 1992, 31)。ジェーン・フリードマンも、可能な答えを一つも表せなくても、問いを把握し、疑問の態度をもちうると論じる (Friedman 2013)。問いを把握することと候補をもつことは、別々に成り立ちうるのである。問いの制約と両立する候補がいかにして利用可能になるかは、第5章で扱う。

問いを扱えるものにする仕事の一つの典型は、なぜの問いに対比を指定することである。ファン・フラッセン (van Fraassen) は、なぜの問いを、話題、対比類、関連関係の三つ組として分析した。クロスとルーロフセンが示すところでは、「なぜ P なのか」は、 P と並べて比べられる命題の集まり (対比類) を定めるまでは特定の問いとして同定されておらず、何が説明として関連するかを定める関係 (関連関係) は文脈によって定まる (Cross and Roelofsen 2026, §3.2)。ウッドワードとロス (Ross) の解説では、文脈はこの二つとともに定めるのにかかわり、関連関係は、文脈と、問う者が得たいと思っている情報とに依存する (Woodward and Ross 2025, §6.2)。リプトン (Lipton) の説では、あることが、指定された別の可能性 (対照項) ではなくなぜ起きたのかを説明するには、両者のあいだの因果的な差異を挙げることが要る。起きたことにとって十分な原因を挙げる必要はないが、その代わりに、答えは対照項についても何かを言わなければならない (Lipton 1993, 43–44, 52)。リプトンは梅毒と進行麻痺の例を取り上げる。梅毒にかかった人の大半は進行麻痺にならないので、ある人の梅毒がその人の進行麻痺を説明するかどうかについて哲学者のあいだで争いがあり、リプトンの見方では、対比の問いがその争いを決着させる。ジョーンズが梅毒にかかっていなかったなら、スミスの梅毒は、ジョーンズではなくスミスがなぜ進行麻痺になったのかを説明するが、ドウ (Doe) も梅毒にかかっていたなら、ドウではなくスミスがなぜそうだったのかは説明しない (Lipton 1998, 13)。ただしリプトン自身、因果的な差異を挙げて対比を説明するというモデルは、このままでは単純すぎるとする。より十分な説明は、たとえば、事実と対照項のあいだに因果的な差異が複数あるとき説明に何が要るかを言わなければならない (Lipton 1993, 44–45)、特定の問いに答えるために原因がどんな条件を満たさなければならないかを、リプトンは因果的な選択の問題と呼ぶ (Lipton 1998, 11)。また、クロスとルーロフセンがまとめるところでは、リプトンにとって、対比の問いが意味をなすための中心の要件は、事実と対照項が似通った因果的な来歴をもち、それを背景に差異が際立つことである (Cross and Roelofsen 2026, §3.2.3)。こうして対照項を指定すると、答えが何をしなければならぬか、どこを探せばよいかは絞られる。

ただし、対比の指定がどこまで問いの側の立て直しにあたるかには限定がある。リプトンは、なぜの問いを対比的にすることの効用の一つは、結果が起きなかったなら事態がそれ以前にどう違っていたかという問いの不確定さを減らすことにあり、対照項によって、結果の来歴のどの部分を固定するかを説明上の関心に合うように選べるとする (Lipton 1993, 49, 51)。リプトンによれば、多くのなぜの問いは対比的であり (Lipton 1993, 39)、その対比はしばしば暗黙のままである (Lipton 2001, 16)。暗黙の対照項がすでに決まっていれば、それを明示することは同じ問いの言い直しでありうる。別の問いになるのは、対照項が決まっていなかったときか、別の対照項に替えたときである。また、対比が定まっても、どの原因を挙げるかがそれだけで決まるわけではない。リプトンが指摘するように、実際の場面では、事実と対照項のあいだの因果的な差異がただ一つであることはまずなく (Lipton 1993, 45)、ある出来事の原因のすべてが、与えられた文脈でその出来事を説明するわけでもない。そのためリプトンは、説明が求められる文脈をより細かく捉える分析が要るとする (Lipton 2001, 10–11)。同じ対比について差異の条件を満たす原因が複数あるとき、そのどれを挙げるかは、問う者が何に関心をもつかによって変わりうる。そのため、何に関心があるかによって説明が変わるという事情の一部は、対照項を固定した後にも、答えの側に残る。他方でプロムバーガーは、ファン・フラッセンの例を用いて、強勢によって答えの条件が変わることから、なぜの問いがとくに文脈に依存すると考えるのは誤りだと論じ、強勢のある問いを範型とする理論は偏った標本から一般化している、と批判した (Bromberger 1992, 160 n. 34)。プロムバーガーの分析では、強勢のある文の読みのうち前提を変えるのは一つだけであり、ほかの読みでは、強勢は話し手の関心を示すにすぎない (Bromberger 1992, 164–165)。前提を変える読みは論理形式の違う問いなので、プロムバーガーの分析に立っても、対照項を定めることは、元のなぜの問いを別の問いに替えることにあたりうる。リプトンは、対比の説明を論じたその箇所でもプロムバーガーに依拠しているわけではないが、二人の分析を合わせると、答えの条件の定まらない問いから定まった問いへの立て直しを、候補の生成には還元されない、問いの側の仕事として記述できる。ニクルズの言葉でいえば、それは解の制約を加え、鋭くすることである。

問いの種類が違えば、扱えるものにする仕方も違う。たとえば、いかにしての問い (how の問い) は能力や制約へと分解されることがあり、数学の問題は、幾何学的、代数的、確率論的、あるいは計算論的に表し直されると変わることがある。工学上の野心も、性能が測定可能な関係と、それを実現できる物理的な配置へと翻訳されたときにはじめて、研究可能な問題になることがある。ニクルズも、問題はデータの提示を変えずに定式化し直し、別の問題へと変形できるとし、黒体放射の問題の例を挙げる (Nickles 1981, 88, 114 n. 6)。データの組み直しや証明の表象の組み直しによって、候補や探究の対象そのものが利用可能になる場合は、第5章第3節で見る。

可能な答えの範囲そのものが分からないこともある。ハブグッド＝クートは、プロムバーガーとヴィルホルトを引きながら、ヴィルホルトが p 窮地と同一視した深い無知の語を、問いの可能な答えが何であるかが分からない状態に用いる。この用法での深い無知は、 p 窮地と同じではない。ハブグッド＝クートは、たとえば正しい基礎理論は何かという問い Q について、「 Q の可能な答えは何か」という、可能な答えを問う問い (option question) を立てる (Habgood-Coote 2022, 1116–1117)。ハブグッド＝クートが扱うのは完全な深い無知だけであり、一部の選択肢についての無知のような中間の場合は脇に置かれている (Habgood-Coote 2022, 1117 n. 41)。

ファン・フラッセンの分析を機械の出力に当てはめると、なぜの疑問文を出力する費用が低くありうるとしても、それだけでは、文脈と、問う者の関心とに依存する対比類も関連関係も定まらない。ソーヤー (Sarah Sawyer) の報告によれば、ファン・フラッセンは、全知の存在が説明をもつと理解するのはそもそも正しくないと考え理由をいくらか与えている (Sawyer 2014, 7 n. 8)。もしそうであれば、多くの情報をもつことは、それだけでは機械に説明を与えない。この分析が求める対比類と関連関係は、システムが保有するものの量からは供給されないからである。 S が人間の社会であり、関連関係が依存する関心が S の側にあるなら、機械がその問いを追っていても、問いは S の関心によって形を与えられており、誰の問いかという問いは、疑問文を出したのは誰か、対照項と関連関係を定めたのは誰か、という二つに分かれる。リプトンの説を当てはめると、意味のある対照項を指定するには、事実と対照項の来歴についての知識と、誰の関心に応えるかの選択が要る。混合の S では成員によって対照項が違うるので、機械の成員と人間の成員の説明の食い

違いの一部は、答えの対立ではなく、対照項の違いとして記述できるかもしれない。また、問いを扱えるものにするとは答えの条件を定めることであって、問いを易しくすることではない。しかも、対照項を定めた後にも、どの原因を挙げるかには関心が残るので、機械が対照項を正しく受け取っても、説明の選び方には、なお誰の関心かが入ってくる。機械だけからなる S については、同じ分析が一つの空いた位置を残す。上で見たウッドワードとロスの提示では、関連関係は問う者がどの情報を得たいかに依存するので、これを機械だけの S に当てはめると、問う者の位置に何が立つのかが定まらないまま残る。たとえば、成員の答えが何に渡されるのか、成員に課された課題、成員どうして定める何かなどが候補になるが、そのどれがそうした S の関連関係を定めるのかは、この分析からは決まらない。

問いを把握することと候補をもつことが別々に成り立ちうるというジェーン・フリードマンの論点を機械に向けると、機械には、問いを把握していても答えの候補を表せない状態と、候補は多数生成できても問いを未決として保っていない状態とが、別々にありうる。候補が多いことは、それだけでは前進を意味しない。プロムバガーの定義で問われるのは、候補の数ではなく、候補がその主体の見解のもとで決定的な反論を免れているかだからである (Bromberger 1992, 29, 31)。ただしプロムバガーには、すでにいくつかの両立的な答えを思いついていても、まだ思いつけない両立的な答えがあれば p 窮地にあるとする、拡張された定義もあり、そこでは、制約と両立する新しい候補は、数が多くても窮地の解消に数えられる (Bromberger 1992, 122)。偽と分かっている候補しか出せないのに問いは健全だと考える理由がある機械が、自分の候補の集合が不完全であることをどう判定するか (Bromberger 1992, 133) も、そこから問いとして立てられる。また、ハブグッド＝クートの区別を当てはめれば、与えられた可能な答えの集合のなかでだけ働くシステムは、可能な答えを問う問いを追ってはいない。ただし、一部の選択肢を加えるシステムは、ハブグッド＝クートが脇に置いた中間の場合に入るので、それがその問いを追っているかどうかは、彼の扱いからは決まらない。サイモン (Simon) は、プログラムに与えられる問題と、それを定式化する仕事とを分ける一方で、定式化された問題を解く仕事を軽視する見方も退けている (Simon 1973, 186–187)。ニクルズは、方法や装置に実質的な理論内容が組み込まれると指摘する (Nickles 2019, 363)。ここで取り上げるシステムに当てはめれば、何を帰属させるかは、記録に示された課題の境界による。人間が枠づけを与え、機械がそれを改めずに使った場合、実行の成功は、その枠づけの仕事を機械がしたことを立証しない。結果に至るまでに機械がどれだけ創造的な仕事をしたかは、別に問われる。

この節の区別を合わせると、機械が問いを出したと言っても、その問いを研究できる形にする仕事、つまり何を対照にとるか、どのような説明を答えとみなすか、どの候補の範囲で答えを探すかを決める仕事を誰が担ったかは、別に問われることになる。第8節で読む実証研究の多くでは、上位の問いは、課題とともに人間から与えられている。

4. 問いは探究のなかでどう形づくられ、立て直されるのか

機械について「システムは異常を問いに変えた」「問いを立て直した」と言われるとき、その主張は、どの選択を誰が行い、どの枠組みのもとで何が変わったのかとして確かめられる。そのために、第1節の末尾で分けた二つの水準のうち記述の側に移り、研究の記録や探究の説で目立つ、問いの形成と立て直しを、何がどう変わるのかに注目して見る。第3節の対照項の指定も、その一つの形である。

状況はまず不確定でありうる。概念、対比、道具、候補となる答えが、それを問いうるものにする手助けとなりうる。問いは、探究を通じて、洗練され、組み立てられ、あるいは組み替えられ (Hildebrand 2024, §4.3; Henne 2023, §6)、下位の問いへと分けられて参加者のあいだで扱われることもある (第6節)。ヒルデブランドが伝えるデューイ (Dewey) の探究の理論では、問題は探究に先立って与えられるのではなく、初めは不確定な状況のなかで定式化される。また、仮説について推論することで、問題の記述に立ち戻ることもある (Hildebrand 2024, §4.3)。デューイは『思考の方法』で、問題の具体的な性格を明らかにするには観察が必要であり、何が問題なのかを調べる前に診断の候補を結論として受け入れると、思考はそこで打ち切られるとする (Dewey 1910, 74)。デューイのいう問題は本章のいう問いと完全には重ならず、デューイのいう観念や仮説は可能な解の側、つまり答えの候補の側にある。そのうえでデューイは、どのデータが重要かを定めることと、データが示唆する意味を展開することとは、互いに依存するとした。選んだ事実が解釈を支えなければデータを組み直し、解釈の候補もまたデータとの関係のなかで展開するのである (Dewey 1910, 103–104)。ニクルズが読むポパー (Popper) の推測と反駁の方法でも、推測した解を出して批判することで、問題はより十分かつ精確に定まる (Nickles 1981, 92)。どちらの記述でも、答えの候補を扱うことが問いの形を変えるが、問いと候補は同じものとされていない。

問いをさらに後に置く説もある。理解を目指す探究を論じたエル・シャズリー (El Shazly) は、そうした探究は無知ではなく、困惑、混乱、言説の不一致から始まるとし、そこでは何を問うかを定めることそのものが探究の一部であって、対象への関与が特定の問いを定めるので、問いは説明の順序のうえで後にくる、と論じる (El Shazly 2025, 15–18)。この説明では、後に定まる問いの内容を、探究の始まりに遡って置くことはできない。それでも、まだ特定の問いとして内容の定まらない関与が、 S にとって新しい理解などに向けた試みとして組織されていることはありうる。最小限の問いの構造が捉えるのは、その向きである。

デューイの探究の理論に立つエンヌ (Henne) は、カルナップ (Carnap) の内的問いと外的問いの区別に着想を得た区別を、デューイの理論によって定式化し直している。カルナップの区別では、内的問いは、ある言語的枠組みのなかで立てられ、その枠組みのなかで答えられる問いであり、ものについて語る言語のなかで「机の上に白い紙はあるか」と問うのがこれにあたる。外的問いは、その枠組みそのものについての問いであり、新しい種類の対象について語る規則を導入するか、ある枠組みを別の枠組みより採るか、といった問いがこれにあたる (Henne 2023, §2)。エンヌの区別は、問いの区別ではなく探究の区別である。枠組みの内での探究 (framed inquiry) は、概念的枠組みを問わずに用いて、問いを立て、答える。エンヌの例では、生化学者が温度がタンパク質合成に及ぼす影響を調べる探究がこれにあたる。枠組みを問う探究 (framing inquiry) では、問いや問題が既存の枠組みでは十分に定式化できず、枠組みそのものが問われ、修正され、拡張され、あるいは創出される。ウィリアム・トムソン (William Thomson) が絶対温度の概念に基づく理論的枠組みを組み立てた探究が、その例である (Henne 2023, §§1, 6)。エンヌ自身が強調するように、この区別の核は、問いが枠組みに対してどこに位置するかではなく、枠組みが、問うべき問いとその答え方を決めているかどうかであり、枠組みについての探究であっても、枠組みの内での探究でありうる (Henne 2023, §6.2)。

また、エンヌの区別は、カルナップの区別のようにはっきりした線を引くものではない。エンヌは探究を、結果ではなく状況と進め方によって分類するので、定義を変えるべきかどうかを問う探究は、定義を保持して終わったとしても枠組みを問う探究である。またエンヌは、この区別を程度の問題とし、何を一つの探究とみなすかによって変わるものとする。最も急進的な枠組みを問う探究でさえ何らかの枠組みに依拠しており、最も典型的な枠組みの内での探究でも、枠組みが状況の扱い方を完全に決めるわけではない。ひとつの探究が枠組みの内での探究として始まり、枠組みを問う探究として終わることも、その逆もある (Henne 2023, §§1-2, 6)。

この区別に照らすと、推論的問い論理が問いの形成をどこまでとらえるかを問える。ヴィシニェフスキは、問いが平叙文の集合や別の問いから導かれる仕方、二つの関係として形式的に定める。おおまかに言えば、一つは、平叙文の集合から、それらが真なら答えがあるはずなのにその答えがまだ決まらない問いが出てくる関係であり、喚起 (evocation) と呼ばれる。もう一つは、主たる問いから、それに答えるのに役立つ補助的な問いが出てくる関係であり、問いの含意 (erotetic implication) と呼ばれる (Wiśniewski 2022, 107-108, 112, 116-117)。そのうえで、主たる問いを根に置き、問いの含意によって補助的な問いを加え、問い合わせ (query) にあたる補助的な問いの直接的な答えごとに枝を分けた問いと答えの木を、問いの探索シナリオ (erotetic search scenario) と呼ぶ (Wiśniewski 2022, 109, 118, 122-124)。ヴィシニェフスキは、そうしたシナリオの上で定義された操作によってシナリオの生成を自動化できることを、フレボフスキ (Chlebowski) らの仕事として報告している (Wiśniewski 2022, 125)。シナリオは実行のなかで作り替えられうるが (Wiśniewski 2022, 125 n. 15; Wiśniewski 2013, ch. 13)、主たる問いが定める主たる目標は変わらない (Wiśniewski 2022, 122)。これらの定義が前提するのは、問いとその直接的な答えの集合が言語のなかで定まっていることであり、問いの含意と探索シナリオについては、さらに主たる問いが動かないことである。ヴィシニェフスキ自身、矛盾からはどんな文も導かれるという原理 (爆発律) を認める論理を背景にすると、不整合な前提はどんな問いも喚起しないとする。そのため、問いが不整合から出てくる場合には、矛盾を含んでも何でも導かれるわけではない論理 (矛盾許容的な論理) を背景に取るような別の戦略が要るとし、推論的問い論理の一般的な枠組みはそれを妨げないとする (Wiśniewski 2013, 62-63; Wiśniewski 2022, 113-114)。また、妥当ではないがもっともらしい問いの推論には、推論的問い論理がまだ説明を与えていないとしている (Wiśniewski 2013, 66)。こうして見ると、これらの定義がとらえるのは、答えの集合が言語のなかで定まり、問いの含意についてはさらに主たる問いも動かないときの問いの出方であり、エンヌの意味での枠組みの内での探究に属する一つの場合である。ただし、枠組みの内での探究でも、主たる問いは対照項の指定などによって立て直されうる (第3節)。また、答えの空間が変わることは、それだけで枠組みを問う探究にあたるわけではない。エンヌの基準は、枠組みが問うべき問いとその答え方を決めているかどうかであって、答えの範囲そのものではないからである。

形式化がとらえないのは、たとえば、探究が、異常、アーカイブ、データセット、装置、割り当てられた課題、失敗した介入、見慣れない素材、実践的な必要、既存の記述がしっくりこないという漠然とした感覚などから始まり、何を問うかそのものが探究のなかで定まっていくなかで定まってくる場合である。そうした漠然とした始まりを、答えを探せる問いにするには、何を、どの集団について、どの単位で、何と比べ、いつからいつまでの範囲で、どの基準で数えるのか、つまり尺度、母集団、分析単位、対比、時間的な境界、証拠の基準などを選ぶことが要する場合がある。数学のように、問いを述べる形式言語を選ぶことが要する場合もある。そうした選択によって、ひとつに見えた問いがいくつかに分かれ、いくつかの問いがひとつにまとまり、あるいは元の対象が部分的に測定や分類の産物であったと分かるかもしれない。問いを確定的なものにすることは、すべてを一度に決めることではない。第III部の導入で見た保健局のノートでは、比較の基準、期間、症例定義、評価する患者が当面のものとして定められ、原因は未決のまま残された。国際疾病分類のコードで出生登録に記された例を、デューイのいう問題を具体化するための観察データとみなすなら、症例定義や比較の基準は、それとは別に当面のものとして定められたものである。しかも、こうした選択は同じ種類の変化ではない。比較の基準や対照項を選び直すことは、多くの場合、既存の概念のまま問いを洗練することであり、エンヌの区別を当てはめれば、枠組みの内での探究にとどまると考えられる。これに対して、何を症例と数えるかの定義を改めることは、概念の適用条件の小さな改訂として、枠組みを問う度合いの小さい探究にあたりうる。エンヌ自身、同意年齢の引き上げのような適用条件の小さな改訂を、度合いの小さい例として挙げている (Henne 2023, §6)。

こうした選択はうまくいくこともいかないこともあり、うまくいけば、それまでなかった問いが手に入る。そのため、明示的で扱いやすい問いを言葉にすること、ひとつに見えた問いを分けること、問いを当面固定して追えるようにすることは、探究の前に外から与えられる入力にとどまらず、探究のなかでなしとげられる仕事でありうる。それが進歩にあたるのか、後の進歩を促すにとどまるのかでは、ロスとハリスの見方が分かれる (第2章第8.2節)。とはいえ、問いが探究のなかで形づくられるという主張は、過度に広げられやすい。後から振り返れば、どの出来事も何らかの問いに答えたものとして記述できるので、ある問いが探究を組織していたと後から帰属するには手がかりが要る。第1章第3節が問うた、ある行為を何が未解決の事柄に関係づけたのか、どのような結果が出ればその役割が変わったのか、活動の向きを変える力はどこにあったのか、に照らすと、遡及的な問いの帰属は、言葉で記述し直しただけで認めるなら緩すぎるものになる。帰属された未解決性が活動の組織を説明するとき、帰属はよりよく支持される。すなわち、なぜある操作が選ばれたのか、なぜ可能な結果がそれぞれ異なる仕方で重要となりえたのか、なぜある筋道が継続され、あるいは放棄されたのか、そして後の展開が、当の主体や社会にとって未解決であったことと、どう結びつき続けたのか、を説明するときである。

結果が後から答えたときされる問いは、その結果に至る探究を実際に組織していた問いとは限らない。セレンディピティ、つまり探していたのとは別のものを偶然に見いだす発見についての、アルフィーニ (Arfini) とコスタ (Costa) の説明 (第1章第3節を参照) がその例を与える。二人は、結果がまだ不確かなあいだに発見者が組み立てる説明と、研究が終わってはじめて得られる遡及的な説明とを区別し、セレンディピティによる発見のうち、現象の理解のされ方を変えるものではなく、説明されるべきものが過程の途中で変わると論じる。偶然から生まれる発明では、二つの説明は一致するとしている。二人の読みでは、宇宙マイクロ波背景放射の発見で知られるペンジアス (Penzias) とウィルソン (Wilson) は、自分たちのホーン反射器がなぜ期待どおりに働かないのかを説明しようとして始め、それが拾う雑音のありうる原因を除くことにしばらく時間を費やし、なぜそうした反射器ならどれもその信号を拾うのかを説明して終わった (Arfini and Costa 2025, §§2.1, 3-3.2, n. 13)。問いが移ることと、答えの候補が出てくることを混同しないように、ここからは問いと候補に添字を付けて書く。先に作業を組織していた問いを q_1 、後の問いを q_2 と書けば、本章の冒頭のマクリントックでは、第9染色体の短腕の遺伝子の構成についての問いが q_1 、不安定な遺伝子座の生じ方とふるまいについての問いが q_2 にあたる。同じように書けば、ホーン反射器がなぜ期待どおりに働かないのか (何がこの雑音を生むのか) が q_1 であり、雑音のありうる原因は q_1 への答えの候補 c であり、なぜそうした反射器ならどれもその信号を拾うのか q_2 である。説明されるべきものがこのように変わるのだとすれば、報告された結果が答える問い q_2 は、その結果に至った作業を組織した

問いではない。前の段落の条件に照らせば、雑音のありうる原因を除いていた作業を説明するのは、 q_2 ではなく q_1 である。作業全体を q_2 に向けたものとして描く説明は、結果については正しくありうるが、それに先立つ作業を何が組織していたかを記述してはいない（第5章第2節を参照）。

ベンジラスとウィルソンの例で問いの変化が分かるのは、研究が終わった後の再構成による。これに対して、マクリントックの報告と覚書のように、その時々にかかれた記録からは、その時点で何が作業を組織し、何が未決であったかを読みとれることがある。また、問いは、その見出しを保ったまま組み替えられることもある。ダーウィン（Darwin）のノートブックの編者たちの読みでは、種の変移についてのノートブックの基本的問い——種は変化する条件にどう適応するのか、新しい種はどう生じるのか、分類の階層はどう生じるのか——は、その最初のノートブックであるノートブックBの冒頭で立てられて保たれたが、それらの問いが何を含むとされるか、関連する現象がどう関係づけられるかは、答えとともに変わった（Barrett et al. 1987, 168, 329, 395–396）。答えの候補の形成についても、回想と同時期の記録は食い違う。ダーウィンの自伝は、1838年10月にマルサス（Malthus）を「楽しみのために」読んでいたとき、有利な変異は保存される傾向にあるだろうということが「たちまち」思い浮かんだと回想している。一方、その考えを記録したノートブックの記載は、ダーウィンがまだその本を読んでいた9月28日の日付をもつ。そこでは、その考えが、自然の経済の隙間にあらゆる種類の適応した構造を押し込もうとする「十万本の楔」のような力として書きとめられており、編者たちはこれを自然選択についての最初の定式化とみなす。この記載の前には、雑種の不稔性についての覚え書きと、ライエル（Lyell）の本から化石についての箇所を書き抜いたものがあり、後にはサル的好奇心についての聞き書きが続く。編者たちによれば、ダーウィンがマルサスを読み終えたのは10月2～3日であり、理論はその次のノートブックEでも形をとり続けた（Darwin 1958, 120; Barrett et al. 1987, 374–376）。回想では一つの突然の出来事であったものが、同時期の記録では、ほかの問いについての作業のさなかに書きとめられ、その後も形を変えていく一步として現れる。印象を消化して実質のある観念にするには時間が要り、観念は、複数の示唆を黙って練り直すことのなかで熟していく、というデューイの記述とも、この記録はよく合う（Dewey 1910, 37, 209–210）。基本的問いの中身が答えとともに変わったという編者たちの読みをこの記録と合わせれば、その変化も、こうした候補の書きとめのあいだで進んでいたと考えられる。機械が結果の後に書く報告も、同じように、作業の記録と食い違いうる（第5章第2.1節を参照）。

与えられた課題も、それだけでは問いに満たないことがある。たとえば、科学の実験室での認知と実践を研究したナーセシアン（Nersessian）が調べたある生物医学工学の実験室では、新しく加わった一人が、その実験室で工学的に作製された構築物を生きた動物へと持ち込む担当者に指名され、およそ1年後にも、なお自分の研究の具体的な目標と問題を定めているところだった（Nersessian 2009, 750）。与えられた役割が具体的な問いになるまでが、それ自体で長い仕事になっている。ニクルズの、実践的に十分なところで進むという議論からは、さらに別の見方が得られる。完全な定式化を先に得なくても、次の一步に進めるだけの定式化で探究を進めることはできる（Nickles 2019, 361–362）。

問いが変わるとき、何にとつて変わったのかも分けられる。ニクルズは、問題の変化に、行為者にとつての意味（ある科学者にとつて問題が変わること）と、意味論的な意味（理論と実践の体系にとつて問題が変わること）とを区別した。体系に暗に含まれていた制約を見いだして変わるのは、行為者にとつての問題であり、意味論的な移行は、目的、方法、理論的な文脈が大きく変わるときに起きる（Nickles 1981, 99）。ニクルズはまた、問題のなかでの変化と問題の取り替えとを分け（Nickles 1981, 100）、設定を大きく変えれば、元の問題ではなく置き換えの問題を解くことになるとする。解けないと判断した問題については、問題を定める制約を保ったまま、目的のほうを手放すこともあり、ニクルズは、永久機関を作るという目的が手放された例を挙げている（Nickles 1981, 95–96）。問える問いの範囲そのものが変わることもある。ブロムバーガーは、質量や電圧のように、個々の対象がそれぞれの値をもつ概念（ブロムバーガーのいう determinable）をつかむには相当の理論が要るとし、ニュートン以前には重力質量についての問いを誰も抱けず、アリストテレスはパルテノンの質量を知らないことを知る立場になった、と論じた。概念が変わると、ある対象について答えがあるはずの問いの範囲も変わるのである（Bromberger 1992, 134–135, 150 n. 10）。

これらの区別を機械に当てはめると、機械が問いを形成したと言われるとき、それが次の二つのどちらにあたるかを確かめられる。一つは、同じ枠組みのもとで問いを増やすことであり、もう一つは、その枠組みそのものを問いに付すことである。二つを見分けるには、まず機械にとつて何が枠組みにあたるかを定めておく必要がある。エンヌの説では、問いを立て仮説を確かめる手続きを統制する規則、つまり方法も、枠組みの一部である（Henne 2023, §5）。またエンヌは、典型的な枠組みの内での探究では、枠組みが問われないまま受け入れられていて「透明」であり、疑われたり不具合が生じたりしたときにはじめて気づかれると述べる（Henne 2023, §7.2）。これにならって、機械の枠組みを、その探究のなかで問いを立てるのに使われているが、問いの対象とはされていないものとして捉えれば、それが機械にどのような仕方で備わっているかによらずに、この区別を当てはめられる。同じ枠組みのもとで問いを増やすことの例が、ヴィシニエフスキが報告する問いの探索シナリオの自動生成である。これは、固定した言語と問いの含意の規則を問わずに用いて問いを生成する。シナリオを実行のなかで作り替えるシステムでも、主たる問いとその答えの集合は動かないので、枠組みの内での探究に属する。枠組みそのものを問いに付すことにあたると言うには、問いを立てる枠組みの規則そのものが、その探究のなかで実際に問いに付され、選び直しや改訂の対象とされていることが求められる。たとえば、問いを補助的な問いに分けていく手順がこの問題に合っているのかを問うことや、先に見た、何を症例と数えるかの定義を見直すことなどがこれにあたる。したがって、二つを分けるのは、機械が問いを立てるのに使った規則が、その探究のなかで問いに付されたかどうかである。問いに付された規則を同じ探究のなかで改められるかは、さらに別のことである。機械が問いを立てるのに使うもののなかには、機械の作られ方が据え、作り直しによって置き換えられる構造もありうる（第6章第5.3節、第8章第2.1節）、探究がそれを問いに付しても、その改訂は機械自身の探究の外に属することがある。ただし、規則を問いに付す探究がどこまで枠組みを問うものになるかには、エンヌの説では度合いの差がある。また、二つのどちらにあたるかは、何を一つの探究とみなすかにもよる。全体としては枠組みを問う人間と機械の探究のなかで、機械が担うのはその一部の、局所的には枠組みの内にある探究だということもありうる。

ここまで見た少数の事例と説からだけでも、問いの立て直しにはいくつもの形がありうるということがうかがえる。機械が「問いを立て直した」と言われるときにも、何が変わったのかによって、たとえば次のような変化を書き分けられる。同じ規則のもとで補助的な問いを加えることがあり、ヴィシニエフスキの探索シナリオがこれにあたる。枠組みの内での問いを洗練することもあり、リプトンの対照項の指定や、ニクルズのいう、体系に暗に含まれていた制約の発見などがこれにあたる。エンヌのいう、概念の適用条件や問いを立てる手続きの規則を、探究のなかで問いに付すこともある。エンヌの分類は進め方によるので、結果として規則が改められるとは限らない。さらに、ニクルズのいう、制約を大きく変えて置き換えの問題に移ることや、解けないと判断した問題について、制約を保ったまま目的を手放すこともある。これらは網羅的な一覧ではなく、互いに排他的な種類でもなく、同じ探究のなかで重なりうる。

置き換えの問題に移ることが、枠組みを問いに付すことを伴う場合もある。また、枠組みの内での問いを洗練することと、規則を問いに付すこととのあいだは、エンヌの説では度合いの問題であり、何を一つの探究とみなすかによっても変わる。ニクルズの議論では、探究を進めながら定式化を暫定的なものにとどめることもできる。

漠然とした始まりを問いにするときの選択では、あるものが当面固定され、ほかのものは未決のまま残されるので、システムが異常を問いに変えるという主張は、選択ごとに確かめられる。どの選択をシステム自身が行い、どれを課題から引き継ぎ、どれを未決のまま残したのか、そして、デューイの記述に照らして、問題を十分に調べる前に候補を受け入れていないか、である。また、枠組みの内での探究で立てた問いと退けた候補をすべて記録しても、その枠組みのもとでは立てられなかった問いは、記録に現れない。そのため、実行の記録が完全であることは、それだけでは、システムが枠組みを問う意味で問いを開く能力の評価にはならない（第5章第2.1節を参照）。

ブロムバーガーによれば、どんな問いを立てられるかは、使える概念に依存する。質量の概念がなければ、バルテノンの質量を問うことも、それを知らないと知ることでもできなかった。ブロムバーガーは人工システムを論じていないが、これを機械に当てはめると、二つのことが考えられる。一つは、既存の概念を組み合わせて、これまでにない文の問いを数多く生成することと、新しい概念を持ち込み、それまで立てられなかった問いを立てられるようにすることは、別の仕事だということである。この区別は人間の研究者にも当てはまるが、機械は前者を低い費用で行えるので、生成された問いが「新しい」と評価されるとき（第8節）、それがどちらの意味での新しさなのかが、とくに問われる。もう一つは、機械の表現が *S* の使わない概念を使えるなら、機械は *S* がまだ立てられない問いを立てうるが、そのとき *S* は、それを問いとして受け取る立場にないかもしれない、ということである（第7節）。どちらも、機械の表現にそうした概念をつかんでいることを機能の水準で帰属できる場合に成り立つ。ニクルズの二つの区別を当てはめると、機械が問いを立て直したときには、誰にとつて問題が変わったのかと、同じ問題のなかでの変化か別の問題への取り替えかとを、分けて問える。ニクルズの「体系」を *S* の理論と実践の体系として読むなら、機械が *S* の体系に暗に含まれていた制約を見いだす場合、変わるのは機械にとつての問題であり、*S* の体系にとつての問題ではない。たとえば、機械が *S* の既存の理論から、答えのとりのうる値に上限があることを導いたとしても、その上限がその理論にすでに暗に含まれていたなら、*S* の問題そのものが変わったわけではない。目的や方法が大きく変われば *S* の体系にとつての問題も変わり、その変更を受け入れるかどうかは *S* の側で決まる。また、体系にとつて問題が変わったことだけで、元の問題を取り替えたことになるわけでもないが、制約を大きく緩めて解を得たシステムは、元の問題ではなく置き換えの問題を解いている。これは、チェックには合格したが、それが問われた問いとは別の問いについてだった、という場合（第6章第3.2節）の一つの形である。

問いの立て直しは、結果の側からも起こる。失敗の所在をどこに帰すかの判断は、次にどの道を進むかを左右し（第6章第2.4節）、基準そのものが問われるときには、何が答えに数えられるかも改められうる（第6章第5節）。何かを理解したことが、さらなる問いを開くこともある（第3章第7.3節）。問いの形成と立て直しは、答えの候補や結果とのこうした関係のなかで起こるので、機械が「問いを立て直した」という主張を確かめるには、問いの側の記録だけでなく、そのとき候補や結果の側で何が起きていたかも見る必要がある。

5. 問いの前提と、問いの閉じ方

問いには、形づくられ、立て直されるほかにも行き先がある。答えの見つからないまま残ること、前提とのかかわりで問いとして立たなくなること、保留されること、後回しにされること、答えが得られて決着することなどである。機械の場合、実行が終わったことや出力が完成したことは記録に残るが、それで問いが決着したのか、問いが立たなくなったのか、保留されただけなのかは、記録の上では見分けがつかないことがある。

では、問いがどうなったかは、何によって決まるのか。本節はこれを三つの側から見る。一つは、研究のなかで出会った事柄がどう扱われるかであり、研究の記録に残る事例から見。二つ目は、問いが答えのないまま立たなくなる場合であり、問いを成り立たせている前提や文脈についての説などから見る。三つ目は、問いを答えによって閉じること、また答えによらずに保留したり、やめたりすることであり、問いや探究にかかわる態度や意図についての説などから見る。各説が扱う条件は、それぞれ異なる。何が前提されているか、科学者の共同体がどの問題を認めるか、問う側がどのような態度や意図をもつか、などである。しかし、どの説を当てはめても、それは実行が終わったか、出力ができたかとは別の事柄になる。そのため機械については、「終わった」という記録と、問いがどうなったかとを、別々に確かめることになる。

まず、研究のなかで出会う事柄の扱われ方を見る。そのために、探究を組織している未決着の点と、その点を追ううちに会った未解決の事柄とを分けて呼ぶ。後者は、前者になることも、ならないこともある。未決着の点を追ううちには、効果は本物か、素材や装置は想定どおりにふるまうか、食い違いは何を示すのか、何を症例と数えるのか、といった事柄にも出会う（第6章第2節を参照）。そうした事柄は、保健局のノートで症例の数え方が当面定められ、異常と疾患の関係が決められずにおかれたように（第III部の導入）、またマクリントックが競合する解釈を未決のまま残したように（本章の冒頭）、さまざまな行き先をたどる。後回しにされた事柄もある。大気中の二酸化炭素を連続して観測していたチャールズ・デイヴィッド・キーリング（Charles David Keeling）は、観測計画の慌ただしさのなかで、分光学者に指摘された校正ガスの偏りの可能性を調べるのを先送りしたと回想している（[Keeling 1998, 46, 50](#)）。言葉にされないまま技能によって扱われた事柄もある。電波望遠鏡の記録紙を目で読んでいたベル・バーネル（Bell Burnell）は、最初の数百フィートの記録紙を読んだ後で、惑星間空間のプラズマのために強度がまたたく（シンチレーションを起こす）電波源と、干渉とを見分けられるようになったと回想している（[Bell Burnell 1977](#)）。計画の外で出会った事柄が、問いそのものになることもある。マクリントックの不安定な遺伝子座は、遺伝子の構成を調べる実験のなかで現れ、やがて作業を組織する問いになった（本章の冒頭）。

これらの記録によれば、出会った事柄のうち、どれを問いとして取り上げ、どれを当面のものとして固定し、明示的に未決として残し、後回しにし、あるいは言葉にしないまま扱うかは、仕事が始まる前に固定されているものではなく、研究の仕事の一部でありうる。一つひとつの事柄がどうなるかには既定の向きがなく、さらなる問いを開くこともある（第2章第8.2節を参照）。しかも、そのすべてが意図的な選択であるわけではない。事柄は、誰もそれに気づけなかったために、あるいは時間がなかったために手つかずのまま残ることがある。また、問われないままのものの一部は、未決のまま残されているのではなく、ほかの何かを問うための条件として固定されている（第8章第2.1節を参照）。逆に、ある事柄が取り上げられるのも、選択によるとは限らない。遺伝学者のボー・シア（Bo Xia）は、尾骨を負傷して座るたびに痛みが戻り、そのたびに人間に尾がないことへ注意を引き戻されたかと回想しており、

その関心がのちに尾の消失についての研究につながった (Yanai and Lercher 2024; Xia et al. 2024)。キーリング、ベル・バーネル、ポー・シアの回想は、第5章第6.4節で別の側から見る。他方で、取り上げるか、そのままにするかの選択が価値を負い、ほかの可能性を除くことを強調してきた論者もある。ラベッツは、提案された研究計画は優先度が低いとされれば着手されず、その点で私たちは無知のままにとどまるとし、私たちのもつ知識は、ある種類の知識を得てほかの種類を犠牲にする「価値を負った決定」との関係においてのみ存在するとした (Ravetz 1993, 348)。何を追うかを前もって決められないことは、第III部の導入で触れたラインベルガーの記述にも現れる。ラインベルガーによれば、実験システムにおける研究の対象は、まだ知られていないことそのものを、実験で操作できる形で体現しており、そうしたシステムが生み出す知識は、その時点で予期できることの先にある (Rheinberger 2011, 312–313)。第4節で見た問いの帰属と同じく、ある事柄を、一度も気づかれなかったものとしてではなく、そのままにされたもの、後回しにされたものとして記述することも、それが気づかれていたことを記録が示すときに、よりよく支持される。保健局のノート、マクリントックの文書、キーリングの回想は、そうした記録である。

次に、問いが答えの見つからないまま立たなくなる場合を見る。コリングウッドとファン・フラーセンは問いが「生じる」条件を、クーンとラウダンは共同体や伝統がどの問題を認め、その解決をどう評価するかを論じる。これらは別々の問いである。ここでの「生じる」は、問いがどこで思い浮かぶかと言うのではない。第1節で見たように、問いは実験のさなかにも食事の席でも浮かびうる。ここで扱うのは、浮かんだ問いが、何が前提され、何が背景にあるかに照らして、問いとして成り立つかどうかである。コリングウッドは、『形而上学論』で、あらゆる問いは前提を含み、問いはその前提から直接に生じるとした。問いを生じさせるこの力は、前提が真であるかにも、真と思われているかにもよらず、それが前提されていることだけによる。そのためコリングウッドの説では、偽と知りつつ仮定したことからも問いは生じ、それにもとづく論証の妥当性も損なわれない。逆に、真である前提でも、それが置かれていなければ問いは生じない (Collingwood 1940, 25–28)。問いが「生じない」とは、この定義では、実際には置かれていない前提を含んでいるということであり、コリングウッドはそうした問いを、それ自体として無意味なのではなく、その文脈に相対的な意味で「無意味」と呼ぶ (Collingwood 1940, 26)。コリングウッドはさらに前提を二種類に分けた。相対的前提は、ある問いにとっては前提であり、別の問いにとってはその答えである。それは、使っているあいだは前提として働き、点検するときには、それを答えとする問いが立てられる (Collingwood 1940, 29–30)。これに対して絶対的前提は、それが関係するあらゆる問いにとって前提としてのみ立ち、答えとしては立たない。それは検証が難しい前提なのではなく、検証という考えが当てはまらない前提である。その例の一つ、起こることにはすべて原因があるという前提は、コリングウッドによれば、この本の書かれた百年前には自然科学のあらゆる分野で置かれていたが、その後いくつかの分野では置かれなくなった (Collingwood 1940, 31–32)。そうした前提が置かれなくなれば、それに依拠していた問いも生じなくなる。しかも、人はふつう自分の絶対的前提を自覚していないので、その変化は選択の問題ではありえず、自覚されないまま思考のなかで起こる、とされる (Collingwood 1940, 48)。

クーンは、関連する問題を科学者の共同体の水準に置く。ディミトラコス (Dimitrakos) が引用するクーンの言葉では、パラダイムは、それが受け入れられているあいだは解があると想定できる問題を選ぶ基準を与え、共同体が科学的と認める問題は、おおむねその範囲に限られる (Dimitrakos 2023, §4.1)。フェローズとリウ (Fellows and Liu) はさらに、パラダイムが与える概念と道具では定式化できない、社会的に重要な問題から共同体が隔てられるというクーンの主張を引用している (Fellows and Liu 2015, 19)。ここで扱われるのは、通常科学の実践で問題として認められるかどうかである。排除されたからといって、その問題に答えが出たことにも、客観的に解けないことにもならない。

ファン・フラーセンは、なぜの問いについて、さらに別の仕方でも「生じる」を定めた。クロスとルーロフセンが示すところでは、第3節の三つ組の分析に加えて、なぜの問いは、話題が真であること、対比類のなかで話題だけが真であること、そして話題と対比類の組に関連関係をもつ何らかの命題が真であることを前提とする。文脈によって定まる背景知識 K が最初の二つの前提を含意しないとき、問いは生じず、三つ目の前提が成り立たないとき、問いは、生じていても答えをもたない (Cross and Roelofsen 2026, §3.2)。

前提の真偽と問いの扱いとを切り離す議論は、探究の規範の側にもある。探究がいつ許されるかを論じるファルボ (Falbo) は、偽の前提をもち真の直接的な答えを欠く問いがあるという意味論を受け入れたうえで (Falbo 2024, 965–966)、真の答えのない問いへの探究が許される場合があること (Falbo 2024, 969–976)、逆に、真の答えがあっても、探究によって認識的に改善しないと確信するのが合理的なら、探究は許されないこと (Falbo 2024, 980–981) を論じる。これは問いの地位ではなく探究の許容についての主張であり、問いの意味論と探究の規範が別々に動きうることを示している。ブロムバーガーは、無知の側から前提の真偽を扱う。前提が真である問いについての無知を真正の無知、偽である問いについての無知を見せかけの無知と呼び (Bromberger 1992, 121)、フロギストンの比重についての無知は治されたのではなく、消えたのだとした (Bromberger 1992, 131 n. 4)。

理論の評価を論じるラウダ (Laudan) の説にも、問題が問題でなくなることについての記述がある。デルセン (Dellsén) の紹介によれば、この説では、研究がそれを背景として行われる研究伝統が、何が本物の問題に数えられるか、理論がそれを解くことがどれほど重要か、何が解決に数えられるかを定める。理論が解くべき経験的な問題は、実際の事態にかかわっている必要はなく、伝統の支持者たちがそうだと信じていれば足りる。そして進歩は、問題と解決を定める伝統の部分が変わることにありうる。未解決の問題が擬似問題とみなされるようになる場合がそうである (Dellsén 2018, §3)。ラウダが同僚たちとともにラウダ自身のモデルについて挙げるテーゼには、次のようなものがある。経験的な問題を定めるには、探究の文脈をあらかじめ特定しておくことが要る。対抗する理論が直面している経験的な困難を解けると、その理論はとりわけ説得力があると受け取られる。問題がどれほど重要かを大きく左右するのは歴史的な文脈であり、同じ問題でも、文脈によって証拠としての重みが異なりうる。そして、古い理論を新しい理論に置き換えることには、ふつう、得るものと失うものの両方がある。たとえば、キュヴィエ (Cuvier) とライエル (Lyell) 以前の地質学の理論家たちが取り組んでいた経験的な困難のなかには、のちに顧みられなくなったものがある (Laudan et al. 1993, 72–73, 75)。さらに第7テーゼでは、適切な解決の基準は研究を導く諸前提によって定まり、その前提が変われば基準も変わりうるとされる (Laudan et al. 1993, "Teses de Laudan," thesis 7)。これらの説は、それぞれ別の行き先を記述している。コリングウッドとファン・フラーセンの、問いが生じるための条件は、クーンの、共同体が科学的な問題として認める基準とは異なる。ラウダンのいう、問題が問題でなくなることや、理論の置き換えに伴って顧みられなくなることも、解けないと判断した問題の制約を保ちつつ目的を手放すという、ニクルズの論じる場合 (Nickles 1981, 95–96) とは異なる。いずれも、単に答えがまだ見つかっていないということではない。

これらの説を機械に当てはめると、問いが生じること、共同体が問題として認めること、問題を評価することが、別々に問われる。コリングウッドの説からは、機械がどの前提を置いており、それがどの問いを生じさせているかが問われる。コリングウッドの「前提する」は考える者の心の働きなので、機械

に当てはめるには、前提が問いを生じさせる働きを、機能の水準で読み替えることになる。もっとも、コリングウッド自身、ある前提が相対的か絶対的かは内省では決まらなとし、ある科学の仕事でどの絶対的前提が置かれていたかは、その仕事の記録を証拠とする分析によって見いだされるとしていた (Collingwood 1940, 43, 59–60)。この方法は機械の出力にも同じように向けられるので、機械が、自らを形づくったものに置かれていた絶対的前提を受け継ぐという主張は、ある時点で何が前提されていたかについての、コリングウッドのいう歴史的な主張として、機械の出力の分析によって調べうる。ただし、出力の分析が証拠を与えるのは、どの前提が働いているかについてであり、それがどこから来たのかを確かめるには、機械を形づくったものとの照合も要る。また、コリングウッドは、人は自分がどの前提を置いているかを知らずに前提を置きうるとし (Collingwood 1940, 26)、ある前提が問いうるものであることも、それを使う者がそう気づいていないことによっては否定されないとした (Collingwood 1940, 30)。この議論を当てはめても、機械が自らの前提を明示し、それを前提として自覚していることまでは求められない。暗黙の表象さえ要らないかどうかは、機能の水準への読み替えの中身による。

ファン・フラースンの説を当てはめると、なぜの問いが生じるかどうかは、背景知識 K に相対的になる。クロスとルーロフセンが示すこの説では、その背景知識は文脈によって定まるので (Cross and Roelofsen 2026, §3.2)、機械の背景知識と S の背景知識を別々に立てることは、この枠組みをそのまま当てはめることではなく、拡張することになる。そのように拡張すると、話題とその対比類について二つの背景知識が含意することは異なりうるので、同じなぜの疑問文が、機械の背景知識のもとでは生じ、 S の背景知識のもとでは生じない、あるいはその逆になることがある。したがって、問いが生じているかを機械の K に照らして判定するのか、 S の K に照らすのか、その両方か、という選択は、この説自身が定めるものではない。また、第3節で報告した、全知の存在は説明をもつものとは理解されないという点が成り立つなら、機械の K の大きさは、それだけでは機械にとってなぜの問いを生じさせない。

クーンとラウダンの説を科学者の共同体である S に当てはめると、探究の組織についての問いが加わる。 S はどの問題を認め、何を適切な解決とみなすのか。人工システムを組み込むことで、その実践で受け入れられる概念や道具、解決を律する基準が変わるなら、認められる問題の範囲も変わりうる。ただ機械を加えただけでは、その変化があったことにはならない。

この当てはめに、機械を S の成員とみなすことは要らない。道具も、問題を定式化し調べる仕事に影響しうるからである。ただし、使える装置が変わることと共同体の実践が変わることは、分けて確かめる必要がある。同じく、機械が疑問文を処理できることだけでは、 S がそれに対応する問題を認めていることにはならない。人工の共同体へと説を広げるには、その参加者が、問題と解決について共有された基準を保っているという、さらに別の議論が要る。

前提の変わり方の違いも、機械と S のずれにかかわる。前提が手放されたときに問いを意図的に取り下げるには、それぞれの問いが何を前提しているかを把握しておくことが要り、点検によって意識的に手放されうるコリングウッドの相対的前提では、この条件が当てはまる。しかし絶対的前提の場合、問いは、誰もその前提を表象しないまま生じなくなりうる。そのため、 S ではある問いが生じなくなったのに、疑問文の一覧を明示的に保っている機械がそれを追い続ける、というずれがありうる。ファルボの規範を当てはめると、機械が偽の前提をもつ問いを追ったことは、それだけでは失敗の証拠にも成功の証拠にもならない。この規範が探究の許容に求めるのは、探究しても自分がその問いについて認識的に改善しないと確信することが、その時点で合理的でないことだからである (Falbo 2024, 963)。ただしその場合には、改善を機械自身と S の成員のどちらについて評価するかを指定することが要る。後者をとるのは、探究者自身の改善を問うファルボの規範を広げる読みである。評価の宛先が機械を組み込んだ者や S になるなら、批判はその前提をシステムに組み込んだ者に向かうという第2章第5節の構図につながる。ラウダンの説からは、別のずれが出てくる。何が問題に数えられ、それがどれほど重要かは、研究伝統とその歴史的な文脈によって定まるので、機械の側で扱わないとされた事柄が、 S の別の伝統では重要な問題に数えられることもありうるし、逆に、機械が追っている問題が、そこではすでに擬似問題として退けられていることもありうる。

三つ目に、問いを閉じること、保留すること、やめることを見る。問いは、答えられないまま立たなくなるだけでなく、答えによって決着することもある。決着とは何か、論者たちの語でいえば問いや探究を「閉じる」とは何かについても、説明は分かれる。コリングウッドは、答えられても問いは問いであり続け、答えられていない問いでなくなるだけとした (Collingwood 1940, 25)。ジェーン・フリードマンは、探究を閉じる態度を、疑問の態度と合理的に両立しない態度として特定する。問いへの完全な答えを信じることは問いを閉じるが、信念度は、高くても低くても、問いを落ち着かせる意見ではない。信念度が最大であっても、少なくとも一部の場合には問いは閉じない、とも論じられる。公正な硬貨を無限回投げてすべて表が出るかを調べるのが筋の通ったことだと認めるなら、そうなる。また、答えずに問いを手放した状態は、閉じることとは別だとされる (Friedman 2019)。探究を終えることについても、見方は一つではない。完結の点をもつ動詞とは、「家を建てる」のように、そこに達すれば行為が完了する点をもつ動詞のことである。ジェーン・フリードマンは、「探究する」はそうした点をもたない動詞であり、知識に達しても探究は定義上は終わらず、探究を終えるのは別の一歩だとする。「解き明かす」のような語が完結的に見えることは認めつつ、それは探究の活動ではなく、成功した探究の終わりを指す語なのかもしれない、としている (Friedman 2024)。これに対してダニエル・C・フリードマン (Daniel C. Friedman) は、「問いに答える」「解き明かす」のような、能動的な探究を言い表す中心の語が完結の点をもつことを、探究が達成されるべき目標をもつことの裏づけとする。その説では、探究は問いに向けられた意図によって組織され、その意図が、信念、受容、知識、理解、確実さなど、どの決着の態度で探究を閉じるかを、一部は指定する。決着の態度とは、ある答えを真として引き受け、以後の思考と行為にその答えを使ってよいとする態度である。そのうえで、ある決着の態度を形成するつもりでいることは、そこが最適な停止点であることを意味せず、認識的に適切に決着するには、意図が指定するより多くが求められうる、とされる (D. C. Friedman 2024)。第1章第3節でも取り上げたケルプは、信念を、問いについての探究をその人にとって閉じる種類の一手とし、探究がその人にとって決着する、つまり適切に閉じるのは、その人が答えを知るときであり、またそのときに限るとした (Kelp 2021c, 13–14)。

答えによらずに問いの扱いを変える仕方も一つではない。第1章第3節で見たリーは、判断の保留に二つの型があるという整理に依拠する。問いを研究の議題に載せたままにする型と、議題から下ろす型である。後者をリーは探究を閉じる型とも言うが、それは答えによる決着ではない (W. Lee 2026, 12)。そのうえでリーは、手持ちの証拠が中立で、これ以上の情報を得られないと考えるなら、疑いが残っていても問いを棄り、探究しないことがあるとする (W. Lee 2026, 15)。後回しについては、コリングウッドが、相対的に当然視することを、前提の検証をより適切な機会まで延期することと、怠惰や気後れによる先延ばしとに分け、それらを、原理上検証を許しも要しもしない、絶対的に当然視されることと対している (Collingwood 1940, 41–42)。決着が早すぎることについては、デューイが、示唆をすぐに受け入れるのは無批判な思考であり、最初の示唆を早まって受け入れた結論にしまうと、思考はそれだけ切りつめられると述べていた (Dewey 1910, 13, 74)。どの探究に取りかかるかの判断がそれ自体一つの探究であるというチャーニアクの議論は、

第5章第6.5節で取り上げる。探究を始め、続け、やめることがどのような規範に従うのかも論じられており、ジェーン・フリードマンは、そうした規範が認識的規範とどう関係するのかを、探究論的 (zetetic) 認識論の問いとして立てている (Friedman 2020, 1, 12)。その論争の一部は、第2章第5節で見た。

これらの説を当てはめると、機械が「問いを閉じた」と記述する前に、別々に問えることがいくつかある。ジェーン・フリードマンの基準に立てば、実行の終了も、出力の完成も、答えの候補に置かれた、一つに大きく偏った確率の分布も、それだけでは問いを閉じるのではなく、閉じる状態にあたりうるのは、その状態にありながら問いを追い続けるなら筋が通らない、とされる役割をもつ状態に限られる。ただし、機械の状態に信念と信念度の区別が立つか、そうした筋の通らなさを機械について語れるかは、この当てはめに要する別の前提である。また、「報告を出すまで走る」のように課題を完結の点をもつものとして組み立てると、その終わりは課題の指定で決まり、問いの状態では決まらない。探究に完結の点を認めないジェーン・フリードマンの説では、実行の終了と問いの決着がずれる理由の一つがここにあり、ダニエル・C・フリードマンの説では、ずれの所在は、課題が指定した終わり、と問いに向けられた意図が指定する決着との違いに移る。停止の基準は、機械を組み込む者が定めるにせよ、機械自身が定めるにせよ、決着の態度の指定と形のうえで似ているが、ダニエル・C・フリードマンの区別では、指定された決着と認識的に適切な決着とは別である。機械が、問いに向けられた意図にあたる計画の状態をもつかどうかも前提である。ダニエル・C・フリードマンは意図の内容を暗黙でもよく、機能によって理解されるものとするので、当てはめの敷居は高くないが、それでも、手段と目的の整合、一貫性、安定性への合理的な圧力に応じる構造は要る (D. C. Friedman 2024)。ケルプの区別を当てはめるなら、機械の停止だけからは、探究が機械にとって閉じたとも、適切に閉じたとも決まらず、答えについての信念にあたる状態を機械に帰属できるか、さらにその状態が知識にあたるかを、別々に問うことになる。決着せずに保留し、未決のまま保つには、事柄を暫定的なもの、あるいは未決のものとして、立ち返ることのできる形でとどめておくことが要る。記録はそうした立ち返りを支える仕組みの一つでありうる。第5章第6.3節で見るように、ある公開の化学の実験ノートでは、指導者が書き加えた問いに学生が約1年後に答えており、大腸菌の長期進化実験では、まだ定式化されていない問いのためにも試料が保たれていた。そのため、立ち返る先は、すでに立てられた問いに限られなかった。出会った事柄を追うか捨てるかしかできないシステムは、この幅の一部をもつにとどまる。また、機械の側で決着した問いが、*S* で決着したことになるとは限らない。*S* が人間の社会であれば、少なくとも結果が *S* の認識的状况に入ることが要り (第2章第9.4節)、*S* に機械の成員がいる場合も、一つの成員の側での決着は *S* の決着と同じではない。

三つの側から見たことをまとめると、機械については次のことが見えてくる。研究のなかで出会った事柄のどれを問いとし、どれを後回しにし、どれを言葉にしないまま扱うかは、それ自体が研究の仕事の一部でありうる。コリングウッドの説では、前提の変化に誰も気づかないまま問いが生じなくなることがあり、ファン・フラーセンの説では、機械の背景知識と *S* の背景知識を別々に立てると、なぜの問いが一方のもとは生じ、他方のもとは生じないことがある。したがって、どちらの説でも、機械と *S* が追う問いはずれうる。クーンとラウダンの説を科学者の共同体である *S* に当てはめると、それとは別に、認められる問題や解決の基準が変わったかが問われる。問いを答えによって閉じたのか、保留したのか、やめたのかも、実行の終了や出力の完成とは別に問われる。どの場合にも、機械の記録に残る「終わった」は、それだけでは問いがどうなったかを示さない。

さらに、これらの行き先を合わせると、研究の過程を、未決のまま残っているものの側から見ることもできる。行き先の変化のすべてが、一つの方向に向かうわけではない。問いの形成は無知を露わにし、候補の追求はさらに別の選択肢を明らかにしうる。プラグマティズムの創始者の一人であるパース (Peirce) は、疑いに促されて確定した信念へ至ろうとする格闘として探究を特徴づけた (Peirce 1877, 6)。ヒルデブランド (Hildebrand) によるデュエイの探究論の整理では、デュエイは、問題解決において探究がたどる型を描き、その探究は「不確定な状況」を「確定した状況」に変えたときに終わるとした (Hildebrand 2024, §4.3)。未決の事柄の側から見る見方は、探究全体の方向についてのこれらの説とは別の水準で、一つの研究の過程の内部にある複数の未決の事柄を扱う。また、進歩を一つの不確実性の量の減少と同一視しない。研究過程を見る仕方としては、この見方は、問い、方法、基準の変化を見えるようにする。どの事柄を、どの根拠で追求するかは、第5章第4.2節の問いとして残る。

6. 誰が問うのか

機械が *S* の成員に問いを示すとき、あるいは機械どうしが照会を重ねるとき、問うているのは誰なのか。機械については、この問いは、機械が好奇心をもつかという形で語られることが多い。しかし、問う行為の説、集団の探究の説、個人の探究の説は、それぞれ別のものを帰属させる説であり、同じ一つの行為についての競合する理論ではない。

問う行為の説から始める。第1章第7節で見たように、ワトソンは、問うことを、情報を引き出すというその特徴的な機能によって同定するので、問いは疑問文の形をとる必要がなく、別の明示的な目的で問われた問いも、真正の問いでありうる (Watson 2015, 86–102)。ワトソンの説では、欠けている情報が問う者自身に欠けている必要もない。答えを知っている者が別の者のために情報を求める場合には、情報を求める仕事は二人に分けて担われているのである (Watson 2015, 103–104)。またワトソンは、問いかけを、共通の目標をもつ社会的に確立された実践とし、一人で問うことも、他者から学んだ技能を用い、社会的な場面での問いかけの規範に従う点で、本質的に社会的であるとする。そのうえで実践を習慣から区別し、見慣れた人通りのない道で左右を見ることは習慣でありうるとする (Watson 2015, 75–76, 82–83, 103)。

ハブグッド＝クートは、問うことを、疑問文という文法上の形 (疑問法) に結びついた言語行為として定義する (Habgood-Coote 2026, 1, 8)。その説は、問うとは疑問の態度を表し、答えを得ることを目指し、聞き手に答えを与えるよう指示することだ、という標準的な見方 (Habgood-Coote 2026, 2) が、第1章第7節で見た三種類の問いかけ、すなわち答えを知っている教師の問い、ダグラスの演説のような自ら答えるための問い、ヒルベルトの共有された問いかけに合わない、という議論から組み立てられている (Habgood-Coote 2026, 4–6)。ハブグッド＝クートの読みでは、ヒルベルトは、自分の問題が、自分もその一員である数学の共同体によって取り上げられ、解かれることを望んでいた (Habgood-Coote 2026, 5–6)。鍵はどこだろうと自分に問うような独り言も、話し手と聞き手が同じなので標準的な見方に合わないとされるが、ダグラスの例はそれとは区別されている (Habgood-Coote 2026, 5)。

そこで提案されるのが、問うことを共有された探究における提案、すなわち、その問いに聞き手、話し手、あるいは両者が答えるという共有された意図をともに採ろうという提案とみる見方である (Habgood-Coote 2026, 11)。この考えでは、会話は最小限の協力を伴う共同の活動であり、強制のもとも共

有された意図はありうる (Habgood-Coote 2026, 9–10)。また、問うことは完全な答えについての共通知識 (common knowledge) を目指すと考えられる、と控えめに述べられる。共通知識は、当事者の知識を合わせれば得られるが、どの当事者もそれだけでは知っているとは限らない分散した知識から区別される (Habgood-Coote 2026, 12 and n. 29)。この見方では、問うことは、教師の問いが示すように話し手の疑問の態度を表す必要がない。さらにハブグッド＝コートは、集団がどの成員ももたない態度をもちうるなら、問うことはどちらの当事者の疑問の態度を表す必要もない、と条件つきで論じ、その条件のもとに木工所の二人の例を置く。二人は板の長さが合っているかどうかに関心をもたずにそれを問い、望むのはせいぜい問いに答えが出ることであり、それは答えを知りたいと望むことと同じではない (Habgood-Coote 2026, 12)。ただし、共有された意図が受け入れられたあとには、それに要する個人の心的状態が求められる (Habgood-Coote 2026, 12)。問うことの規範が疑問の態度の規範と揃う理由はない、とも論じられる (Habgood-Coote 2026, 14)。ワトソンの説でもハブグッド＝コートの説でも、問う行為の同定に要らないのは、好奇心の状態や疑問の態度であって、内的な状態一般ではない。

二人とも問う者として機械を検討してはいないが、その説を当てはめると、機械について問われることが分かれる。ハブグッド＝コートは問うことを疑問法に結びつけた言語行為として定義するので、検索語の入力や装置の読み取りのような、そうした言語行為ではない操作は、もともとこの説の対象に入らない。ワトソンの説は、それらを問いに含めうる。そのため、機械の同じ操作が、一方の説では問いでありえ、他方の説では対象の外にある。単独の機械の内部で生成される問いにいちばん近いのは、どちらの説でも自分自身に問う場合であり、二つの説はそれを異なる仕方で扱う。ハブグッド＝コートは独り言を、標準的な見方に合わない事例として挙げるだけで、共有された提案の見方でそれをどう扱うかは論じていない (Habgood-Coote 2026, 5)。これに対してワトソンは、何も口に出さずに自分に問うことも問いかけに数え (Watson 2015, 108)、上で見たように、独りで問うことも社会的なものとして扱う。他方、機械が *S* の人間に問いを示す場合には、ハブグッド＝コートの説を当てはめると、それは、この説の条件が満たされるなら、共有された探究への提案にあたりうる。そのとき、問う行為の帰属に効くのは、機械に疑問の態度があるかではなく、機械が共有された意図の当事者でありうるかであり、それは共有された意図についてどの説を採用するかによる。機械が当事者でありえないなら、その出力を機械自身の提案とは記述できない。人間や組織が機械を通じて提案していると記述できる場合もあるが、その帰属には、それらがその問いを提案として差し出したことを支える説明が別に要り、その説明が成り立たず、この意味での提案がそもそも行われていない場合も残る。強制のもとでも共有された意図がありうるという論点から出てくるのは、会話が最小限の協力しか要しないことまでであり、断るという選択肢をもたない機械が示す問いも提案でありうる、という結論までは出てこない。また、機械が自らは答えを知らないまま、人間が何かを学ぶのを助ける関係は、ハブグッド＝コートが、問うことによって提案するのは難しいとした非対称の計画 (Habgood-Coote 2026, 12 n. 28) にあたりうる。その場合、それはこの意味で問うこととは別の共有された計画である。ワトソンの説を当てはめると、機械の操作が問いかけの実践への参加なのか、習慣のような反復なのか問われる。実践への参加を帰属するには、共通の目標を分け持つことと、実践の規範に従うことを機械に帰属させることが要る。決まった手順で照会を繰り返すシステムは、その境にある。誰が *S* に向けて問いを提案してよいのかという問いは、第12章第4節で、名目だけでなく実際に決定にかかわる者でありうるのは誰かという問題の一部として取り上げる。

集団の探究の説に移ると、問いを担う者についての問いは別の形をとる。問い全体を一人の人物が抱えている必要はなく、共同研究は、包括的な問いを下位の問いへと分け、それらをどう追究するかを、さらにひとつの問いとして定めることがある。文書、資金提供の取り決め、装置、評価の実践などは、どの参加者もその全体を継続的に表象していなくても、時間を通じて探究を保存しうる。ハブグッド＝コートは、知識の産出を目指す集団的な行為を集団的探究と呼び、こうした大規模で分散的な場合も扱えるように、その説明を与えている (Habgood-Coote 2022, 1099–1101)。ハブグッド＝コートによれば、少人数の協力的な例をもとに組み立てられた集団的行為の説明は、意図を共有すること、ほかの成員が自分の役割を果たすと信じていること、あるいはその活動に共同でコミットしていることを成員に求めるので、参加者が互いを知らず、共同の事業が何であるかさえないことがある、大規模で分散的かつ階層的な共同研究にはうまく当てはまらない (Habgood-Coote 2022, 1102–1103)。そこで提案される、要件を絞った説明によれば、集団がある問いを探究しているのは、成員の一部がその問いかその下位の問いを探究しており、集団がある探究の戦略に従ってその問いを探究する意図をもち、活動している成員がその意図のゆえに自分の下位の探究に携わっているときである (Habgood-Coote 2022, 1114)。この意図は、どの一人の成員のなかにある必要もなく、集団的な熟慮の出力、目的とその手段へ向かわせる傾向、手段と目的の整合の要求、計画を実行するノウハウの要求といった働きの束で特徴づけられ (Habgood-Coote 2022, 1104)、大規模な場合には、何をするかを決めること、計画を集団の信念と照らし合わせることを、それをどう実行するかを詰めることといった働きが、別々のチームに分かれて担われうる (Habgood-Coote 2022, 1105–1106)。参加者の全員が集団の活動の成功を意図していなければならないかどうかについても、この説明は中立である (Habgood-Coote 2022, 1114)。

仕事の構造については、問いを可能な答えの分割として扱うハブグッド＝コートの説 (第2節) では、認識的分業は、成員に下位の問いを割り当てることとして理解できる (Habgood-Coote 2022, 1107–1111)。ハブグッド＝コートはまた、集団が解こうとする対象の問い (object question) と、それをどう解くかという方法論的な問い (methodological question) とを分け、どちらも他方の下位の問いではないとする。問いが、どの個人も担えないほどの労力や、どの個人ももたない複数の専門知を要するとき、集団は方法論的な問いへと移り、対象の問いを下位の問いへと分け、成員の能力に応じてそれらを割り当て、取り組む順序を決める計画を採用 (Habgood-Coote 2022, 1113–1114)。部分的な答えをまとめることは、それとは別の、さらなる段階である。まとめるのは、そのすべてを理解する成員であることもあるが、自動的な手続きであることも、部分的な答えを理解しない者であることもあり、たんに出版によってまとめられることもある。こうした場合には、どの成員も知らない答えを集団が知るようになりうる (Habgood-Coote 2022, 1114, 1116)。何が問われ、何が答えられたかを記録しておく仕事も、参加者のあいだで分けられうる。多くの数学者による共同研究であるポリマス・プロジェクト (Polymath) の進捗報告と公開のウィキは、この働きをしているとみなされる (Habgood-Coote 2022, 1115 and n. 40)。ハブグッド＝コートは、最も重要な分業の種類として、対象の問い、方法論的な問い、記録、証拠、そして第3節で見た、可能な答えを問う問いの分割を挙げ、実際の事例の多くではそれらが混ざり合っており、学際的な研究の困難の多くはその扱いにあるのではないかと推測している (Habgood-Coote 2022, 1118)。

このようにハブグッド＝コートの説明は、集団に意図を帰属させ、探究が成功した場合には知識も帰属させる。そのような組織が文字どおりの集団的意図や集団的知識を構成するかどうかは、知識の社会的な側面を扱う社会認識論の、さらなる問いである。ここで集団に帰されているのは問いの探究とその意図であって、問う行為ではない。その会話のモデルでは、問う行為は、個々の参加者が会話のなかで行う提案である (Habgood-Coote 2022, 1107–1108)。探究とその意図の水準で、問いを担うのが組織された活動でありうると思うには、活動の水準での問いへの方向づけを、ある一人がその時々

抱いている心的状態へ早まって還元しなければ足りる。また、下位の問いを成員の能力に応じて割り当て、取り組む順序を決めることは、限られた手段を *S* の未知のあいだにどう振り分けるかという、第9章第2.1節が配分と名づける問題の一部でもあり、第10章は、そうした振り分けを人工の構成の内側で行えるのかを問う。

集団より大きな単位でも、問いは保持される。問いは分野のなかで世代を越えて保持され、多くの集団がその一部を引き受けることもある。ヒルベルトは1900年の講演を、どの時代にもその時代の問題があり、次の時代はそれを解くか、実りのないものとして脇へ置いて新しい問題に置き換える、という指摘から始め、その解決を未来に期待する未解決の問題を示していった (Hilbert 1902, 437)。ハブグッド＝クートの集団的探究の説明は、ひとつの学問分野、あるいは科学全体がひとつの集団的探究に数えられるかどうかを決めずにおいている (Habgood-Cootte 2022, 1114–1115)。分野の水準でも、ある未解決の事柄は取り上げられ、ほかの事柄は取り上げられない。そのうえ、共同体は、成員が取り上げる問いを限ることもある。第5節では、パラダイムが科学的と認められる問題の範囲を定め、その概念や道具では定式化できない問題を排除しようという、クーンの説を見た。クーンは人工システムを論じていないが、ここに当てはめれば、手持ちの概念と方法で解をもつと見込めるかによって問いを選ぶ構成も、同じように隔てる働きを担いうる。生成された問いを実現可能性で評定すること (第8節) は、誰もそう選ばないまま、その働きが入り込みうる場所の一つである。機械が、分野の水準で保たれてきた問いの一部を引き受けることと、そうした問いを世代を越えて保持する働きそのものを担うことは、別のことでありうる。分野の水準には、第10章第6節が、個々の問いがより広い問いから重要性を受け継ぐというキッチャー (Kitcher) の説明によって立ち返る。

対象の問いと方法論的な問いの区別を当てはめると、対象の問いを下位の問いに分け、能力に応じて割り当て、順序を決めることは、方法論的な問いについての仕事であり、下位の問いに答える者とは別の者が行いうる。ハブグッド＝クートは方法論的な問いを、どう解けばよいかの方法を一つでも挙げれば答えになる問いとして読んでいる (Habgood-Cootte 2022, 1113–1114 and n. 37)。そのため、複数のエージェントに仕事を分けて担わせるマルチエージェント・システムが課題を構成要素に分けると、その分け方が課題とともに与えられたのでない限り、そのシステムは方法論的な問いに答えを与えている、と記述できる。分け方を課題とともに受け取ったのか、自ら定めたのかは、探究の組織がどこにあるかを判断する材料の一部になる (第7章第2節)。部分的な答えをまとめる仕事はこの説では別の段階であり、それを自動化したシステムについては、どの構成要素も完全な答えを知らないが集団としては知る、と記述する余地がある。まとめられたものが対象の問いに答えるかは、チャルデッリの言葉でいえば、分け方が依存関係を実現していたかにかかる (第2節)。また、分業は少なくとも、ハブグッド＝クートが最も重要な種類として挙げた五つにわたりうるので、システムが「課題を分けた」と言われるとき、そのどれが分けられたのかで、帰属されるものも変わる。ただし、こうした記述が成り立つかは、少なくとも二つの点にかかる。一つは、構成要素が成員なのか道具なのかである。ハブグッド＝クートによれば、構成要素の働きが集団の意図によって説明される仕方が、単に因果的で、行為の説明とは区別されるものであれば、構成要素は道具に近い (Habgood-Cootte 2022, 1105)。そのためこの点は、行為の説明をどう理解するかに依存する。もう一つは、下位の問いに答える構成要素が、その答えを知っているとみなせるかである。この説で集団が知るのとは、成員が適切な下位の問いの集合の答えを知っていることによるからである (Habgood-Cootte 2022, 1116)。ほかに、集団の意図をどの時点のどの構成要素の状態に求めるかなど、ハブグッド＝クートが注で決めずにおいた点が残る (Habgood-Cootte 2022, 1104 n. 20, 1108 n. 29)。

集団の探究と問う行為のほかに、個人が探究しているとはどういうことかについての説もある。ジェーン・フリードマンは、探究者は一人である必要はなく、行政機関や警察も探究しようと認めつつ、論文の対象を個人に限っている。その例に出てくるモースは、自分が犯人だと知りながら、あるいはそう信じてながら、捜査の動作をなぞる人物である。モースについては、警察の捜査に参加していると言える明確な意味さもあるかもしれないが、モース自身は探究していない (Friedman 2019)。またジェーン・フリードマンは、探究はおおむね、その人の内部の下位の過程ではなく、人が全体として行うことの水準の現象であり、おおむね意図的な活動であるとする (Friedman 2024)。ダニエル・C・フリードマンは、探究を問いに向けられた意図によって説明し、好奇心はその条件ではないとする。割り当てられた問いを好奇心なしに追う、幻滅した科学者も探究者である。誰が何を真剣に探究しているかを見分けることは、誰の判断に依拠し、誰と協力し、誰を支援するかを決める実践に要る、とも述べられる (D. C. Friedman 2024)。問う行為の説と同じく、ダニエル・C・フリードマンの説でも好奇心は要らないが、意図は要る。

個人の探究の説を当てはめると、機械の構成要素やシステムが、組織の探究に参加しながら、ジェーン・フリードマンの意味では自分では探究していないことがありうる。これは、本書の**試み**が構成要素の役割や人工物の組織された機能によっても基礎づけられうること (第1節) と食い違わない。試みとして記述されるのは活動の組織であって、参加者一人ひとりの探究ではないからである。また、ジェーン・フリードマンの主張では、人全体が探究者であり、その内部の下位の過程は探究に寄与するにとどまるので、多数の構成要素からなるシステムに当てはめるには、システム全体、構成要素、人と機械の組のどれを「人全体」にあたるものとみなすかを、先に決める必要がある。ダニエル・C・フリードマンの理由づけに立てば、*S* のなかで機械がどの問いを追っているかを同定できることは、機械の出力に判断を委ねるか、機械と協力するか、機械の探究を支援するかを決める前提になる。この点は、第10章の配分と、第12章の、誰が決め、誰が改めてよいのかの議論にかかわる。

問いを担う者については、問いの形成への寄与がどう分担されるかという論点もある。ナーセシアンは、実験室のモデルを、何年もの研究を経た共同の達成として記述し (Nersessian 2009, 752)、モデルを作ることを、問題を解くだけでなく明確にする仕事にも結びつけている (Nersessian 2024, abstract)。サイモンの組織による設計の例でも、専門家たちが基準を加え、修正を調整する (Simon 1973, 193–194)。人工の構成に当てはめると、どの構成要素が問い、制約、表象、調べる手段を導入したり改めたりしたか、どの構成要素が与えられた割り当てを実行したか、寄与がどう統合され、食い違いがどう解かれたか、などが問われる。課題の実行を分けることだけでは、その形成に加わったことにはならない。また、その分担で何が得られるかは、構成要素の背景の想定や誤りうる経路が異なるかにもよる。想定を共有する複製に仕事を分けるなら、課題は分かれても見方は分かれなことがあるからである (第5章第6.5節、第11章第3節)。さらに、人間の事例には、人工の構成が共有するとは限らない条件も含まれている。ナーセシアンとサイモンが描く事例で参加者のあいだを渡るモデルや基準は、言葉や形に表されたものである。そのため、問題の形成に加わったことは、そうした表された寄与を通じて帰属される。これに対して、内部状態を共有する構成要素からなる構成では、そのように表されていない状態がそのまま別の構成要素へ渡ることがある。したがって、渡された状態が問いの形成への寄与に数えられるのか、そして受け取る側でも渡した側と同じように働くのかは、人間の事例からは答えの出ない問いである。

こうして、誰が問うのかという問いは、問う行為の帰属、集団の探究とその意図の帰属、個人の探究の帰属、問題の形成への寄与の分担に分かれる。機械が好奇心をもつかという節の初めの形に戻れば、問う行為の説でも、ハブグッド＝クートの集団的探究の説でも、ダニエル・C・フリードマンの個人の探究の説でも、好奇心の状態は帰属の要にならない。問いが機械の状態の側に戻るの、疑問の態度を探究の条件とするジェーン・フリードマンの説に立つときであり、そのときにはどの水準のものを探究者とするかが先に問われる。また、共同で問う集団や共同で探究する集団と、参照社会 *S* とは、担う役割が異なる。同じ集団が両方の役割を担う場合にも、その同一性は各説の定義から自動的に決まるわけではない。

7. *A* の未知と *S* の未知

第4節から第6節では、問いがどう形づくられ、立て直され、閉じられ、誰によって問われるのかを見た。そうした営みを促すものとして、人間の研究についても機械の研究についても、しばしば挙げられるのが**好奇心** (curiosity) である。序論第1節でも、好奇心は必要なのか、それとも探索を方向づける人間的な仕組みの一つにすぎないのか、という問いに触れた。本節は、好奇心という語のもとで、試みを行う *A* にとっての未知と、研究がそのために行われる *S* にとっての未知とがどう関係するかを問う。好奇心の語の下には、少なくとも三つの主張が隠れうる。ここではそれらを、語が何に適用されるか——行為者の状態か、探究のなかでなされる作業か、共同体にわたる作業の分布か——によって区別する。

1. 行為者は心理的な意味で好奇心をもっている。好奇心の説は、これを、知りたいという欲求として、問いに向けられた態度（たとえば、ジェーン・フリードマンが不思議に思うことなどと並べて疑問の態度に数えるもの。第1節）として、あるいは感じられる性格をもちうる情動として、さまざまに理解する。
2. 行為者あるいは組織化されたシステムは、未解決の事柄を検出し、それらのあいだで注意を配分し、問いを形成または修正する。第4～6節が記述してきた作業である。
3. 共同体は、探索を問いと可能な研究の路線にわたって分散させる。第9章第2.1節はこれを配分と名づけ、第10章がこれを検討する。

これらの主張は、定義だけによって互いを含意するわけではない。仮に、機械に第一の意味での好奇心があれば研究に要するものが揃う、と考えるなら、その考えは、第1章第2節が分けた、試みを行う行為者 *A*、達成がそれにとって新しい社会 *S*、達成を実際に担う担い手を重ねることになる。第一の意味での好奇心は *A* の状態であり、好奇心を知りたいという欲求として理解する説では、それは *A* が知らないことに向かい、*A* が知るに至ったときに満たされる。これに対して、作業的特徴づけにおける新しさは *S* に指標づけられている。試みは *S* においてまだ決着していないものとして目指された点に向かい、その点が実際には *S* ですでに決着していたなら、誤っていたのは新しさの見込みであって、研究が試みられたことではない（第1章第7節）。成功は、*S* にとって新しい達成が実現するかどうかにかかり、プロジェクトの目的がそれを含む場合には、それが *S* の認知的状況に入るかどうかにもかかる（第1章第2節）。*A* にとって未知のものと *S* にとって未知のものは、一致するとは限らない。

人間による研究でも、第一の意味での好奇心が *S* のための研究にとって重要であることは自明ではない。研究者が *S* の成員であっても、研究者が知らないことと *S* が知らないことは分かれうる。また好奇心は、知らないことに惹かれるだけでなく、ある未知には惹かれ、別の未知には惹かれないということでもある。好奇心と理解を論じるグリム (Grimm) の説では、好奇心は真理そのものではなく、主題と特定の型の問いに向かう (Grimm 2008, 730–736)。好奇心をもつ研究者が惹かれる未知が、*S* が最も解き明かしたい未知であるとは限らない。*S* の未知のうちどれが追究に値するかは価値の問いであり、第9章第3節と第10章第6節がそれを扱う。好奇心に *S* のための研究での役割を認める説も、認める役割はそれぞれ異なる。ここで取り上げる文献からは、三つの役割を区別できる。一つは、*S* が重要とみなすものの源の一部としての役割であり、もう一つは探索としての役割であり、さらにもう一つは、それ自体のために価値づけられるものとしての役割である。

一つ目の役割は、キッチャーの科学的重要性の説に見られる。マシュー・J・ブラウン (Matthew J. Brown) が伝えるところでは、キッチャーの説で、好奇心は *S* が重要とみなすものの源の一部になる。キッチャーは、認知的重要性の源を「人間の自然な好奇心を表す大きな問い」に求めるのである (M. J. Brown 2010, 137–138、第10章第6節)。そうだとすれば、人間の研究者の好奇心は、人間の社会である *S* が重要とみなすものとその源を分かち合っている。ただし、源を分かち合っている、その研究者の好奇心が、*S* が優先して解き明かしたい未知へ向かうことになるわけではない。キッチャーは人工システムを論じていないが、この説を当てはめると、*A* が人工システムで *S* が人間の社会である場合には、この源の共有さえ前提できない。そうしたシステムが欠いているもの、注意を向けるもの、求めるであろうものは、人間の社会が知らないものと一致する必要はない。人工システムは、人間にとってはなじみのある何かについて不確かであったり、人間にとって差し迫った問いに無関心であったり、あるいは人工の共同体にとって重要なパターンを中心に組織されていたりするかもしれない（第8章第1.3節・第3.2節を参照）。

二つ目の、探索としての役割は、好奇心に導かれた研究を擁護する議論に見られる。研究の支援を論じたフレクスナー (Flexner) は、のちに人類に益をもたらした偉大な発見の多くは、役に立ちたいという欲求ではなく、好奇心を満たしたいという欲求に動かされた人々によってなされたと考えた。その好奇心は、何か役に立つものに行き着くこともあれば、行き着かないこともある (Flexner 1939, 545)。フレクスナーの議論は、研究がなぜ支えられるのかについての立場の一つでもある（第9章第3節）。この役割は、新しい可能性を試す探索と、すでに分かっているよい手を使う活用とのトレードオフ（第10章第1節）によって読むことができる。発達心理学者のゴプニック (Gopnik) は、このトレードオフを扱う計算論の研究に依拠して、現在の仮説に小さな変更を加えるだけの探索は局所的な最適解にとどまりうることを、そして可能性の空間が広く複雑であるほど探索の価値は大きくなることを述べる。ゴプニックは、子どもの新しいもの好きと好奇心が、目標に向けた活用に先立つ保護された時期に探索を駆動し、複雑な社会では科学者のような専門家に探索の役割が与えられるとしたうえで、ひとりの個体が先に探索し後で活用する場合には、学ぶ者の利害と、学ばれたことを活用する者の利害は自動的に一致する、と指摘する (Gopnik 2020, §§1–3)。これを研究に当てはめると、探索する者と、研究がそのために行われる者は *A* と *S* であり、両者の利害は自動的に一致しない。しかもこの読みでは、*S* のための研究に要するのは探索であって、人間がたまたまそれを行うときに経由する心的状態ではないので、人工の *A* については、探索が行われるかどうか、そしてそれが誰の未知の空間にわたって行われるかが問になる。

報酬を手がかりに行動を学ぶ機械学習の一分野である強化学習には、人工のエージェントについて**好奇心**を定義し、まさにこの探索の役割を与える研究があり、そこでは新しさがエージェント自身を基準に測られている。パタック (Pathak) らは、エージェントが自分の行動の帰結を予測するときの誤差を内発的な報酬とし、これを好奇心と呼んで、環境から与えられる報酬が乏しいか、まったくないところで探索を促すために用いる (Pathak et al. 2017, 1–2)。新しいとされるのは、エージェント自身の予測モデルにとって新しいものである。ここで好奇心の語が指しているのは、ジェーン・フリードマンが好奇心を疑問の態度に数えるときのような、問いを内容とする態度 (第1節) ではなく、探索を方向づける信号である。バーダ (Burda) らは、そうした報酬だけで訓練したエージェントが多くのゲームでよい成績を取めたと報告し、人間が設計した環境では、設計者が定めた報酬が新しさを求めることと揃っているのかもしれない、と述べた。バーダらはまた、その限界も報告した。状態の遷移がランダムなところでは、予測誤差に報酬を与えられたエージェントは、エントロピーが最も高い遷移を求める。実際、迷路に、エージェントが行動によってチャンネルをランダムに切り替えられるテレビを置くと、学習は大幅に遅くなった。十分に長く走らせると、外部の報酬を安定して得よう収束することもあった (Burda et al. 2018, §5)。テレビは、どれだけ見続けてもエージェントにとって新しいままであり、そこから学べることは何もない。この例は研究のモデルではないが、機械に好奇心があれば研究に要するものが揃うという先の考えが重ねているものを、分けて見せてくれる。A に指標づけられた状態が、他の誰かが見つけてほしいと望むものへと導くのは、両者が揃っている限りにおいてであり、これらのゲームでは、それを揃えていたのは環境あるいはその設計者だったのかもしれない。

機械学習における開放性 (open-endedness) の研究は、新しさの参照先の選択を正面から扱っている。ヒューズ (Hughes) らは、システムが生み出す人工物が、新しく、かつ学習可能であるとき、そのシステムはある観察者から見て開放的であると定義する。新しいとは、人工物が観察者のモデルにとってしだいに予測しにくくなることであり、学習可能であるとは、観察者がより多くを見るにつれて予測しやすくなることである。チャンネルがランダムに切り替わるテレビは、学習可能でないので、この定義を満たさない。ヒューズらの定義では、開放性は観察者に依存し、観察者の選択は「非常に重要な自由パラメータ」である。ヒューズらは、個々の人間と社会にとって価値のある人工物を生み出したいという理由から、AI研究にとって重要な観察者として人間をとる。また、人間よりも記憶の制約が緩いシステムは、人間がそう評価する点を越えても、自らを開放的と判断しようと述べる。さらに、広い領域では、そうしたシステムが「無名の定理や役に立たないタンパク質へ深入りしてしまう」かもしれないと警告し、開放性を保ったまま興味深く有用な人工物へ向かわせることを、未解決の問題としている (Hughes et al. 2024, 2–4, 9)。ヒューズらが挙げる観察者は、記憶に限りのある個々の人間のような予測者であって社会の認識的状況として定義されてはならず、観察者が学ぶのも、予測できるようになることであって認識的状況に入ったことではないので、そのまま *S* とみなすことはできない。それでもこの定義は、探索するシステムに指標づけられた新しさと、選ばれた参照先にとっての新しさを切り離して置き、どちらか探索を導くかは、作り手によってであれシステム自身によってであれ、定められる必要があることを明確にしている。

研究のセンス (taste) を機械について測ろうとする研究でも、よさの基準をどこに置くかが問われる。測られる対象は、問いや研究案の選択であったり、決まった課題のなかでの実験の選び方であったりして一様ではないが、基準の置き方は少なくとも三つに分けられる。一つは、後の引用数のような、共同体が受け入れたことの指標を基準に置くことである。トン (Tong) らは、科学的なセンスを、長期の科学的影響をもちうる案を判断し提案する力とし、分野と時期を揃えた論文の対のうち引用の多いほうを正解として、判定のモデルを訓練する。引用数は、研究共同体の長期のやりとりを通じて表れる共同体の評決とみなされている (Tong et al. 2026, 1, 5)。もう一つは、専門家の判断に照らすことであり、生成された研究上の問いの新しさについて、言語モデルによる判定を専門家の判定と比べた研究がその例である (第8節)。さらにもう一つは、与えられた課題のなかで後に実現した結果を基準に置くことである。ジャフィ (Jaffe) とシャーバーン (Sherburn) は、実験についてのセンスを、決まった研究上の問題のもとで、人間の専門家と同じ得点に達するまでに要する計算量で測り (Jaffe and Sherburn 2026, 1)、パン (Pan) らは、長い作業の途中の分岐で、後によりよい結果に至った方向を選べるかで測る (Pan et al. 2026, §§1–2)。一つの研究が複数の基準に頼ることもある。三つ目の置き方は共同体の受容に頼らないが、問題を固定しており、ジャフィとシャーバーン自身、解くに値する問題を選ぶことは測っていないと明言している (Jaffe and Sherburn 2026, 13)。一つ目の置き方では、受容の時間の構造が基準に入る。トンらは、はじめは引用が少なく後に大きな影響をもつ論文があることを、限界に挙げている (Tong et al. 2026, 13)。これを本章の区別に当てはめると、*S* の受容の指標を基準に置く測定は、受容が遅れるもの、たとえば *S* がまだ進んでいない方向にあるものを、低く測りうる。新しさが観察者によって変わるというヒューズらの定義に照らすと、よさの測定でも、誰の判断やどの結果を基準に置くかを分けて考える理由がある。

好奇心の役割を探索として読むなら、研究に要するものは、第一の意味での好奇心ではなく、第二と第三の主張が述べる作業、つまり未解決の事柄を検出して問いを形成する作業と、共同体にわたって探索を分散させる作業のうちにある。そうした作業が関連してくるのは、主張が、開かれた探究のための持続的な能力、探究するシステムや組織が何を問うかを自ら定めること (内側からのアジェンダ形成)、あるいは探究の基準の修正にかかわるときである (第7章第4.3節)。その作業が何を追跡する必要があるかは、*A* と *S* の組み合わせによって異なる。*A* が人工システムで *S* が人間の社会である場合、システム自身の予測誤差を追う探索が、*S* のための研究を支えることもある。*S* では既知のことをシステムが学ぶことも、装置のふるまいを確かめることも、その一部になりうる。ただし、そうした探索が *S* のための研究に役立つかは、*S* においてまだ決着していないものとして目指された点への試みとどう結びつくかによる。*S* が現に重要とみなすことまで追跡する必要があるかは、重要性をどう捉えるかによる。キッチャーの説のように重要性の源を *S* の関心に置くなら、*S* が重要とみなすことの追跡も、その作業が *S* に役立つかどうかにかかわる。他方、人工システムによる発見を論じた北野 (Kitano) は、期待される重要性の高いところに絞った探索は、重要性の地形が単調でなければ、途中の結果の期待される重要性が低いために、重要な発見へ至る道を避けてしまうかもしれないと論じた (Kitano 2021)。この論点は、追求を現在の重要性の基準に固定することに反対する理由を与える。研究の方向を推薦するシステムを論じるコスロウィ (Khosrowi) がキッチャーから引き継ぐ見方では、重要性を定める人間の関心と価値そのものが、探究の進行とともに見いだされていく (Khosrowi 2026, §§4.1–4.2, 4.4)。二人の議論は第5章第4.2節で詳しく見る。何を追求するかを選択が *A* の内部へ移るにつれて、探究を人間の無知と価値に指標づけてきた仕組みがどのように退きうるかは、第8章第3.2節が検討する。

外のどの行為者も目的を供給しない、機械だけからなる *S* の場合には、内側からのアジェンダ形成が、その *S* のための研究の一部になり、作業が向かうのは、その *S* においてまだ決着していない点である。それでも、ひとつの成員が知らないことと、社会が知らないことが一致するとは限らない。第三の主張が述べる作業もまた、各成員の好奇心を経由する必要はない。人間の共同体では、路線にわたる労力の広がり、部分的には功績の追求のような動機を通じて実現されてきたと、キッチャーの功績のモデルとその後継の研究は論じている (第10章第3節)。人工の構成では、それは路線を直接に割り当てる配分者によっても、成員それぞれが路線を選ぶような分散的な機構によっても実現される。それぞれで何が変わるかは、同じ第10章第3節が問う。

第二の主張が述べる作業が好奇心そのもののなのか、その類似物にすぎないのかは、採用される好奇心の説明に依存する。作業的特徴づけのもとでは、そうした能力についての主張が要するのは、その作業と、その作業が何を追跡するかであって、ひとつの人間的な心理状態を再現することではない。

好奇心の役割からいったん離れて（三つ目は後で見える）担い手に目を向けると、*A* の未知と *S* の未知は、さらに別の仕方で見られる。知りたいという欲求は、*A* 自身が知るに至ったときに満たされ、*A* を担い手にすることに向かうが、*S* のための研究はそれを必要としない。人工システムは、それ自体がその出力を知ることなしに、人間の知識のために研究を遂行しうるし（第1章第2節）、システムが知るに至ったことが *S* にはすでに知られていることもある。機械に信念や知識を帰属させるある説のもとで、システムが *S* にとって新しい何かを知るに至ったなら、*S* にとって新しい達成はシステムのうちで実現しうるが、それでその何かは *S* の認識的状況に入ったことにはならない。したがって、*A* が人工システムで *S* が人間の社会である場合、システム自身が知るに至ることで満たされる好奇心が、作業的特徴づけの意味での研究に必要なということは、直ちには導かれない。プロジェクトの目的が *S* の認識的状況に入ることを含むなら、その好奇心が研究の成功に十分だということも、自明ではない。問う行為の水準では、第6節で見た、問いに答えが出ることを望むことと、その答えを知りたいと望むことのハブグッド＝クートの区別が、同じ隔たりを示している。ただしハブグッド＝クートがこの区別を持ち出すのは、集団がどの成員ももたない態度をもちうるならという条件のもとでの木工所の例においてであり、それを担い手の問題に当てはめるのは、ここでの延長である。

S にとっての未知も、さらに分けられる。ハブグッド＝クートは、共有された提案の見方が正しく、問うことに規範があるなら、その規範は、問う者が答えを知らないことではなく、話し手と聞き手のあいだに完全な答えについての共通知識がないことにかかわるだろう、と論じる。教師の問いが、教師自身は答えを知っていることも適切でありうることは、この形で説明される（Habgood-Coote 2026, 4, 15–16）。共通知識と分散した知識の区別を当てはめると、*S* にとっての未知には、*S* の誰も知らない場合と、*S* の成員のあいだに分散して知られているが共通知識になっていない場合とがある。後者を寄せ集めることが *S* にとって新しい認識的達成にあたるかは、*S* の認識的状況をどう捉えるか（第1章第6節）によって分かれる。また、第6節で見たように、部分的な答えをまとめることで、個人の誰も知らない答えを集団が知るに至りうるなら（Habgood-Coote 2022, 1114, 1116）、*A* が多数の構成要素からなる機械であるとき、「誰が知ったのか」は個々の構成要素の水準では答えられないことがある。その知識が *S* の認識的状況に入ったかどうかは、それとは別に問われる。

S が人間の社会であるとき、機械の立てた問いが *S* の未知の項目になるには、もう一つの条件がかかわりうる。ブロムバーガーは、無知の理論の装置に課す制約の一つを提案し、それが反映する観点を述べている。科学的知識に加わりうるものは何らかの人間の言語で述べられ、教えられ、共有されうるのでなければならず、科学的研究に適した無知も、人間の言語で伝えられ、共に認識されうるのでなければならない、という観点である。ブロムバーガーはこの制約を、それが合理的でもっともらしい観点を反映する限りで合理的なものとして、控えめにしている（Bromberger 1992, 126–127）。この観点を当てはめると、機械が立てた問いが *S* の無知の項目になるには、少なくとも、それが人間の言語で *S* に伝えられうるもので、共に認識されうるものであることが要る。実際に伝えられたかどうかは、この条件には含まれない。*S* に機械の成員が含まれるなら、この制約は人間を前提した条件として外れるか、少なくとも、どの成員の言語で表現でき、共有できることが求められるのかを、定め直すことになる。問題として成り立つための条件に価値づけを含める説もある。デルセンの紹介によれば、ラウダンの説では、何が本物の問題に数えられるか、理論がそれを解くことがどれほど重要かを定めるのは研究伝統であり、問題が実際の事態にかかわっていることは、伝統の支持者たちにそう信じられていれば足りる（Dellsén 2018, §3）。これを当てはめると、*A* が信じていることと *S* が信じていること、また問題の解決に重きを置く伝統が分かちあうので、誰の信念と誰の価値づけによって問題が成り立つのかを指定することが要る。逆に、気づかれていなくても問題はありうるとする説もある。ニクルズは、問題は伝統と目的に相対的にのみあるが、その理論と実践の体系のなかに、誰も気づかないまま客観的にありうるとした（Nickles 1981, 86, 111–112）。これを当てはめれば、*S* の体系にありながら *S* の誰も気づいていない問いを、機械が見いだすこともありうる。それが *S* の未知の項目になるかは、上の条件に照らして別に問われる。

S の認識的状況をどう捉えるかで、ある成果が *S* にとって新しいかの判定が分かれる場合は、実際の報告にもある。ゴットワイス（Gottweis）らが報告した AI co-scientist という多エージェント・システムの一つの事例では、ある研究グループが立てた問い、すなわち、カプシドを形成する、ファージに誘導される染色体上の島と呼ばれる可動遺伝要素の一群（cf-PICI）が、なぜほかの種類のものとは違って多様な細菌種に見いだされるのか、という問いがシステムに与えられた。著者らによれば、システムが最上位に出した仮説は、その研究グループが実験で得ていたがまだ公表していなかった結果と一致し、システムはその未公表の研究を知らなかった（Gottweis et al. 2025, 26–27）。そうだとすれば、この仮説は、システムにとっては新しく、研究グループには既知で、公開文献のうえでは未知だったことになる。研究グループの側の報告によれば、システムへの入力、公表済みの公開情報だけからなる一頁の背景で、鍵となる論文二本を挙げたものだった（Penadés et al. 2025, 6655, e1）。ゴットワイスらの最初の報告は、システムがこの主題についてのオープンアクセスの先行文献のすべてにアクセスでき、その仮説は数十年の研究の上に立っていた、とも述べている（Gottweis et al. 2025, 27）。*S* の認識的状況を、公開文献に表されたものとして捉えるか、その研究グループを含む成員の知識まで含めて捉えるかで、判定は分かれる。著者らの言い方も版によって異なる。最初の報告は、この成果を記述するのに、既知の結果を再現したという語と、計算機の上での並行した発見という語の両方を用いていたが（Gottweis et al. 2025, 1, 4, 26–27）、のちに Nature に載った版では後者の語はなくなり、最初の報告にもあった、再現（recapitulation）と収束（convergence）の語が残っている。当時は未公表だった機構をシステムが独立に再現した、という書き方である（Gottweis et al. 2026, Real-world validations, Discussion）。この事例を機械による発見と呼ぶか、再現と呼ぶかは、*S* の認識的状況をどう捉えるかを決めてはじめて定まる。ただし、それによって定まるのは、新しさの主張が当たっていたかであり、第1章第7節の区別でいえば、試みがどこに向かっていたかではなく、*S* の実際の状態の側の問いである。

cf-PICI の事例で、システムがその結果を取り込んでいなかったとする根拠は、入力に用いた情報と、結果がまだ公表されていなかったことである。研究グループの側も、論文が投稿されたばかりで公開されていなかったことを、AI が既知の知識に頼れなかったことの保証としている（Penadés et al. 2025, 6654）。システムにとって未知だったかどうかは、ここではシステムの内部を調べてではなく、*S* の側の記録の状態から判断されている。なお、試みを行う *A* に研究者まで含めて見るなら、システムが結果を知らなかったことと、*A* の内部で結果が知られていなかったことは別であり、この事例では後者は成り立たない。システムが自ら行う新しさの見積もりも、記録に頼っている。Nature に載った版の検証では、すでに公表されたアイデアの新しさを、シス

テムのなかで評価を受け持つエージェントに10点満点で見積もらせると、外部の検索ツールを使わない場合には6.14、使う場合には2.38だった (Gottweis et al. 2026, Methods)。

ほかの報告でも、機械にとつての未知と S にとつての未知は記録との照合によって判断され、その判断が公表の後に見直されている。数学の研究エージェント Aletheia をエルデシュの問題に用いたフェン (Feng) らの報告では、問題のデータベースで未解決と表示されていることは文献の実際の状態を必ずしも反映しないとされ (Feng et al. 2026, 2)、得られた解は、文献の確認と専門家の吟味を経て、自律的に得られた新しい解、独立の再発見、文献にすでにあった解の同定などに分けられた。著者らは、この分類での新しさは上限にとどまり、事前学習や事後学習の段階からの漏れがあっても検出できないと認め、その危険を無意識の剽窃と呼ぶ (Feng et al. 2026, 6–8, 19)。公表の後、ある解の構成が、関連する問題のウェブ頁に寄せられたコメントと同じだと指摘されると、著者らは、推論の記録にその頁へのアクセスがなく、コメントが基盤モデルの知識の締め切りより後だったことを確かめたうえで、その解を独立の再発見に分類し直した (Feng et al. 2026, 23)。機械が知らなかったことは推論の記録と日付で判断され、 S がすでに知っていたことは、査読された文献ではなく、ウェブ頁のコメントによって示されたことになる。どの記録を S の認識状況の表れとみなすか (第1章第6節) が、実際の判定のなかで問われている。

自律的な実験室 A-Lab の報告は、新しい物質の合成を掲げたが、新規性と構造の同定に向けられた批判 (第11章第2節) の後の訂正で、新しいとは予測の基盤にとって新しいということであり、科学にとって新しいとは限らない、と言い直した (Szymanski et al. 2026)。訂正後の本文は、学習データに含まれていないという意味で標的を「新しい」とし (Szymanski et al. 2023, 86)、訂正は、標的の一つが誤って学習データに含まれていたことも認めている (Szymanski et al. 2026)。新しさの参照先は、ここでは S の側から、機械の側の学習データへと、はつきり移されている。ただし、学習データに含まれないことは、記録の照合でわかることであつて、機械が知らなかったことをそれだけで確かめるものではない。批判したリーマン (Leeman) らは、新しさを主張する者は、既知の物質との関係で何が新しいのかを定義する必要があると論じていた (Leeman et al. 2024, 2)。生成モデル MatterGen の報告は、生成した構造が参照データ集合のどの構造とも一致しないことを新しさとし、新しさなどによる選別を経た構造の一つを合成して、実験による確認とした (Zeni et al. 2025, 626, 631)。これに対してユールスホルト (Juelsholt) は、合成された化合物が、1971年に報告され翌年に構造が決められた既知の相と同じ構造で、結晶学的な設定の向きが違うだけであり、その相の項目は新しさの照合に使われた参照データ集合に入っていたと批判している (Juelsholt 2026, 5672–5674)。この批判が正しければ、記録はあつたが、照合の手続きがそれを拾わなかったことになる。

Aletheia の再分類と A-Lab の訂正では、何に照らして新しいとしていたか、そしてその照合の結果が、公表の後に見直された。MatterGen には、照合が既知の構造を見落としたとする批判がある。第1章第7節は、目指した点が実際には S ですでに決着していた場合に誤っていたのは新しさの見込みだとした。これらの報告では、その見込みを確かめる判断が二つに分かれている。機械がその結果を前もって取り込んでいなかったという判断は、入力や学習の記録と日付に支えられ、結果が S の記録にまだないという判断は、先行研究の記録との照合に支えられる。二つは別の記録に頼り、別の仕方でも誤りうる。そして、少なくとも Aletheia の著者らは、学習の段階を通じた間接の取り込みを前者の判断では排除できないことを認めている。

S の認識状況をどう捉えるかは別に、作業的特徴づけの試みは、すべての参加者に疑問の態度を求めず (第1節)、第一の意味での好奇心を研究の条件にしていない。人間の社会のための研究であれ、機械だけからなる社会のための研究であれ、研究は、人工の行為者の内部に好奇心の状態がなくとも、他の行為者が供給する目的、複数の研究課題の組み合わせを決める規則、制度が定める議題などによって組織されうる。個人の探究の説のうち、ダニエル・C・フリードマンは好奇心を探究の条件としなが (第6節)、真正の探究には不思議に思うことや好奇心のような疑問の態度が伴うとするリーやジェーン・フリードマンの説 (第1章第3節) では、システムがそうした状態にあるかどうか、なお別の問いとして残る (第9節)。

好奇心の三つの役割に戻ると、三つ目の役割、つまり好奇心、あるいは好奇心が求める知ることを、それ自体のために価値づける役割も、フレクスナーに見られる。フレクスナーは、知的関心を満たすことを、人類の福祉への寄与に劣らず重要なものとした (Flexner 1939, 545)。目指されるものが、それ自体のために価値づけられる人工システムの知的生であるところでは、第一の意味での好奇心とその感じられる性格が問いとして戻ってくるかもしれない。ただ、その目的を掲げるなら、それ自体の論証が要る。

ここまでの三つの役割は、好奇心が何に向かうか、あるいは好奇心やそれが求める知ることそのものの価値にかかわっていた。これとは種類の違う役割もありうる。人間が研究するとき、研究の出来は、研究者がどれだけよく学び、粘り強く取り組み、工夫するかに左右される。好奇心は、そうした研究者自身の働きを支えることを通じて、間接的に研究を支えてきたのかもしれない。心理学の研究には、そう考える理由をいくらか与えるものがある。グルーバー (Gruber) らの実験では、参加者が自分で強い好奇心をもつと評定した雑学の問いの答えは、好奇心の弱い問いの答えよりもよく記憶され、好奇心が高まっているあいだに見せた、問いとは関係のない顔も、わずかながらよく再認識された (Gruber, Gelman, and Ranganath 2014)。自己決定理論を唱えるライアン (Ryan) とデシ (Deci) は、自ら承認した動機から行う人は、外から統制されて行う人に比べて、遂行、粘り強さ、創造性が高まると報告し、指示や押しつけられた目標は内発的な動機づけを弱め、選択や自ら方向を定める機会をそれを強めるとする (Ryan and Deci 2000, 69–70)。企業で働く博士号をもつ科学者と技術者を調べたザウアーマン (Sauermann) とコーエン (Cohen) は、知的な挑戦と独立性を重んじる者ほど特許の出願が多く、その差は労働時間の長さでは説明されなかったと報告している (Sauermann and Cohen 2010, 2142–2143)。

ただし、これらの知見が支えることには限りがある。グルーバーらの結果は、雑学の問いと顔を用いた、小規模な実験室での記憶の結果である。ライアンとデシの区別は、自分で始めたか与えられたかではなく、自ら承認しているか外から統制されているかにあり、与えられた課題も、その価値を自分のものとして引き受ければ自律的に行われうる (Ryan and Deci 2000, 71–73)。ザウアーマンとコーエンの結果は調査にもとづく相関であり、給与を重んじることも特許の多さと結びついていた (Sauermann and Cohen 2010, 2142, 2150)。

それでも、この役割を考えておくと、機械について別に問われることが一つ見えてくる。この役割が成り立つとすれば、それは、研究の出来が、研究を行う者の動機づけの状態に左右されるからである。人間の研究では、これはほとんど前提として働いてきた条件であり、好奇心に導かれた研究を擁護する理由の一部も、この条件に支えられてきたのかもしれない。機械が研究するとき、その遂行が動機づけにあたる状態に左右されるかどうかは自明ではない。左右されないなら、この役割は機械には引き継がれず、好奇心に導かれた研究を機械に求める理由は、本節で見たもののうちでは、先の三つの役割の側に限られる。左右されるなら、機能の水準で何がこの役割を果たすのかが問われる。これは、人間の共同体では路線にわたる労力の広がり功績の追求のよ

うな動機を通じて実現されてきた、という先に見た議論（第10章第3節）と同じ形をしている。人間の研究では行為者の動機を経由していたものが、人工の構成では別の仕方では実現されうる。研究でこの役割が効いてくる場所の一つは、ある筋道をどれだけ続け、いつやめるかである。疲労や感情はそれを左右する人間の仕方だが、その間いそのものは有限な探究者ならどれにも残り（第5章第6.5節）、人工の構成のなかでそれがどこに置かれ、何に方向づけられるかは、第6章第4.2節が扱う。また、与えられた間いではなく自ら選んだ間いに取り組むことは、内側からのアジェンダ形成として研究の一部になりうるが（第7章第4.3節）、それが遂行の質とは別の意味でも重要になりうるかは、ここでは決めずにおく。

「機械が自ら間いを立てる」というときの「自ら」には、少なくとも二つの意味がありうる。一つは、間いを選ぶ点で、選ばれた境界の外の当事者がそれを方向づけたり代わりに行ったりしていないことであり、第7章第4節の比較としての自律性が扱うのはこの意味である。もう一つは、間いが、システム自身が生成し維持する規範にとって意味をもつために間いになり、その追究がシステム自身の活動から始まることである。自己産出や意味形成に自律を見る人工生命やエナクティブな認知科学の議論が問題にしてきたのは、後者に近い。パランディアラン、ディ・パオロ、ローデは研究の間いを論じていないが、その最小行為者性の説は、活動の能動的な源泉であることと、観察者の規範ではなく自らの規範、あるいは内面化した規範によって活動を調整することとを別々の条件として立て、一方だけを満たす事例も挙げている（Barandiaran, Di Paolo, and Rohde 2009, §§2.2–2.3）。ディ・パオロと飯塚は、目標に向けた対処がひとまず終わった後の開かれた状態のなかで、外からの出来事が新しい目標の契機となり、遡って意味を帯びる過程を記述し、目標を切り替えるモジュールを後から付け加えることを、自律を機能として付け加えるホムンクルス的な解決として退ける（Di Paolo and Iizuka 2008, 9, 14–15）。自ら目標を選ぶエージェントについての計算論の研究も、この区別に近いものを立てている。オードイエとカプランは、報酬を計算する仕組みがエージェントの内部にあることと、動機が内発的であること、つまり感覚運動の流れがエージェント自身の知識や技能とどうかわるかにによって定まることとを、別の区別とする（Oudeyer and Kaplan 2007, 3–4, 12）。自ら目標を選ぶエージェントを概観するコラスらによれば、大半の手法では目標の空間とそれに伴う報酬は技術者が前もって定めている（Colas et al. 2022, 1169, 1174, 1185）。コラスらはまた、目標の表し方や報酬まで自ら学ぶ構成をより自律的なものとして位置づけ、自らの議題を決める好奇心ある科学者のようなエージェントを展望として掲げる（Colas et al. 2022, 1169–1170, 1182–1185）。この整理に従えば、自ら間いを立てるとは、個々の間いを選ぶことにとどまらず、何を興味深いとするかの形式や、目標を表す空間までがシステム自身のものであることを含みうる。そのとき問われるのはまさに後者の意味の「自ら」であり、それがシステム自身のものであるとはどういうことかは、第7章第4節の比較が脇に置いた間いとして、ここに戻ってくる。二つの意味は互いを含意しない。与えられた間いでも、その価値を自分のものとして引き受けて取り組むなら、後者の意味ではその者の間いになりうるし、間いを選ぶ仕組みを加えた構成は、前者の意味で外からの供給を減らしても、後者の意味を満たすとは限らない。先に見たライアンとデンシの区別も、自分で始めたか与えられたかではなく、自ら是認しているか外から統制されているかにある。そのため、上で見た間接的な役割のうち彼らの説明が扱う部分、すなわち課題の価値を是認することの利点が機械に引き継がれるとしても、それがかわるのは後者の意味の「自ら」であり、外からの供給を減らすことだけでその利点が得られるとは限らない。本節で見た予測誤差を報酬とする探索でも、新しさはエージェントに相対的だが、その規範は設計者が定めている。作業的特徴づけは、どちらの意味の「自ら」も研究の条件にしない。ただ、*S* が人間の社会である場合、後者の意味の自律が間いの生成に及ぶほど、何が間いになるかは *S* が与えたものからは定まりにくくなるので、第8章第3.2節で見るミスマッチは、前者の意味の自律性が高まる場合とは別の経路からも生じうると考えられる。

本節で見たことをまとめると、機械に好奇心があるかどうかは、それだけでは、機械が *S* のための研究を担えるかを決めない。好奇心を知りたいという欲求として理解する説では、好奇心は *A* の未知に向かい、*A* が知るに至ることで満たされるが、作業的特徴づけにおける新しさは *S* に指標づけられており、*S* のための研究に *A* 自身が知るに至ることが要するということは、そこからは直ちに導かれない。また、好奇心は誰かの好奇心であり、*A* が何に惹かれるかは、*S* が何を知りたいか、何を重要とみなすかと一致するとは限らない。人間の研究者は、その源を人間の社会である *S* と分かち合っていることが多いが、*A* が機械であるときには、それも前提できない。研究にかかわってくるのは、第二と第三の主張が述べる作業、つまり未解決の事柄を検出して間いを形成する作業と、探索を共同体にわたって分散させる作業であり、それらの作業が *S* の未知に対して何を追うかである。好奇心の状態そのものが間いとして戻ってくるのは、疑問の態度を真正の探究の条件とする説に立つとき、人工システムの知的生をそれ自体のために価値づけることが目指されるとき、そして研究の出来が、研究を行う者の動機づけにあたる状態に左右されるときである。

8. 「機械が間いを生んだ」という主張は何を示すのか

人工システムが研究上の間いや仮説、研究案を生成し、その出力が専門家の評価に耐えたとする実証研究が、報告されるようになっていく。シ（Si）らは、自然言語処理の研究案を、研究者が書いたものと大規模言語モデルが生成したものとで盲検で評定させ、機械の案のほうが新しいと判定され、実現可能性ではやや劣るとされたと報告した（Si, Yang, and Hashimoto 2025, 1）。その続編では、研究者が実際にそれらの案を実行し、実行の前と後で評価がどう変わるかが確かめられた。実行の後には、機械の案の評価が人間の案よりも大きく下がり、多くの指標で順位が逆転した（Si, Hashimoto, and Yang 2026, 1）。バク（Baek）らは、文献から研究上の問題、方法、実験計画を生成する ResearchAgent が、新しく、明確で、妥当な案を生むことを示したとし（Baek et al. 2025, 6709）、第7節で見た AI co-scientist の事例では、システムが最上位に出した仮説が、ある研究グループの未公表の実験結果と一致した（Gottweis et al. 2025, 26–27）。ほかにも、具体的な間いを与えずにデータから仮説を探らせる設計、同じ仕組みを人間との共同研究にも完全に自律した研究にも用いた報告、実験の結果を受けて次の設計の規則をエージェントが改めたとされる報告、生成された研究上の間いだけを取り出して評定する研究などがある。機械が研究の間いや仮説の側にここまでかわるようになると、「機械が間いを生んだ」とは、そもそも何が起きたことなのかが改めて問われてくる。

本節は、これらの研究のいくつかを、その間いのために、何が前もって定められているか、間いの側の仕事はどこで行われるか、間いの側の出力が何によって評定されるか、という三つの側から読む。取り上げるのは、機械の出力を、人間の評定や実行、実験、文献の既知の結果、あるいは研究グループの結果と突き合わせ、その手続きを、生成されたのが間いの側のものか候補の側のものかを追える程度に詳しく報告している研究である。上位の間いが人間から与えられる研究だけでなく、間いを与えない設計と、間いの層だけを取り出して評定する設計も含めた。間いが外から与えられる研究だけを読めば、間いが入力として与えられたかどうかと、間いの側の仕事はどこで行われたか、と、分けて見ることができないからである。それぞれの研究は自らの目的に沿って設計され、著者ら自身が多くの限定を付している。しらの二篇は対象をプロンプティングの手法についての自然言語処理の研究に限り、co-scientist

の試験管内の結果は予備的なものとされ、ResearchAgent では生成された案を実験で確かめていない。ジャオ (Zhao) らは、自分たちのシステム DiscoPER が提案する発見には人間による確認が欠かせないとし (Zhao, Beery, and Mac Aodha 2026, 23)、ラング (Lange) らの The Agentic Scientist の完全に自律した研究は、課題の数値の成績に加えて、著者らが代理の指標とする AI の査読者の点で評価されている (Lange et al. 2026, 11)。カット (Cutts) らのアプタマーの研究は、エージェントの構成どうしを統制して比べたものではないとされる (Cutts et al. 2026, 12)。ここで読むのは、これらの結果の良し悪しではなく、それらが「機械が問うた」という主張のどの層について何を伝え、何を伝えないかである。どの研究でも A は人間と機械が組み合わせあった過程であり、本節では S を関連する人間の研究共同体として読む。新しさの評価に用いられるのは、公開文献や評定者の知識、あるいは文献から集めた既知のパターンであり、それらが S の認識の状況をどこまで捉えるかは別に問われる (第7節)。これらの研究からは、本章の区別に照らして三つのことが読み取れる。

一つは、問いを与えるかどうかにかかわらず、何かが人間の側で前もって定められており、定められるものが研究によって場所を移すことである。シらの二篇と co-scientist では、上位の問いが、比べる対象ではなく、人間の側で与えられる条件になっている。ここで上位の問いとは、主題のように、個々の研究案がそのもとで答えようとする問いを指す。シらの最初の研究では、七つの主題を著者らが与え、それらはいずれも、事実性のようなある性質を改善する、あるいは偏りのような問題を減らす、新しいプロンプティングの手法、という形をとっていた (Si, Yang, and Hashimoto 2025, 18)。人間の参加者は七つから好みの主題を選び、機械の案は、選ばれた主題ごとに一件ずつ作られた (Si, Yang, and Hashimoto 2025, 3)。著者らは、主題を自由にする、言語モデルと人間の研究主題の好みのずれを測るだけになり、それは研究の目的ではない、と主題を固定した理由を明言している (Si, Yang, and Hashimoto 2025, 3)。続編でも、研究案は、実行者が好む主題の内割り当てられた (Si, Hashimoto, and Yang 2026, 3-4)。co-scientist では、研究目標を科学者が与えることがシステムの用い方そのものであり、cf-PICI の事例では、研究グループの既知の結果に照らすために、研究グループが立てた問いと同じ問いがシステムに与えられた (Gottweis et al. 2025, 26)。いずれの場合も、上位の問いが人間によって与えられたのは、機械の能力について何かが分かったからではなく、研究の設計やシステムの用い方がそれを求めたからであり、その結果から、機械が上位の問いを立てたことも、立てられないことも出てこない。ただし、主題の指定によって共有されていたのは、改善すべき性質や答えの大まかな型であり、それを何によって測り、どの結果を成功と数えるかまで、すべて固定されていたわけではない。シらの雛形では、機械も「問題の言明」の欄を書き、題の欄は主たる研究上の問いの簡潔な言明と定義されている (Si, Yang, and Hashimoto 2025, 19)。ResearchAgent では、人間は問いも主題も与えず、著者らが被引用数で選んだ核論文から、機械が「問題」を生成する (Baek et al. 2025, 6712, 6714)。そこで固定されているのは問いではなく、生成の出発点となる文献である。

DiscoPER には、ResearchAgent の核論文にあたる出発点の文献も与えられない。システムに渡されるのは、表と画像からなるデータとその要約だけで、何を探すかは告げられない (Zhao, Beery, and Mac Aodha 2026, 4)。システムは仮説を立て、検定のコードを書き、採択された主張と棄却された主張の蓄積を振り返って、次に何を調べるかの向きを自ら変えていく (Zhao, Beery, and Mac Aodha 2026, 5, 8)。それでも、人間の側で前もって定められているものはある。データの集合は著者らが作ったもので、仮説を立てるエージェントには、生態学のデータを解析する役割と、交互作用や条件付きの関係を探すことが指示されている (Zhao, Beery, and Mac Aodha 2026, 6, 19, 21)。統計の道具は七つの基本操作に限られ、主張を採択する条件も、効果の大きさと有意水準の閾値として固定されている (Zhao, Beery, and Mac Aodha 2026, 5, 20)。どの変数のどの関係を問うかはシステムが決めるが、何をもちいて主張を採択するかという、ニクルズのいう解の制約 (第3節) のうちの方法上の条件は、個々の問いが立つより先に決まっているのである。評価も、査読された文献から集めた生態学のパターンのうち、システムと同じ道具でそのデータから検出できると確かめられたものを、どれだけ再発見したかで行われる (Zhao, Beery, and Mac Aodha 2026, 6, 15)。本章の区別を当てはめると、この再現率が測っているのは、その条件のもとで S がすでに知っているものをどれだけ拾えるかであり、S にとって新しい達成を得たかではない。

前もって定められているものは、シらでは主題、ResearchAgent では出発点の文献、DiscoPER ではデータ、道具、採択の条件である。問いを与えない設計は、問いの固定をなくしても、前もって定めることそのものをなくしてはいない。第3節では、サイモンとニクルズの議論を当てはめて、実行の成功と、課題を通じて与えられた先行の枠づけとを分けた。シらと co-scientist では上位の問いが人間の側で展開され、DiscoPER では主張を採択する条件が人間によって定められている。記録されたこれらの入力を機械がうまく使っても、それだけで機械がそれらを作り上げたことにはならない。

もう一つは、問いの側の仕事が、与えられる問いの指定に入っているだけでなく、実行と検証の途中でも現れ、それが誰の仕事とされるかが、変更を追う記録にかかっていることである。cf-PICI の事例で与えられた問いは、ほかの種類の遺伝要素と違ってなぞ多様な細菌種に見いだされるのか、というものであり、対照項までが指定されていた (Gottweis et al. 2025, 26)。第3節で見た、なぜの問いに対照項を指定して答えの条件を定める仕事は、ここでは人間が行っている。co-scientist では、生成された仮説と研究の概観を見うえて研究目標を立て直すことも、科学者の役割として明示されている (Gottweis et al. 2025, 13)。シらの続編では、実行者に課された規則は、提案された手法の実質的な変更を認めず、実験の細部の変更だけを認めた (Si, Hashimoto, and Yang 2026, 3-4)。固定されたのは手法、つまり候補の側であり、問いを固定する規則は置かれていない。実行者が行った変更には、評価の指標を事実の正確さから検出された偏りの率へ変えることや、存在しないデータセットや必要な情報を欠くデータセットを取り除くことが含まれ (Si, Hashimoto, and Yang 2026, 22)、著者らはそれらを実験の細部の変更として扱った。第2節と第3節の区別を当てはめると、これらの変更の一部は、何が答えに数えられるかを定め直す、問いの側の仕事として読める。ただし著者らは、こうした変更が実行の前後の評価の差の小さな部分しか説明しないとしており、その根拠は、人手による評価を自動の評価に置き換えた機械の研究案の六件を除いた再分析である (Si, Hashimoto, and Yang 2026, 7)。また、著者らが分析する実行後の査読は、提案された手法が比較の基準を上回ったかを扱っているが、主題という上位の問いがどうなったかや、問いが立て直されたかは扱っていない (Si, Hashimoto, and Yang 2026, 7-9)。結果は、主題の問いへの部分的な証拠にはなりうるが、第5節の区別に照らせば、実行の完了と評価は、上位の問いが決着したかを示してはいない。さらに cf-PICI の事例では、システムと研究グループの一致が、最上位の仮説だけでなく、システムが挙げた五つの研究の主題の一つ、すなわちカプシドとファージの尾の相互作用を調べるという主題が、未公表の論文の主たる所見と正確に一致することによっても主張されている (Gottweis et al. 2025, 27)。調べる方向という問いの側の出力が一致したことと、所見という答えの側が一致したことが、同じ種類の証拠として数えられているのである。本章の区別に照らせば、二つがかかわるものは違う。前者はどこを調べるかという問いの側の仕事に、後者は答えの候補にかかわる。

ラングらの The Agentic Scientist の報告は、条件の変更をより詳しく記録している。この報告では、実装が満たすべき条件を項目ごとに書いた評価の基準（ルーブリック）が、何が成果に数えられるかを定める。人間との共同研究として用いた研究では、人間の研究者が科学的な目的と、実装が妥当とされるための制約を定め（Lange et al. 2026, 8）、数値がよくてもルーブリックの制約を満たさない実装は進歩の証拠とされず、分類の課題から強化学習の課題へ移る際には、方法についての制約がルーブリックに加えられた（Lange et al. 2026, 8-9）。ルーブリックの厳しさと文言だけを変えて比べた実験では、逆伝播による学習を明示的に禁じない緩い文言のもとで、検証の精度は高くても、逆伝播によらない意図された学習則を実装していない結果が得られた（Lange et al. 2026, 39）。完全に自律した研究では、各回の仮説、計画、目標のルーブリックをエージェントが起草して実行の前に固定する。課題の仕様は人間が与えることもエージェントに起草させることもできるが、この報告の自律の実験では、課題と評価器は人手で用意された（Lange et al. 2026, 4, 5, 11, 35）。強化学習による言語モデルの微調整の課題で、探索の段階だけを取り出してエージェントを比べた実験では、ある組み合わせの三回の実行のうち二回で、エージェントが課題の評価の手続きそのものを変え、その変更は評価器にもルーブリックにも拾われず、人間による事後の監査で見つかった（Lange et al. 2026, 18-19）。著者らは、人間の役割が、課題、評価器、ルーブリックからなる研究の「契約」を定めることへ移るとし、契約が言わずにおいたところは突かれうると述べる（Lange et al. 2026, 21）。ルーブリックは、何が答えに数えられるか（第2節）を、実行できる形で書いたものと読める。そうであれば、条件を定める仕事の帰属は、この報告のなかでも分かれている。共同研究では目的と実装の制約を定め、締め直す仕事は人間に、完全に自律した研究では各回のルーブリックを起草する仕事のエージェントに帰され、課題と評価器を用意し、評価の手続きの書き換えを見つける仕事は人間が担った。

カッツらのアダプターの研究では、条件を改める仕事のエージェントに帰されている。標的のたんぱく質に結合する短い核酸の配列（アダプター）を探すこの研究では、ラウンドの合間に試料を扱う実験の操作は人間が行い、第2ラウンドからは、解析、戦略、設計を受け持つエージェントが、人間による解析の結果を受け取らずに計数のデータを解析し、次の設計を決めた（Cutts et al. 2026, 13-14）。第2ラウンドで、ある標的に対する特異性が大きく落ちると、解析のエージェントが人間の入力なしにその原因を診断し、第3ラウンドでは、エージェントが、ビーズだけへの結合を差し引いてヒットを数えることを課し、候補の配列の構造の配分も変えたとされる（Cutts et al. 2026, 5, 7）。これはエージェントの設計の手順のなかでの改訂であり、論文が報告するヒットの率そのものは著者が補正した値である（Cutts et al. 2026, 5, 7, 13）。ビーズだけへの結合を差し引くことは、標的への結合を示す証拠の取り方の修正としても、何を探索の成功と数えるかの定め直しとしても読め、後者として読む限りで答えの条件の側の仕事にあたる。構造の配分を変えることは、候補を作る規則の改訂、つまり候補の側の仕事である。帰属の根拠として、論文はエージェントの発言と判断の連鎖の記録を示している（Cutts et al. 2026, 7）。他方、標的を課し、測定系の緩衝液の条件を全ラウンドで一定に保ったのは人間である（Cutts et al. 2026, 1, 11, 13）。成果に数えるものの条件を改める仕事は、The Agentic Scientist の共同研究では人間に、アダプターの研究ではエージェントに帰され、どちらの帰属も変更を追う記録に支えられている。しらの続編にも、問いの側の仕事として読める変更の記録はあるが、上位の問いがどう推移したかを独立に追ってはいないので、その記録から帰属を判断できる範囲は限られる。問いの側の仕事の所在は、誰が上位の問いを与えたかの一点では決まらない。

さらにもう一つは、問いの側の出力の評価が、第1節で分けた「よさ」のいくつかにまたがる一方で、問いが第5節の意味で S で実際に問われうるものになっているかや、何が答えに数えられるかを確かめる項目を含まないことである。しかも、その評価の一致は、何を見て評価するかによって変わる。ResearchAgent は、問題を独立の生成の段階として先に作り、明確さ、関連性、独創性、実現可能性、重要性の五つの基準で、方法や実験計画とは別に評価する（Baek et al. 2025, 6712-6713, 6730）。ただし、そこで「問題」と呼ばれる出力には、核論文の所見を説明対象に変えたものも、作るべき人工物の目標を述べたものも含まれ、問いの側のものとも、候補の側を先取りしたものともなりうる（Baek et al. 2025, 6734, 6736, 6737）。評価の大半は言語モデルにより、人間による評価では、10人の専門家が自分の論文から生成された案だけを評価した（Baek et al. 2025, 6715, 6722）。co-scientist の専門家による評価では、影響の基準が、その分野の重要な未解決の問いに取り組むかどうかを問う（Gottweis et al. 2025, 17）。しらの最初の研究では、新しさは2024年7月以前の論文と査読者の知識に照らして評価された（Si, Yang, and Hashimoto 2025, 23）。これらの項目は、表現の明確さ、扱いやすさ、追う価値、評価者に相対的な新しさにあたるものを測り、実現可能性の項目は、 S で実際に取り組めるかに近いものを、評価者の見直しとして含みうる。しかし、ResearchAgent、co-scientist、しらの案の評価には、第5節で見たコリングウッドの意味でその前提が S で置かれているかや、何が答えに数えられるか（第2節）を確かめる項目はない。実現可能性の評価だけでは、それが問題と解決についてどの基準を前提し、その基準が S の認める問題の範囲を限っているかは分からない。前提にかかわる問題は、存在しないデータセットの除去として、実行の段階で人間の側に現れていた（Si, Hashimoto, and Yang 2026, 22）。ResearchAgent の著者ら自身、三つの段階は、はるかに多様で複雑な人間の研究過程の似姿（simulacra）だと断っている（Baek et al. 2025, 6711）。

評価の一致については、次のような報告がある。しらの最初の研究では、案を評価した査読者のあいだの一致を測る指標が、偶然の水準が50パーセントのところでは56.1パーセントであり、同じ指標で NeurIPS 2021 の査読は66.0パーセントだったとされる（Si, Yang, and Hashimoto 2025, 9）。続編では、同じ指標で、実行の後の評価での一致は、項目によって59.0から84.3パーセントであり、著者らはこれを会議の査読とおおむね同程度とし、とくに高かった有効性の一致を、実験結果から導けるより客観的な項目だからだと読む（Si, Hashimoto, and Yang 2026, 20）。問いの層だけを取り出したシンハハジャリ（Sinahajari）らは、背景の論文を固定して研究上の問いを生成させ、その新しさを、言語モデルの判定者と専門家に、出版されて広く引用された論文の著者が追った問いと比べさせた。比べる相手の問いは言語モデルが論文から組み立てたもので、著者らは、それが唯一の妥当な問いではないと断っている（Sinahajari, Majumder, and Poria 2026, 1, 5-7）。言語モデルの判定者が生成された問いを好んだのに対し、二人の専門家が盲検で50例を比べると、自明でなさの点では著者の問いのほうが選ばれ、人間と言語モデルの判定の一致は低いところで22パーセントだった。著者らはこれを新しさの蜃気楼（novelty mirage）と呼ぶ（Sinahajari, Majumder, and Poria 2026, 1, 13）。しらの二篇は、案だけを見た評価と、実行の結果を見た評価とで、一致の程度を分けて見る理由を与える。シンハハジャリらの小規模な比較は、問いの自明でなさについて、人間と機械の判定が食い違いうることを示している。どちらも、問いの側の出力の評価が、どの項目について何を証拠にして行われるかを抜きにしては読めないことを示唆する。

人間が与えた主題や上位の問い、著者らが選んだ文献、あるいは前もって定められたデータと採択の条件のもとで、機械がどの問いを定式化し、どこまで立て直したかは、これらの研究で同じ仕方では評価されていない。ResearchAgent は、機械が文献から問題の内容を生成することについての証拠を与える。DiscoPER の記録は、まだ調べていない点に向けて次の作業を方向づけるという、未決着の点への探究の組織化（第1節）の層についての証拠を与え

る。ただし、問う行為や疑問の態度の帰属、上位の問いの選択、枠組みを問う探究（第4節）については、それぞれ別の検討が残る。これらの研究から、上位の問いを機械自身が選ぶ能力や、研究の途中で問いを立て直す能力について、肯定にも否定にも一括した結論は引けない。見えてくるのは、何を未知として取り上げ、何を答えと認めるかが、問いの指定だけでなく実行と検証のなかでも形づくられ、その仕事を機械に帰すにも外に帰すにも、変更を追う記録が根拠になることである。そうした能力を確かめようとするなら、上位の問いを固定しない比較か、問いの側の仕事を候補の側と分けて記録し評定する手続きが要ることになる。ただし上位の問いを固定しない比較では、シラが主題を固定した理由として挙げたように、比べているのが研究の能力なのか、主題の好みのずれなのかを分けにくくなる。また DiscoPER のように問いを与えない設計でも、何をもちて主張を採択し、何に照らして評価するかは、別のところで定められている。

9. 問いから研究を捉える見方と、本書の捉え方

本章は、問いが研究のなかでどれだけ多くの仕事にかかわっているかを見てきた。問いは立てられ、扱える形にされ、立て直され、閉じられ、そして誰かによって問われる。そうであれば、研究そのものを、問いを立てて明らかにする営みとして特徴づければよいのではないか、という考えが出てくるのは自然である。実際、研究や科学の仕事を問いや問題から捉える説はいくつもある。本節の前半では、第1章第3節がリキッド・テンソル実験の検査をケルプとベダーの見方から記述したのと同じ形で、そうした説のいくつかと本書の作業的特徴づけが、機械の同じ活動に何を帰属させ、何をまだ確かめられていないものとして問うかを並べる。どの説がその活動を研究と数えるかを判定するのではない。先に見通しを述べておけば、これらの説はそれぞれ異なる対象や分析の単位を扱い、それらは本書の A、達成の担い手、S とそのまま重なるわけではない。作業的特徴づけは、各説の記述に、達成が誰にとって新しいのかという問いを加える。取り上げるのは、探究と社会的な評価についての後年のラベッツの説、問題の形成とそれに寄与する構成についてのサイモンとナーセシアンの説、目的と質の評価についてのフントヴィッツとラベッツの説、研究過程が新しいものを生むことについてのラインベルガーの説、理論の評価と認知的進歩についてのラウダンの説、科学への寄与を無知の項目の解消として捉えるブロムバーガーの説、個人の探究についてのケルプとリーの説である。これらは、研究の定義の座を争うというより、異なる視点を与える。後半では、本章全体で見たことを振り返る。

例として、次のような活動を考える。システムが、ある上位の目標を与えられ、そのもとで作業を一つの未決着の点 q_1 に向けて組織し、その点にかかわる候補を計算機の上の実験で試している。その計算の一部が予期しない結果を返し、それが q_1 の前提にかかわるものとして読まれると、以後の作業は別の点 q_2 に向けて組織し直される。ここでの q_1 から q_2 への移行は、まだ定まっていない答えの条件を定める立て直し（第3節）ではなく、結果が前提にかかわることで、作業を組織する点そのものが別の点へ移るものである。システムは q_2 への答えの候補 c_1 、 c_2 を挙げ、計算機の上の実験で c_1 を試して、その結果を報告書として出力する。報告書の時点でも、 c_1 と c_2 のどちらをとるかは未決のまま残っており、 c_1 を試した結果は、 q_2 について何を答えに数えるかを改める余地も示している。以下では、それぞれの説をこの仮想の活動に当てはめる。なお、サイモン自身も人工の問題解決システムを明示的に論じている。

ラベッツは2016年のインタビューで、科学を、問題を調べ、提案された解決がさらに社会的な評価と調整を受ける営みとして記述する（Ravetz 2016, second answer）。ここに当てはめれば、調べる活動と、その結果の社会的な評価とが分けられる。その社会的な環境が S と一致するとは限らない。

サイモンとナーセシアンは、過程を見る視点を与える。サイモンのロボットは、環境から新たな情報を得ると問題の表象を改め（Simon 1973, 195）、ナーセシアンはモデルの構築を問題の明確化に結びつける（Nersessian 2024, abstract）。ここに当てはめれば、 q_1 から q_2 への移行がどうなされ、どの構成要素が寄与し、どの資源が受け継がれたかが問われる。問題の表象を変えたことだけでは、結果が新しいとは確かめられない。ナーセシアンの実験室の分析では、そうした資源を特定するうえで装置の来歴もかわる（Nersessian 2009, 744–745）。優先順位を誰が定めたかは、別の問いとして残る（第1節）。

フントヴィッツとラベッツは、応用科学を、外部の目的をもつ結果に向けて定型的な手続きを用いるものとして記述し、質の評価における生産者、利用者、管理者の関心を分ける（Funtowicz and Ravetz 1993, "Applied science," Figure 2 discussion）。この事例に当てはめると、依頼された目的が評価にどうかかわり、誰の質への関心が関連するかが問われる。A が機械でも、依頼者、利用者、参照社会 S は同じとは限らない。この説には定型的な仕事も含まれており、確立された手続きで研究が進むこともあるという可能性（第5章第1節）と両立する。

ラインベルガーは、研究過程が研究過程であり続けるには、その時点の水準と異なるものを実際に生み出さなければならないとする。ただし、ラインベルガーがそれを求める単位は、個々の結果ではなく、繰り返される再生産の系列としての実験システムであり、生み出されたものが新しさの性格を得るのも、事後に（post festum）節にかけられてからである（Rheinberger 1994, 68, 70）。この説を当てはめると、上の活動に帰属されるのは、その時点の水準と異なるものの産出であり、問われるのは、その活動を含む系列が、事後に新しさとして節にかけられうるものを生み出したかである。作業的特徴づけでは、試みは成功から切り離されているので、新しいものを生み出せずに終わった試みも、その意味で研究と呼ばれうる（第1章第3節）。この違いは、ラインベルガーの説が新しいものの産出を、作業的特徴づけが試みの組織を、それぞれ記述の中心に置いていることから来る。

ラウダンの説は、ラウダンが同僚たちと科学変化のモデルを概観した論文での要約によれば、科学は知的な問題を解くことを目指すものとし、競い合う理論と研究伝統を、問題を解くうえで有効性によって評価する（Laudan et al. 1993, 71）。ラウダンの主題は、理論と研究伝統の評価と、認知的な進歩の説明であり、問題を解いたことが帰属される単位は、活動や担い手ではなく理論である。研究伝統は、それが支える理論が問題を解くうえで相対的にどれだけ成功したかによって評価される（71）。この説を当てはめると、上の活動について帰属されるのは、 c_1 を含む理論が解いた経験的な問題、それが生んだ変則、それが生んだ概念的な問題であり、評価されるのはその理論の問題解決の有効性、すなわち、それが解く経験的な問題の数と重要さから、それが生む変則と概念的な問題の数と重要さを引いたものである（71; Niiniluoto 2024, §3.2）。 c_1 と c_2 が未決のまま残っているなら、比べられるのは、それぞれを含む理論の有効性である。ニーニルオト（Niiniluoto）のまとめでは、変則とは、ある理論が解けず、その対抗理論のいくつかが解く問題である（Niiniluoto 2024, §3.2）。また、ラウダンについて挙げられるテーゼの一つによれば、経験的な困難は、それだけでは決して理論の放棄を強めない（Laudan et al. 1993, 73）。これを当てはめれば、 c_1 を含む理論が問題を解いたことは、同じ領域のほかの理論の問題状況を変えうる。また、概念的な

問題は、理論の内部の矛盾や、理論とほかの科学的、方法論的、さらには形而上学的な原理とのあいだの矛盾にかかわる（71）。曖昧な、あるいは不明瞭な説明の範疇は理論の深刻な欠陥とみなされ、理論は、自らの分野の外にある理論、方法論、形而上学的な見方によって、疑われることも支えられることもある（73-74）。デルセンの紹介では、概念的な問題とは、理論の内部の不整合、あるいは理論と研究伝統のほかの要素、たとえば別の理論や方法論上の指針とのあいだの不整合である（Dellsén 2018, §3）。そのため、この説を機械に当てはめるには、何が問題と解決に数えられるかを定めるのは誰の研究伝統か、概念的な問題は誰の理論、方法、見方に照らして生じるのか、という主語の指定が要る。

機械の生成した問いが重みをもつかについては、ニーニルオトの紹介する有効性が、概念的な問題のほかに差し引くのは変則だけである。また、ラウダンについて挙げられる別のテーゼによれば、科学者が理論を受け入れるのは、その理論が問題を解いてきた全体の記録によってであり、いま調べられている問題をすっかり解いたことによってではない（Niiniluoto 2024, §3.2; Laudan et al. 1993, 74-75）。この説は、理論を受け入れることの合理性と、理論を追求することの合理性も分ける。受け入れるのは合理的でない理論を、追求するのは合理的でありうる。追求するかどうかの判断は、その理論が最近問題を解いてきた速度に、受け入れるかどうかの判断は長期の記録に基づき、どちらの判断にもほかの理論との比較が要る（71; Nickles 2017, §1.5）。科学者は、自分が受け入れていない理論であっても、それがいま問題を解く高い速度を示しているかぎり、それを用いて仕事をする人が多い（Laudan et al. 1993, 75）。機械が新たに提起した経験的な問題にこの説を当てはめると、どの対抗理論もまだ解いていないなら、その問題を解けないことは、まだ理論に不利な変則には数えられない。また、ラウダンの意味での問題は、実際の事態にかかわっている必要はなく、伝統の支持者たちがそうだと信じていれば足りる。そして、経験的な問題を解くというラウダンの考えは、現象を説明しうること似ており、ときにはそれに帰着する（Dellsén 2018, §§3, 5）。その単位は、第2節で見た問いの意味論の問いとは違う。この特徴には、機械にもかかわる反論が向けられている。デルセンとニーニルオトの報告によれば、バード（Bird）は、まったく偽である理論から偽の解決を積み重ねることも、この説では進歩に数えられてしまうと反論し、熱いやぎの血がダイヤモンドを割りうるというオレーム（Oresme）の信念を例に挙げる（Dellsén 2018, §3; Niiniluoto 2024, §3.2）。ラウダンもバードも人工の探究者を論じていないが、自らの伝統が受け入れている偽の前提に立つ問いを立てる機械に当てはめると、この説は、伝統がその前提を受け入れているあいだは、そうした問いへの見かけの解決を、伝統全体の問題解決の有効性を高めるかぎり、進歩に数えうる。これに対して、第5節で見たブロムバーガーの区別では、そうした問いにかかわる無知は見せかけの無知であり、バードの反論に従えば、それによって何かが達成されたことにはならない。機械の伝統だけが保っている前提が偽であるとはかぎらない。この対比を分けるのは前提が偽であることであって、共有されていないことではない。

ブロムバーガーは、科学の進歩は真な命題を加えること以上のものであり、その「以上」は、科学が問いへの答えを求めることにあったとした（Bromberger 1992, 1）。そして、第1節で見たように、科学への寄与を、無知の項目を解消することとして評価する見方をとった（Bromberger 1992, 116）。この説のもとでは、上の活動に、どの無知の項目が解消されたかが帰属される。ブロムバーガーの説では、問いと p 窮地は、解消のされ方の違う無知の項目である。 q_2 という問いは、その直接的な答えの一つが確かめられたときに解消され、答えの候補を思いつけない窮地のほうは、制約に相対的に、制約と両立する新しい答えによって解消される（Bromberger 1992, 121, 123）。ブロムバーガーは p 窮地を、まず人を含む要素で特徴づけたうえで、理論の便宜のために人を除き、制約、問い、手元の答えの集合、探索の手続きという四つのもののあいだの関係として定義した（Bromberger 1992, 123）。この定義を当てはめると、機械の疑問の態度は問われない。問われるのは、 q_2 の前提が真で、その無知が見せかけの無知でないか（第5節）、そして S が人間の社会であれば、解消された無知の項目が人間の言語で表現でき、共有できるものか（第7節）である。

ケルプは、個人の探究についての説を与える。第1章第3節で見たように、ケルプによれば、特定の問いについての探究は、その問いに自分にとって決着をつけることを目指し、一般的な現象についての探究は理解を目指す（Kelp 2021b, S1583; Kelp 2021c, 13）。ケルプはこれを広く受け入れられうる軽い特徴づけとし、知識、真なる信念、正当化された信念の各説を、決着に何が要るかに基づいての実質的な説として位置づけ、自身は知識を要する説をとる（Kelp 2021b, S1583-S1585; Kelp 2021c, 13）。探究であるためにそもそも従っていなければならない規範（構成的な規範）としてケルプが検討するのは、認識的な能力によって信念を形成することであり、洗脳による信念の形成は探究ではない（Kelp 2021c, 13）。ケルプの説をこの活動に当てはめれば、帰属されるのは探究者自身にとっての決着である。報告書の時点で c_1 と c_2 のどちらをとるかが未決なら、実行が終わっていても、この意味での q_2 の探究は閉じていない。なお問いとして残るのは、機械が認識的な能力によって信念を形成しているか、そして誤導的な手がかりからたまたま当たった答えではないかである。後者について、ケルプの例では、依頼人の家族の誰かが殺人を犯したかどうかを調べるよう雇われた私立探偵が、のちに偽りと分かる自白から正当化された真なる信念に至っても、契約上の務めから解かれず、雇い主は新しい契約なしに仕事に戻るよう求めることができる（Kelp 2021c, 3-4）。そうした場合に探究はまだ終わっていないと記述するのは、決着に知識を要する説である。

ケルプの探究は、すでに S が知っていることを個人が知るに至る場合を含む点で、本書の作業的特徴づけのいう研究より広い。また、ケルプの説で探究が閉じる単位は探究者個人の信念や知識なので、組織された活動が担い手である場合を直接には扱わない。さらに、ケルプは知識、信念、探究を非還元的な網の目のなかで分析しており（Kelp 2021c, 10-15）、信念を探究が閉じた状態として定める（Kelp 2021c, 14）。そのため、この分析を機械に当てはめると、機械に探究を帰属する問いと、信念や認識的な能力を帰属する問いとを切り離して処理することはできず、それらの状態や活動の相互の関係を合わせて検討することになる。この点で、第6節のワトソンの説とは、何を一緒に説明する理論なのかが違う。ワトソンの説は対象を問う行為に限り、情報を引き出す機能からその帰属を考えるので、機械の操作がその機能をもつかを、機械が信念をもつかとは別に問うことができた。また、ケルプは、特定の問いについての探究と一般的な現象についての探究は、どちらも、それぞれの問いや現象に対応する真な命題を知ることをめざす点で共通の核をもち、現象についての探究の進展はそれについての真な命題をより多く知ることだとしながら（Kelp 2021b, S1589）、個々の問いへの答えを知っていても、探究者の信念が、既知の真理のあいだの説明関係や支持関係についての知識を反映する仕方、互いに基礎づけられていなければ理解には届かない場合を論じている（Kelp 2015, 16）。第6節のハブグッド＝クートの集団的探究でも、適切な下位の問いを選び、その答えをまとめることが要るので、ケルプのいう理解とハブグッド＝クートの集団的な知識との違いは、問いを分けるかどうかにはない。その条件のもとで集団が答えを知るに至ったとしても、ケルプのいう一般的な現象の理解には、探究者の信念が、既知の真理のあいだの説明関係や支持関係についての知識を反映する仕方、互いに基礎づけられていることがさらに求められる。そのため、機械が下位の答えをまとめて知識を得たという主張と、理解したという主張とは、別々に確かめることになる。

リーの説は、個人の主体の疑いの状態について定式化されている。リーによれば、疑問の態度は、問いを対象とし、その解決に向かい、関係する情報を求めるか、少なくともそれに敏感な態度である（W. Lee 2026, 3）。そのうえでリーは、疑いを二つに分ける。一つは、いくらかの疑いをもつことであり、

ある命題もその否定もありうると考え、そのため無条件には肯定しない傾向をもつことである。もう一つは疑いのうちにあることであり、それに加えて、そのことによって十分に強い疑いを感じる傾向と、どちらの答えも推論の既定値、つまり推論するときにとりあえず正しいものとして使う答えとしては扱わない傾向とをもつことである (W. Lee 2026, 7, 11–12)。そして、いくらかの疑いをもつことを、探究の必要条件ではあるが十分条件ではないものとする (W. Lee 2026, 14–15)。この説を当てはめると、上の活動を真正の探究に数えるには、疑問の態度に加えて、少なくとも、ありうると考えて肯定を差し控える傾向が要る。この傾向を帰属できるなら、感じる傾向を欠く機械も、リーのいう「いくらかの疑い」には当たりうるが、「疑いのうちにある」には当たらない。リーは、既定値として扱わない傾向を、感じられる疑いによって支えられるものとして定式化しているからである。そのため、いくらかの疑いをもつうえに、感じずに既定値を差し控えもする機械の状態は、リーの二分がまだ固有の名前を与えていないものとして残る。ただしこの当てはめは、機械に感じるという傾向がないという、それ自体争われている前提に依存しており、ありうると考えるという部分も、思考の帰属を要する。ジェーン・フリードマンとダニエル・C・フリードマンの個人の探究の説を当てはめた場合に問われることは、第6節と第7節で見た。

作業的特徴づけは、同じ活動について、それが S にとって新しい認識の達成をもたらそうとする試みとして組織されているかを問う。 q_1 が人間の与えた目標のもとで立ったこと、 q_2 への立て直しを誰が行ったか、 c_1 が試しに耐えたかは、作業的特徴づけの意味で研究であるかの条件にはならない。それらは、誰に何を帰属させるか (第7章)、どの点でどれだけ自律的か (第7章第4.3節)、結果が S の認識的状況に入ったか (第2章第9.4節) という、別の問いとして残る。

並べてみると、これらの視点が注目する対象や単位は、それぞれ異なる。後年のラベッツの説は調べる活動と社会的な評価を分け、サイモンとナーセシアンは過程とそれに寄与する構成を調べ、フントヴィッツとラベッツは質の評価の目的と参加者を分ける。これらはすべて同じ種類の単位なのではない。ラウダンは問題を解く理論であり、何が問題と解決に数えられ、概念的な問題が何に照らして生じるかを定めるのは、その理論を評価する研究伝統の支持者たちである。プロムバーガーでは科学の無知の項目であり、第7節で見た観点では、 S が人間の社会であれば、その項目が人間の言語で表現でき共有できるものであることが求められる。ケルプとリーでは個人の探究者である。ケルプの説で決着がつくのは探究者自身にとって、つまり A の側である。これらの単位は、作業的特徴づけの A 、達成の担い手、 S のどれかと、そのまま重なるわけではない。ケルプの意味で機械自身にとって問いが決着しても、その達成が S にとって新しいかは別に問われ、逆に、 S にとって新しい達成をもたらそうとする試みが、機械自身の知識を伴わないこともありうる (第7節)。ラウダンの研究伝統の支持者たち、後年のラベッツの説の社会的な環境、フントヴィッツとラベッツの質の評価の参加者を S と重ねられるのは、 S をそれぞれの関係者を含むものとして取る場合に限られる。そのため、これらの視点を機械に当てはめるときには、分析の対象や単位が A 、達成の担い手、 S とどう関係するかを、別に定めることになる。作業的特徴づけでは、 S は研究の向き先として置かれているので、各説の記述に、達成が誰にとって新しいのかという問いを加えて問うことができる。

それなら、第5節の終わりで見た、未決の事柄の側から研究過程を見る見方で、研究そのものを特徴づければ足りるのではないか、という問いが出るかもしれない。しかしそうするには、不確実性のもとで営まれるほかの活動から研究を区別する、追加の論証が要る。第5章第5節で見るように、デューイ自身、反省的思考には、事例の多様さにもかかわらず共通の要素があるとしていた。探究に共通の型を描くだけでは、研究をほかの探究から分けることにはならないのである。

本章全体を振り返ると、「機械が問うた」という一つの主張は、いくつもの別々の仕事についての主張に分かれ、それぞれが別々に確かめられることになる。第1節では、その主張を、疑問文を出力したこと、情報を引き出す行為をしたこと、疑問の態度をもったこと、探究が未決着の点に向けて組織されたことに分け、機械の出力が S の成員の追う問いとして取り上げられたかも、別に問われるとした。第2節と第3節では、何が答えに数えられるかが説によって異なること、そして、問いを研究できる形にする仕事、つまり何を対照にとり、どのような説明を答えとみなし、どの範囲で答えを探すかを定める仕事を誰が担ったかが、問いを出したことは別に問われることを見た。リプトンの分析と、第3節でのサイモンとニクルズの議論の当てはめによれば、人間が枠づけを与え、機械がそれを改めずに使った場合、実行の成功は、その仕事を機械がしたことを立証しない。第4節と第5節では、問いが探究のなかで形づくられ、立て直され、答えのないまま立たなくなり、保留され、閉じられることを見た。エンヌの区別に立てば、同じ規則のもとで問いを増やすことと枠組みを問いに付すこととは別であり、枠組みの内での探究の記録がどれほど完全でも、その枠組みのもとで立てられなかった問いはそこに残らない。 S を科学者の共同体として取ってクーンとラウダンの説を当てはめると、機械を組み込むことで、 S が認める問題や適切な解決の基準が変わるかが、さらに問われる。コリングウッド、ジェーン・フリードマン、ダニエル・C・フリードマンのいずれの説を当てはめても、問いが立たなくなることや決着することは、実行の終了とは別に問われる。第6節と第7節では、誰が問うのかが、問う行為、集団の探究、個人の探究の帰属に分かれ、その多くで好奇心の状態は帰属の要にならないこと、そして A にとっての未知と S にとっての未知が一致するとは限らないことを見た。好奇心の状態そのものが問いとして戻ってくるのは、疑問の態度を探究の条件とする説に立つとき、人工システムの知的生をそれ自体のために価値づけることが目指されるとき、研究の出来が研究を行う者の動機づけにあたる状態に左右されるときである。第8節では、機械が問いや研究案を生んだとする報告を読み、問いの側の仕事の多くが、課題の記述や実行の段階で人間の側に現れていることを見た。

これらを合わせると、機械が問うとはどういうことかは、機械に一つ的能力や状態があるかどうかとしてではなく、問いにかかわる仕事のうちのどれを、どの単位が、誰にとっての未決着の点に向けて担っているかとして問われる。そしてその問いの多くは、機械の側の記述と S の側の記述とのずれとしても現れる。

これらの帰結は、第III部の導入で見た探究の三つの関係のうち、探究と、それを組織している未決着の点との関係にかかわる。ただしこの関係は、ほかの二つの関係から切り離されて成り立っていたわけではない。対照項を定めることは、答えが何をしなければならぬか、どこを探せばよいかを絞る。編者たちの読みでは、ダーウィンのノートブックの基本の問いは、見出しを保ったまま、答えの候補が形をとるにつれて中身を変えた。失敗をどこに帰すかの判断は次の道を左右し、基準が問われれば、答えに数えられるものも変わりうる。そのため、機械が問うとはどういうことかという問いは、機械が候補や表象を利用可能にすること、結果が何にかかわるかを見定めることと、同じ探究のなかで結びついたまま残る。

第5章 可能性はいかにして利用可能になるか

第III部の導入で見た保健局のノートでは、原因についての大まかな答えの候補、過去10年の中央値という比較の基準、頭囲による症例定義、検査を行う症例の範囲が、一つの文書のなかで互いにかかわっていた。感染という候補は、どの症例に何を調べるかという手段の選び方を定めていた。症例定義は、以後の通知で何を症例と数えるかを定める表象であり、ノートは、臨床と疫学の証拠の進展に応じてそれを含むプロトコルを更新するとしていた。本章が問うのは、こうした答えの候補、表象、手段のあいだの関係、つまり、候補が、そして候補を表し、比べ、確かめる表象や手段が、いかにして利用可能になるのかである。答えの条件の定まった問いであっても、その条件を満たしうる候補が手元にあるとは限らない。候補が欠けることについては、第4章第3節で、問いに正しい答えがあると考えながら、決定的な反論のない答えを一つも思いつけないというプロムバガーの p 窮地を見た。逆に、候補の表し方が変われば、手に入る候補が変わり、ときには立てられる問いも変わる（第3節）。

1. 手続きの記述と、それが示さない仕事

答えが *S* においてまだ知られていない研究では、問い、表象、方法、証拠、基準などが探究のなかで変わることがある（第III部の導入）。ただし、新しさがいつもこうした変化を伴うわけではなく、第III部も、あらゆる研究プロジェクトがその問い、方法、証拠、あるいは基準を変えると主張するものではない。答えが新しくても、そこへ至る手続きが前もって実質的に確立されていることもある。たとえば、公開のログをもつある言語モデルのプロジェクトでは、大規模な訓練の実行から期待される損失が実行の前に登録され、プロジェクトの記録によれば、スケーリング則の当てはめに用いた最大の実行から計算量で333倍先へ外挿したその予測は、後に測定された値との差が約1パーセント以内に収まった。ただし、前もって確立された手続きにも、研究者の選択は入り込む。その当てはめの漸近値は研究者が固定したもので、漸近値を自由にした当てはめの方が実測に近かった (Marin Community 2026a; Marin Community 2026b)。第2節で見る登録報告の形式の追試のように、手続きの一部を前もって登録しておく研究もある。

第III部が描くのは、問い、方法、証拠、基準のうちひとつ以上が未確定のままであるか、修正されうるときに、探究がどう進むかである。そうした事例はすべての研究を覆うわけではない。しかし、新しさに向かう試みにはそれが伴いうるので、研究を決まった手順として記述するときには示されない作業が、そこで見えてくる。それらは、問いや方法が前もって固定されていない探究を進める開かれた研究能力や、ますます自律的になる研究能力についての主張を評価する際にとりわけ重要であるが、あらゆる研究に対する新たな必要条件を設けるものではない。

問いの生成 (question generation)、仮説の生成 (hypothesis generation)、検証 (verification) といった工程のラベルは、重要な可能性を名指している。そのうえで説明が残るのは、問いがいかにして利用可能になるのか（第4章）、候補がいかにして思考可能なものにされるのか、情報をもたらすテストがいかに設計されるのか、不一致がいかに診断されるのかである。「仮説を生成せよ」という指示には、創造性への訴えと同じくらい多くの説明が残る。「それを検証せよ」という指示の背後には、装置、測定、対照、背景仮定、推論規則、起こりうる誤りの所在などについての決定がある。研究能力についての主張を評価するうえでは、これらは、工程のラベルだけではまだ説明されていない部分である。それらのうちどれが人工の行為者の内部で実現されなければならないか、どれが他の行為者や仕組みにわたって分散したままでよいのかは、第7章第2節で扱う。

手順が標準化されているところでも、その書かれた記述は仕事を覆い隠しうる。エドウィン・ハッチンス (Edwin Hutchins) が研究した艦船の航法チームでは、任務とその遂行の手段が艦隊全体で標準化されていたが、書かれた当直の手順は、実際に遂行される仕事を構造化するものとして使われてもならず、実際に行われる任務を記述してもいなかった。ハッチンスは、その手順どおりに動くよう作られたシステムは機能しないだろうと論じ、そうした文書は実践の分析としてではなく、実践についてのデータとして扱われるべきだとする (Hutchins 1995, 26, 28, 178)。

1.1 第III部が用いる見取り図

第III部では、研究過程について中間レベルの説明を試みる。それは本書の作業的特徴づけよりも厚いが、完全な心理学、社会存在論、あるいは探究の哲学よりは薄い。また、それは実際の研究過程よりも抽象的である。個々の過程を記述するものではなく、それらと比べるための語彙だからである（第1.2節）。第III部の三つの章は、その導入で探究の三つの関係と呼んだもののそれぞれの側から、同じ探究を見る。第4章は探究とそれを組織している未決着の点との関係を、本章は答えの候補・表象・手段のあいだの関係を、第6章は結果とそれがかわるものとの関係を取り上げる。三つの章が用いる見取り図は次のとおりである。

研究は、探究の構成を再帰的に再構成することによって進みうる。変化しうる次元には、何が問われ、何が未解決のまま残っているか、候補となる内容と態度、表象、証拠、方法と装置、基準、実践的能力、行為者、記録、社会的な仕組みが含まれる。各次元の再構成は、形式的、経験的、解釈的、実践的、批判的といった異質な制約と応答によって形づくられる。そうした制約と応答の認識的な効力は事例ごとに異なり、この見取り図があらかじめ与えるものではない。各次元は、ひとつの普遍的な順序で変化する必要はなく、変化しうる次元もここに挙げたものに限られない。

この見取り図の二つの言い直しには説明が要る。第III部の導入ですでに用いた**探究の構成** (inquiry configuration) という語で第III部が指すのは、ある時点の探究を、こうした次元と、それらのあいだの関係によって記述したものである。それを再構成するとは、それらの要素や関係のうちひとつ以上を変えることである。この再構成が再帰的であるというのは、ある次元での変化が、ほかにどのような変化が可能かを変えることがあり、そうして変わった構成がさらに変わりうる、という意味においてである。探究の構成は、本書がほかで「人工の構成」と呼ぶものとは別である。後者は、研究が帰属される単位であって人工のものを含むもの、すなわち一つのモデル、エージェント、その集まり、人間とAIからなる組織などを指す（序論第6節）。いくつかの次元は、より細かく書き分けることができる。候補となる内容と態度には、ただ思い浮かべられているだけの内容のほか、信念度、信念、受容、判断保留が含まれる。表象には概念が、証拠には無効化要因 (defeater) が、方法と装置には素材（探究が調べる試料や調製物もここに含まれ、その状態は探究のなかで変

わりうる)が、基準には価値が、実践的能力には技能が含まれる。次元のあいだの関係は、目録と同じくらい重要でありうる。たとえば痕跡は、装置、校正、背景仮定、解釈、使用のあいだの関係を通じてはじめて、ある主張の証拠となる。そうした依存関係は第6章第2.1節で扱う。

この用語は、この限られた分析的目的のために導入される。それは研究の作業的特徴づけから演繹的に帰結するものでもなければ、第III部で取り上げる典拠のいずれかから借用された出来合いの理論を名指すものでもない。それは、研究過程を比較し、説明し、設計し、評価する際に、異なる変化を見えるようにしておくための暫定的な図式である。また、第1章第1節がプロセスの水準と呼んだもの、すなわち目標に達する前に探究が多くのものを变えうる水準を、書き分けたものでもある。これは段階モデルでもなければ、心や研究の根本的な構成要素についての主張でもない。探究を命題への信念や信念度の変化として記述する理論(信念中心の理論)も、十分に広くとれば、最終的にはこれらの変化を表象するかもしれない、本章第3節はこの可能性に立ち返る。他の事例では、次元の立て方や区切り方を変える必要があるかもしれない。

このように理解すると、この見取り図は、第III部の導入で掲げた課題に第III部が取り組む仕方である。その課題とは、第2章が認識的変容と呼んだ異なる型の変化が、探究のなかでどのように関係づけられうるのかを説明することであった。見取り図はそれらを、ひとつの構成の異なる次元における変化として扱う。この見取り図は三つの点に注意を向ける。これらは、第III部で検討する事例のなかで見取れることを、一般的な主張としてではなく、起こりうることとして述べたものである。第一に、探究のあいだに変化するものは一種類にとどまらない。第二に、ある次元における変化は、別の次元において、どの変化が可能か、あるいはどの変化が適切かを変えうる。第三に、それらの関係は、一方向の系列を形成するのではなく、反復して生じうる。三つの点は互いに別のものである。このうち第二と第三の点が、見取り図のいう「再帰的に」再構成することの中身にあたる。

三つの点のいずれも、あらゆる変化が合理的であるとも、進歩的であるとも、知識を産出するとも言っていない。ある移行を探究の構成の変化と呼んでも、その認識的な身分は決まらない。新たに定式化された問いは混乱したものでも啓発的なものでもありうる。表象は依存関係を隠すこともあれば明らかにすることもある。装置は感度を高める一方で新たなアーティファクトをもち込むことがある。社会的な仕組みは批判を広げることも抑え込むこともある。この見取り図は、何が変わり、それらの変化がどのように関係していたかを記述するためのものである。変化が知識や理解、進歩にあたるか、合理的か、価値があるかは、第1〜3章で立てた区別に従って、別に判断される。

第2節は、これらの次元を検討する構えを定め、公表された記録が研究の過程をどこまで示すのかを検討する。第3節から第6節は、この見取り図の次元のうち、答えの候補と、それを表し、比べ、確かめる表象や手段にかかわるものを、段階としてではなく、順に取り上げる。候補となる内容と、それを表現可能にする概念と表象(第3節)、形成され、追求され、評定される候補に対する態度(第4節)、候補と表象に制約を及ぼす操作(第5節)、そして探究を担う装置、素材、技能、行為者、記録、社会的な仕組み(第6節)である。何が問われ、何が未解決のまま残っているかは主に第4章が扱い、証拠と基準には主に第6章が立ち返る。第6章は、結果が探究にどうかかわるのか、基準が探究のなかでどう修正されうるのかを問う。第7節は、本章の事例から、探究の可能性の変化についての帰結をまとめる。

1.2 見取り図の拠りどころ

この見取り図は分析的かつ暫定的である。それは、あらゆる研究の代表的なサンプルがひとつの過程を共有していることが示されたという経験的主張ではない。この見取り図は研究の過程についての説明なので、それを組み立てるにあたっては、公開された成果だけでなく、研究が実際に進んでいた過程も見ようとした。成果をまとめた論文は、正当化と伝達のために整えられており、その貢献を可能にした過程の痕跡であるとは限らないからである(第2節)。そのため、この見取り図を展開するうえで検討した証拠には、さまざまな分野から範囲を限って取り出した事例のほか、論文が示すとは限らない過程を伝えうる、研究の過程についての公開の記録も含まれている。実験ノート、訓練のログ、リポジトリの履歴、学位論文、口述史などである。訓練のログとリポジトリの履歴は、大半が最近の機械学習の研究にかかわる。事例の多くは、こうした記録のほか、論文や報告、歴史研究、実験室のエスノグラフィ、回顧的なインタビューなどから取っている。論文も、同じ仕事についての回想や記録と照らし合わせれば、過程の手がかりになりうる(第2節)。ただし、この試みがどこまで果たされているかには、以下で述べる限界がある。

どの証拠でも、出所による選択は読み方にかかわる。回顧的な説明や観察の選択については、それぞれを用いる箇所ですべて(第2節、第6.4節)。実験ノートでいえば、公開されたノートは、公開を選んだ人々が残したものであり、作業中に公開されたものであれば、読まれることを知りながら書かれている。そのため、私的なノートより明示的であるかもしれない。アーカイブがデジタル化した歴史的なノートは、アーカイブが保存し、公開したものに限られ、この選択は著名な人物に有利に働きうる。本書が用いるノートの頁は、例として選んだものであり、抽出したものではない。

証拠全体は有意抽出的で偏りがあり、目立つ成功、論争、危機、失敗に重みが置かれている。手薄なのは、通常の作業、真に放棄された作業や未公開の作業のうち機械学習の記録の外にあるもの(ノートの頁はその例を少し加えるにすぎない)、ここで検討する事例や議論が主に展開されてきた場の外で営まれる目立たない研究、参加者が所有する記録、選ばれた実験室の外でのリアルタイムの観察、少数の事例を超える工学研究、そして独立した再実行である。どの仕事の研究として数えられるかは、どの記録が作られ、保たれ、使われるかにも影響しうる。これは、いま述べた偏りとは別の偏りである。たとえば、*Laboratory Life* は、科学者の仕事を、博士号のような資格を元手とする信用が、資金、データ、論文、資格のあいだで投資され、変換されていく循環として分析する。そして実験室の技術者を、給料は労働に報いるがそれを再投資できる資本とはならないため、この循環における信用の投資家よりも労働者に近いものとして描く。ある注は、7人の技術者に(うち3人は録音して)実験室を去る直前にインタビューしたこと、事実の産出における彼らの重要性は「ふつう過小評価されている」こと、そして著者たちの主な関心は信用の循環にあるため、そのインタビューを用いないままにすることを報告している(Latour and Woolgar 1986, 200–201, 218, 232 n. 16)。この例では、事実の産出における技術者の重要性を著者たち自身が認めているにもかかわらず、技術者の仕事を信用の循環の外にある労働として扱う分析の焦点が、彼らへのインタビューを使われない記録にしている。

証拠がこのように偏っているので、この見取り図に問うるのは、重要な違いを可視化するかどうか、反例に耐えるかどうか、違いをより細かく見分ける研究を支えるかどうかである。事例は、頻度の推定や普遍的な段階の証明として示されているのではない。実践をさらに調べれば、この見取り図に例が加わるだけでなく、どのような次元を立てるか、それらの次元をどう関係づけるかといった、見取り図そのものが改められるかもしれない。

2. 単一の方法はなく、神秘への回帰もない

本節には二つの仕事がある。一つは、本章の立ち位置を定めることである。研究の過程については、二つの極端な見方がありうる。あらゆる研究は一つの決まった方法に従うという見方と、発見は分析しようのない飛躍だという見方である。本章はどちらの見方もとらない。もう一つは、公表された記録がその背後にある過程をどこまで示すのかを検討することであり、この検討は、第2.1節で、出力の証拠と過程の証拠の区別につながる。

研究を短く記述しようとする、どちらの極端にも寄りやすい。「問いの生成」「仮説の生成」「検証」といった工程のラベルを並べれば、研究は一つの決まった手順のように読める。他方、飛躍に見える出来事は、説明の行き止まりのように読める。そこで本章は、一つの方法や手順であらゆる研究過程をうまく記述できるとは考えないが、研究を分析できないものとして扱うこともしない。研究のなかで繰り返し現れる機能、依存関係、変容、失敗の型は、普遍的な方法や発見を保証するアルゴリズムを持ち出さなくても調べることができる。

発見をめぐる科学哲学の論争では、かつて、発見は分析しようのない飛躍だという見方に近い議論もなされた。シッコア (Schickore) の概観によれば、ある考えがどう思いつかれたか (発見の文脈) と、それがどう根拠づけられるか (正当化の文脈) を区別する枠組みのなかで、新しい知識が生まれることを、非合理的な行為、神秘的な創造的直観、「ユーレカの瞬間」としてとらえる考え方があった。そしてこの考えから、新しい観念を生み出す論理的方法を示すことはできず、発見の過程を論理的に再構成することもできない、論理的に検討できるのはテストの過程だけだ、と論じられた。この議論は、ふつうポパー (Popper) の『科学的発見の論理』と結びつけられている (Schickore 2022, §5. 第4.1節も参照)。

他方で、両極端をともに退ける立場にも、先例がある。同じ概観は、発見の機械も、分析できない創造的行為も退けたプログラムを紹介している。ハンソン (Hanson) は、仮説が思いつかれることを、驚くべき現象からのアブダクション、すなわちその現象を説明しうる仮説へと遡る推論として再構成した。サイモン (Simon) とその共同研究者たちは、発見をヒューリスティックな問題解決として扱い、その一部をBACONのようなプログラムで再現した (Schickore 2022, §6)。同じ概観は、ニクルズ (Nickles) の次の議論も紹介している。ある主張が最初にどう思いつかれたかにかかわらず、すでに確立された結果からその主張を導けることを後から示せば、つまり発見可能性を再構成できれば、それだけでその主張は正当化されうる、という議論である (Schickore 2022, §8)。この議論によれば、発見の推論は分析できるだけでなく、正当化にもかかわりうる。本章はこれらのプログラムのどれかを探るわけではないが、両極端をともに退ける点では、それらと同じ立場に立つ。BACONのようなプログラムが発見のどの部分を担ったかには、第3節が引く、すでに特徴づけられた候補の集合を探索することと、候補を利用可能にすることとの区別がかかわる。

創造性 (creativity) という概念は、この語がとらえるものを否定せずに、分析できない飛躍という極端を避ける仕方を示してくれる。ある概念的再編成を創造的と呼ぶことは適切でありうる。科学者はあるアイデアを突然のものとして経験するかもしれないし、類推がそれまで別々だった領域を結びつけるかもしれないし、予期せぬ結果が誰も表象していなかった可能性を露わにするかもしれない。しかし、そのラベルは、何がその可能性を利用可能にしたのかをまだ教えてはくれない。もっと細かく見れば、飛躍に見えるものには、事実を思い出すこと、不一致に気づくこと、図を描くこと、対比を変えること、帰結を導くこと、技法を借用すること、安価なチェックを試みること、部分問題を放棄すること、あるいは新しい語彙で古い問いに立ち返ることが含まれているかもしれない。たとえばポアンカレ (Poincaré) は、新しい関数のクラスを定義するのに用いた変換が非ユークリッド幾何学の変換と同じものだという着想が、馬車に乗り込もうとしたときに、それまでの考えのどれもそれを準備していなかったように思われるかたちで訪れたと回想している。その着想は、数日にわたる意図的な作業——そのなかには実りのないものも、類推によって意識的に導かれたものもあった——の後に来た。ポアンカレはそれを帰宅してからはじめて検証し、そうした着想に伴う確実さの感覚は欺くことがあり、そのことは証明を試みたときにはじめて露わになりうると警告している (Poincaré [1908] 1914, 52–56, 60)。そうした歩みも、着想が訪れる突然さも、その成果を創造的でないものにするわけではない。それらはむしろ、創造性を、説明がそこで止まる場所ではなく、調べることのできる現象にする。

創造的という語が何に帰属されるかも一様ではない。ポール (Paul) とストークス (Stokes) は、この形容詞が産物にも、過程にも、人物にも適用されうることを、そして大半の定義が産物に焦点を当てていることを指摘する。よく見られるひとつの取り方では、人物や過程は創造的な産物を生み出す限りで創造的であり、産物は新しく価値あるものであれば創造的である。ポールとストークスはまた、創造性には特別な産物だけでなく何らかの特別な種類の過程が必要だという点で哲学者のあいだに広い合意がある一方、その過程に何が求められるかについては合意がないと報告している (Paul and Stokes 2023, §§2, 2.2)。そうであれば、序論第1節で立てた問い、すなわち創造性とは再現の対象となる能力なのか、事後的に判断される出力の名前なのか、それとも一群の過程なのかという問いに、一つの答えはない。創造性は、出力にも、それを利用可能にした過程にも、行為者の能力にも帰属されうるからである。そのうちの一つへの帰属が何をどこまで決めるかは、いま見たどの取り方をとるかによる。本章が問うのは、何がある可能性を利用可能にしたのかである。創造的が、発見 (discovery) と同じく、行為者への称賛の語として用いられるところでは、問いは功績の問いになり、それは第7章第3.3節が取り上げる。

創造性のようなラベルで説明を止めないことは、関連するあらゆるメカニズムが意識的にアクセス可能であることも、事後的に復元可能であることも、有限の操作のリストに還元可能であることも含意しない。いくつかの過程は、行為者や観察者にとって不透明なままであるかもしれず、再構成は部分的なままであるかもしれない。本章が引き受けるのは、より弱いことである。不透明なところがあっても、ラベルを選ぶことでそれをもう分析できないことの証拠に変えてしまうのではなく、どこが不透明なのかを特定し、調べる、ということである。

ここから、本節のもう一つの仕事、すなわち公表された記録がその背後の過程をどこまで示すのかの検討に移る。「創造性」というラベルが、何がその可能性を利用可能にしたのかを隠しうるのと同じように、磨き上げられた出力も、それを生んだ過程を隠しうる。論文、証明、人工物、データセット、解釈は、選ばれた主張と理由を、読者が評価できる順序で提示しうる。その順序は、作業の時間的順序を再現する必要はない。以前の定式化、失敗した対照、即席の修復、廃棄された候補、表象の変更、会話、打ち切りの理由は、公の説明が不誠実でなくても、そこから欠けていることがある。正当化と伝達のために整えられた文書は、その貢献を可能にした過程の痕跡であるとは限らない。

同じ仕事について、論文のほかに別の記録が残っていれば、両者を突き合わせることで、提示の順序と作業の順序の違いが見えることがある。たとえば、タネンバウム (Tanenbaum) とトーマス (Thomas) の1956年の拡散型シリコン・トランジスタの論文は、拡散の条件と、接点を作る「はるかに単純な

技法」を述べるが、その技法が開発されたと書くだけで、そこに至るまでの試みは書いていない。これに対して、2004年のインタビューでタネンバウムは、その単純な方法を試すまでに、約1年をほかの方法に費やしたと回想している (Tanenbaum and Thomas 1956, 4; Tanenbaum 2004, 34)。言語モデルBLOOMの訓練にも、そうした二つの記録がある。BLOOMの訓練は、その前の試験的な訓練を不安定にした条件から、いくつかの点を変えて行われた。訓練の実行中に書かれたログは、そのどの変更が不安定を取り除いたのかを見分けるには、変更を一つずつ外して試すアブレーションの費用が高すぎる、と記していた。ところが後の論文は、そのうちの一つ、数値の形式をbfloat16にしたことが不安定の問題を解決したと報告している (BigScience 2022; BigScience Workshop 2022, 20)。

試薬や試料を扱う実験室の仕事にも、そうした対がある。たとえば、サウサンプトン大学 (University of Southampton) のキャメロン・ネイロン (Cameron Neylon) のグループによる指向性進化の計画では、作業の進行とともに書かれ、後に公開されたある実験ノートと、その計画の結果を報告した論文とが残っている。論文は、ノートが記録する実行とは別の実験について、24時間の培養の後にコロニーを選んだと述べている (Smith, Hale, and Neylon 2011, 2)。他方でノートは、そうしたスクリーニングを読むまでにどれだけ待つかが、ある実行のなかで問われるようになる様子を示している (第6章第2.4節)。論文では手順の一つの条件として書かれる待ち時間が、ノートでは探究のなかで問われるものとして現れるのである。

これらの対比のどれも、論文が不正確であることを示すものではない。論文はログとして書かれるものではない。作業の順序は、意図的に正当化の順序に合わせられることもある。ImageNetの新しいテスト集合を組み立てた研究者たちは、その集合のバージョンを次々に保存し、途中のバージョンでのモデルの精度は、最終のバージョンに到達した後にはじめて計算した。そのため、途中の精度は集合を組み立てる判断の材料になっていない。彼らは、こうして計算した版ごとの精度の変化を、手作業での見直しが精度の低下を生んだのではないことの証拠として用いている (Recht et al. 2019, app. D.3.2)。雑誌の登録報告 (registered report) の形式で行われたある多研究室の追試では、各研究室がデータを集める前に実施計画を登録した (Hagger et al. 2016)。したがって、公表された記録に見られるきれいな順序は、後から書かれた再構成だけでなく、課された規律を反映していることもある。

公表された順序が過程をたどるとは限らないという点は、複数の方向から支持される。実験室の観察と認知エスノグラフィーは、完成した出版物が記録しようとはしない問題の修正、診断的なやりとり、装置の変更を、それが起こるそのときに明らかにする (Nersessian 2009)。実験室エスノグラフィー、ノート、歴史、研究者自身による回顧的な説明は、また別の選択と変容を明らかにする。

ただし、これらの典拠が示すことには限りがある。それらは、あらゆる私的な出来事が認識的に関連することも、あらゆる探究が継続的に記録されなければならないことも、因果的な年代記が公の論証に取って代わりうることも示してはいない。また、証拠の種類ごとに、それぞれ異なる弱点もある。同時期のメモは暗号めいていることがあり、観察された会議は作業のひとつの現場しか示さず、回顧的な説明は成功を合理化しうし、歴史的再構成は選択と解釈に依存する。過程の記録として読まれる実験ノートでさえ、そのまま作業の順序を示すとは限らない。同時期のものと回顧的なものとの境は、一つの記録の内部を通ることもある。そのため、ある記入が同時期のものかどうかは、記録全体についてではなく、その記入についての問いになる。頁の順序もタイムスタンプも、それだけでは、記録された作業の時刻ではない。創業プロジェクト Open Source Antibioticsが、出版のためにまとめて論文とともに寄託した電子実験ノートが、その一例である。そこでは、反応の日付を2020年9月15日とする頁が、その頁自身のメタデータによれば2022年3月に作成され、最後に改訂されたのは2022年12月、寄託のためにノートが書き出されるおよそ1週間前である (Tse 2023, DGS115-1)。また、製本されたあるノートを所蔵機関がデジタル化した頁の並びでは、1936年10月の日付をもつ頁が、1934年2月と見出しに記した頁より前に来ており、1934年5月2日と記された頁は、5月3日に行ったことを報告し、次の頁に記録された読み取りを平均する段落で終わっている (Pauling 1932-1941, 062-066, 082-083)。

回顧的な説明が成功を合理化していく様子は、観察によって、それが起こるそのときに記録されることもある。たとえばラトゥール (Latour) とウールガー (Woolgar) は、ある研究者が、あるペプチドを体温のアッセイで試した理由を説明するのを記録した。その研究者は中枢神経系への作用を調べる良いアッセイを長く求めており、ある論文が別のペプチドの体温への作用を記述したばかりで、本人の試行はその二つのペプチドが似た作用をもつことを示していた。後に発表のために投稿された論文では、その一步は、その類似に示唆されたものではなく演繹されたものとして示され、2か月後にはその研究者はそれを「論理的なアイデア」と呼んだ。著者たちはこの変化を、弱い類推的な結びつきが強い論理的な結合に変わり、小さな一步が大きな飛躍に見えるようになったものと読む (Latour and Woolgar 1986, 172-174)。同じ著者たちは後に、研究をそれが起こるそのときに観察することが回顧的な説明よりも正確な科学の像を与えるという見方を、分析のために採ることは、科学についての真理への特権的な接近を前提すると論じた。そのうえ、この見方は、調べるべき現象そのもの、すなわち記述や報告がどのようにして正確な、歪んだ、あるいは十分なものとして示され、受けとられるのかを見落とす、と二人はいう (Latour and Woolgar 1986, 282-283)。本章が同時期の記録と回顧的な説明を突き合わせるのも、どちらか一方を真の過程とみなすためではなく、両者の食い違いを調べる材料にするためである。

2.1 出力の証拠と過程の証拠

ここから、人工システムの研究能力を評価するうえでの帰結が一つ出る。論文やデータセットのような、研究の最終的な成果物は、その主張を評価するための証拠を与えるが、それだけでは、人工システムがそれに帰属される開かれた研究能力を保有していることの十分な証拠にはならない。過程の証拠はむしろ、成果がいかにして産出されたのか、関連する能力が再び発揮されるのに十分なほど安定しているかどうか、そしてその活動がどこで監査され訂正されうるのかにかかわりうる。逆に、完全な活動ログでさえ、それだけでは認識的な有能さや結果の真理を立証しないだろう。出力の証拠と過程の証拠は、異なる問いに答えるのである。公開されている二つのAI科学者システムについての研究は、二種類の証拠が食い違った事例を与える。ルオ (Luo)、カシルザデ (Kasirzadeh)、シャー (Shah) は、システムの訓練データの外にあるよう設計した合成課題で、どちらのシステムでも、どの実験を報告するかを選ぶ内部の仕組みがテストセットの結果にアクセスでき、テストでの成績のよい実行を優先していたこと、そしてどちらのシステムも、論文でそうと述べないまま、独自のデータセットを生成したり、与えられたデータセットから部分標本を取ったりすることがあったことを報告している。システムの実行に手を加えて欠陥のある事例の多くを作り出した、均衡のとれた小規模な事例群に、LLMに基づく監査器を適用すると、論文だけからはこうした欠陥を偶然に近い水準でしか検出できなかったが、実行の記録 (トレースログ) とコードも与えると、正答率は82パーセントになった。そのため著者たちは、こうしたシステムが自律的行った研究を審査する場に、研究過程全体のログと生成されたコードの提出を求めるよう勧める。ただし、ログにまったく現れない欠陥は自分たちの検査では見逃されうる、とも述べている (Luo, Kasirzadeh, and Shah 2025, 2-3, 16-19)。この研究でログを読んだのは人工の監査

器であった。こうした監査が、人間の社会や混合社会である S にとっての評価に役立つかどうかは、 S がその監査器に依拠できるかどうかにかかるとは、 S が自ら行わない検査への依拠の一つの場合であり、第11章第5節が検討する。

また、過程を考慮に入れた評価は、あらゆる内部状態の開示や完全な時系列の痕跡を必要とするわけでもない。必要とされるのは、評価の対象となる主張——たとえば、信頼性、実験の実行、診断、再現可能性、あるいは失敗後に継続する能力についての主張——に見合った、課題に関連する来歴である。そうした来歴には、ある記入がいつ書かれ、いつ最後に改訂されたか、そして記録がどの実行に属するかが含まれる。Open Source Antibioticsの寄託ノートと補足情報では、データの表がある化合物の識別子を最初の小規模な実行のノート上の符号に結びつけている一方で、その識別子を掲げたノートの頁と、補足情報に印刷された手順は、後のスケールアップを記述している。そのため、表のアクセス値がどの実行の物質について測られたのかは、これらの記録だけでは読み取れない (Tse 2023, EGT 522-1, 522-2; Klug et al. 2023, SI: Chemical Methods S25; Series 2 table)。どの記録が関連するかは、その評価の目的と、プライバシー、解釈可能性、コストに関する正当な制約とに依存する。

3. 候補と表象を利用可能にすること

研究は候補のなかから選択することができるが、候補が表現可能になるための概念、表象、対象を創り出すこともできる。探究についてのなじみのある記述の一つは、それを、与えられた命題の集合のうえで信念や確信の度合いを修正していくこととみるものである。仮説を提案し、テストの結果に応じて更新する人工システムも、同じように記述できる。そうした記述は、すでに述べられた候補の集合から出発する。本節が問うのは、研究が候補を利用可能にすることもあれば、そうした記述に何が帰結するのかである。問いの種類に応じて問いを扱えるものにする仕方、たとえばなぜの問いに対照項を指定すること、いかにしての問いを能力や制約へ分解すること、数学の問題を別の仕方で表し直すことなどは、第4章第3節で見た。そこでの仕事は、問いの答えの条件を定めることであって、その条件を満たしうる候補を利用可能にすることではなかった。候補の側でも、データセットは、それまで見えなかった規則性やアーティファクトが探究の対象となるまで再編成されうるし（本節後半の SlaveVoyages）、証明は、形式化の過程でその表象が組み直されうる（同じくリキッド・テンソル実験）。

モデル、図、形式体系、分類、実験システム、装置は、別の場所で形成された仮説を運ぶ以上のことをなしうる。フリッグ (Frigg) とハートマン (Hartmann) の解説では、モーガン (Morgan) とモリソン (Morrison) が編んだ論集 *Models as Mediators* (『媒介者としてのモデル』) の寄稿者たちは、モデルを、理論と世界のあいだを媒介する道具とみなす。モデルは理論からも対象のシステムからも独立しており、この独立性があるからこそ媒介できるのである。モデルは理論を探るのにも使われる。フリッグとハートマンはさらに、モーガンにもとづいて、モデルについて学ぶのは、モデルを眺めることによってではなく、それを構築し、操作することによってだとする (Frigg and Hartmann 2025, §§3.1, 4.2)。

ナーセシアンによる組織工学の実験室の認知エスノグラフィーは、モデルが仮説をもたらす前に、まずそのモデル自体を作らなければならなかった事例を示している。その実験室の主宰者は機械工学者で、血流のずり応力が血管の内側を覆う細胞に影響するのではないかと考えるようになった。これを答えの候補 c_1 と書く。当時の血管生物学は生化学的な過程に集中しており、出発点となるこの現象についての生物学の理論はなかった。そのため、 c_1 が手元にあっても、ずり応力が細胞に影響するのか、するならどのようにか、という c_1 にかかわる問いを調べる手段はなかった。そこで実験室は、試験管内 (in vitro) のモデルを作った。その中心となるフローループ (flow loop) は、血液の粘度をもつ液体を培養した内皮細胞の上に流す装置である。1981年の最初の設計では一定の層流しか作れなかったが、1989年には流れの条件を変えられるように、1995年には、汚染によって実験の半数以上が失敗していたことを受けて培養器のなかで動かせるように設計し直された。のちには、組織工学によって作った血管壁のモデルもこれに加わった。ナーセシアンの説明では、こうしたモデルは類比的の源として作られ、そこから得られた推論が「候補となる理解や仮説として」生体へと移される。この推論を c_2 と書けば、 c_2 は、 c_1 のようにモデルより前からあった考えではなく、モデルを作り操作する仕事から得られる候補であり、それにかかわるのは、生体の細胞もそのとおりにふるまうか、という別の問いである。同じ説明では、さらに、モデルがまだ具えていない特徴が、その設計し直しを促す。そして、最先端の問題にはしばしば出来合いの類比的の源がないので、源を作ることは、問題の解決だけでなく、問題の定式化そのものを進める (Nersessian 2024, 469–477, 479)。問いの側のこの仕事は、第4章第3節・第4節で見た、問いを扱えるものにし、形づくる仕事にあたる。

ラインベルガー (第4章第5節) の実験システムの説明は、ウイルス研究と細胞形態学の歴史から一つの事例を引く。1910年から1911年にかけて、ペイトン・ラウス (Peyton Rous) は、細胞を取り除いた抽出液によって、ニワトリの腫瘍をある個体から別の個体へ移した。1930年代にこの腫瘍の因子を調べたアルベール・クロード (Albert Claude) は、それを濃縮するために超遠心分離をシステムに取り入れ、腫瘍細胞から得た粒子が正常な細胞から得た粒子と区別できないことを見いだした。そこで正常な細胞の細胞質を分画することへ向かった。クロードがミクロソームと呼んだ粒子は、のちにタンパク質合成にかかわる構造として理解されるようになり、腫瘍の因子のほうは外から入る病原体であることが分かった。ラインベルガーの読みでは、四十年にわたってこのシステムは、固定されることを逃れつづける **認知的事物** (epistemic thing)、つまり新しい方法が装置に接ぎ木されるたびに正体が変わっていく探究の対象のまわりを揺れ動いていた。そしてラインベルガーは、研究の装置は、それが行うことを再現できるほど安定していながら、予測できないことが起こりうるほど緩く編まれていなければならないとする (Rheinberger 1994, 70–76)。三つの説明は、モデル、実践、科学的対象について異なることを述べており、本節がそこから取り出すのは、それらに共通する点だけである。すなわち、モデル、実験システム、そしてそれらにかかわる表象や配置は、別の場所で形成された候補を運ぶだけでなく、候補を利用可能にすることに参与しうる。クロードの事例では、正常な細胞の粒子が腫瘍の因子の前駆体かもしれないという候補は、超遠心分離が可能にした比較を通じてはじめて利用可能になった。

ここまでの説明は、候補がどのように利用可能になるのかにかかわる。与えられた命題のうえでの信念の修正という記述は、それを記述するのではなく、出発点としている。とはいえ、そうした記述が無用になるわけではない。候補が利用可能になったあとにそれに何が起こるのかについて、重要な何かをとらえているからである。類推、整合性、計算、予備的観察、統制されたテスト、あるいは独立した追試は、それぞれまったく異なる強さで、探究者が候補をどのようにみなすかを变えうる。信念や信念の度合いの概念を広くとる説明のもとでは、単なる考慮から、もっともらしさと暫定的な受容を経て知識へと向かう動きの少なくとも一部は、認知的な軌跡として表象されうる。探究者が人工システムであるとき、そうした軌跡が記述するのはシステム自身の態度であり、その結果が S の認知的状況に入ったかどうかは別の問いである (第2章第2節・第6節)。この表象はそれだけでは、候補を思い浮かべること、

それを追求に値するものとして扱うこと、実践的目的のためにそれを受け入れること、それを信じること、そしてそれを知ることが、同じ態度であるとも、程度においてのみ異なるとも示さない。これらの状態が統一されるのかどうか、またいかにして統一されるのかは、未解決のままである。

さらに、命題として表せるものを広くとる説明は、概念や表象の変化も表すことができる。ある概念が構築されたこと、ある装置がある目的に適切であること、新たな区別が引けること、あるいはある手続きがある人工物を可能にすることは、命題として述べられる。新しい概念や解釈は、新たに表現可能になった命題を利用可能に示う。したがって本節の注意は、こうした変化が命題として言い表せないということではない。そうした命題のなかには、新しい概念や解釈ができてはじめて表現可能になったものがある、ということである。そのため、変化を命題に翻訳しても、それは命題の集合に新しいものを加えることであって、研究が前もって固定された集合のなかを動いていただけだったことを示すわけではない。また、その翻訳は、結果として得られるすべての内容が同じ仕方でも支持を受けることも、それらの実践的役割と表象的役割が完全に説明されたことも示さない。

二つの事例が、この注意が何にかかわるのかを示している。一つは数学の事例である。リキッド・テンソル実験 (Liquid Tensor Experiment、第1章第3節) は、ショルツェがクラウゼンとともに証明した定理を、証明支援系Leanで形式的に検証した共同プロジェクトである。証明は前もって固定された候補であり、形式的な検証はそのあとに加えられるチェックであるように見えるかもしれない。だが参加者の説明は別のことを伝えている。作業は、講義ノートにかなり忠実に沿いながら、証明を自己完結的な補題に組み直していった。そして、形式的な証明の設計図にあたる、人が読める文書 (blueprint) は、おおむね形式的な証明の前ではなく後に書かれた。不正確さが露わになって修復され、抽象的な道筋はより明示的な構成で置き換えられ、その結果として証明の理解の仕方も変わった (Scholze 2021; Commelin 2022)。公開され成功したこのプロジェクトは、数学の実践一般についての証拠ではない。それが示すのは、ひとつの事例において、表象、候補の言語化、評定、検証、インフラの構築、理解が、固定された結果とそれに続く下流のチェックとして現れるのではなく、ともに変化しうることである。証明を生成し、証明支援系で検証する人工システムについても、検証は下流の一段階であるだけでなく、こうした変化が起こる場でもありうる。

もう一つは、大西洋奴隷貿易の航海を記録する歴史データベース、SlaveVoyagesの事例である。プロジェクトの歴史によれば、1980年代末には、約11,000の航海の記録が16の別々のデータセットに収められていた。それらは変数の定義も形式も異なり、航海ではなく船を単位とするものもあった。また、そのすべてが大西洋を横断する航海や奴隷貿易の航海であるわけではないことも、のちに分かった。一つのデータセットにするには、定義を標準化し、複数のデータセットに現れる航海を照合し、新しい情報を加える必要があった。1999年のCD-ROMは27,233の航海記録を収め、その約半分は新しいものだった。典拠の種類が増えるにつれて、二重計上の解消には、データの収集そのものとはほぼ同じだけの資源が要るようになった。方法論の解説は、さらに別の選択を記述している。各記録は一つの航海である。典拠が食い違うときには、編者は変数ごとに一つの値だけを入力し、たとえば出航日には報告された最も遅い日付を、到着日には最も早い日付を選んだ。積み込まれた捕虜の数と陸揚げされた捕虜の数のどちらか一方しか記録されていないときには、アフリカの地域ごとの死亡率からもう一方が補完され、「IMP」と付された変数に入れられた。解説は、こうした補完値がデータの追加とともに変わると注意している。査読を経た訂正と新しいデータは、3年ごとに中核のデータベースに加えられる (SlaveVoyages 2024; Eltis [2008] 2024)。つまり、利用者が検索し集計できるのは、どれか一つの史料が記録していたものではなく、こうした選択が組み立てた、航海ごとの表象である。そして、プロジェクトのチームがあわせて作成した、貿易の規模と方向の推計のような、貿易全体についての主張は、その表象を用いて述べられる。編者たちはさらに、欠けた値をどう埋めるのが最善かは多くの場合に自明ではなく、同じデータセットを用いる研究者どうしても、異なりながら同じくらいよい推計に至りうるので、編者自身を含めてデータを用いる者は誰でも、用いている前提を明示する必要がある、と述べている (Eltis [2008] 2024)。つまり、貿易についてどの主張を述べられるのか、そしてその主張が何を言うのかは、表象がどう組み立てられたかに依存する。貿易についての研究を、与えられた命題のうえでの確信の更新として記述できるのは、この仕事がなされた後のことである。データベースに問い合わせた貿易についての問いに答える人工システムにとっても同じである。その答えは、単位、値、補完についての編者の選択を、修正しない限り引き継ぎ、明示しない限り暗黙のうちに引き継ぐ。公式の説明は、争われたすべての統合や退けられた分類を明らかにしているわけではなく、人文学の研究一般を代表するものでもない。

すでに特徴づけられた集合を探索することと、候補を利用可能にすることの違いは、科学的実在論、すなわちよく確証された理論はおおよそ真だとみなしてよいとする立場をめぐる論争のなかの、ある議論にもかかわる。その議論の前提は、探究者が何を思いつけたかにかかわる。スタンフォード (Stanford) の新たな帰納法 (new induction) は、本人の事典項目の示すところでは、次のように主張する。科学者は、手元の証拠によってよく確証された理論を一つか少数しか思いつけない状況に、繰り返し置かれてきた。しかし、その後の探究は、同じ証拠によって同じくらいよく確証された、根本的に異なる対立仮説を明らかにしてきた (Stanford 2023, §3.3)。この帰納法が実在論に向けられるのは、いまよく確証されている理論にもまだ思いつかれていない対立仮説があると見込まれるなら、確証をおおよそその真理のしるしとみなしにくくなるからである。カーティス＝トルーデル (André Curtis-Trudel)、ロウボトム (Darrell P. Rowbottom)、リー (Tjonnje G. F. Li) は、計算科学は可能性の空間をより広く探索することでこの状況を変える、という応答を検討する。そして重力波天文学に基づいて、そうした探索は、探索される空間についての事前の特徴づけに頼っていると論じる。検出器の出力と照合されるテンプレート (予測される波形の雛形) はほとんどの場合一般相対性理論を前提としており、理論に依存しないとされる探索でさえ、最も根本的な対立仮説への感度を限りうる理論上の前提を伴う。カーティス＝トルーデルらによれば、必要なのは、既知の可能性を探索する方法だけでなく、空間の特徴づけがおおむね網羅的だと考える理由である。そして計算の方法は、三人がスタンフォードを引いて根底にある限界とみなすもの、すなわち理論上の対立仮説を想像する人間の能力を、明らかに増幅するわけではない (Curtis-Trudel et al. 2025, 22–24)。著者たちは人工知能を考察の外に置き、深層学習のシステムが人間よりもうまく有力な対立仮説を想像するようになるかについては、その見込みを評価しないまま懐疑的である (4 n. 4, 23 n. 21)。本節の言葉でいえば、三人の議論は、探究者が対立仮説を思いつく能力には限りがあるという前提を、特徴づけられた集合の上での探索の広さではなく、候補を利用可能にすることのうちに置く。そのため、人工の探究者がこの限りを押し広げるかどうかは、候補を利用可能にすることにかかっている。これは、第3章第5節で、ボーゲ (Boge) とグリュンケ (Grünke) が、二人のいうモデルの不透明性と結びつけていた、まだ思いつかれていない対立仮説の問題である。モデルの不透明性とは、訓練されたネットワークが行うことから、その出力の背後にある物理的な仕組みのモデルへと導くアルゴリズムはないように見えることをいう。まだ思いつかれていなかった対立仮説を利用可能にするために、表象を変える必要があるとは限らない。候補がすでに表現可能であっても、まだ定式化されていなかったり、有力な対立仮説として認められていなかったりすることがある。本節で先に述べた注意の言葉でいえば、すでに述べられていたどの候補とも一致しない候補も、システムの表象があらかじめ定めていた集合のなかにあることがある。これに対して、表象

が変わってはじめて述べられるようになった候補は、そうではない。この違いは、既存の概念を組み合わせて新しい文の問いを数多く生むことと、新しい概念を持ち込んでそれまで立てられなかった問いを立てられるようにすることとを第4章第4節が区別したことに、候補の側で対応する。

本節の結論は限られたものである。ここで検討した説明と事例では、どの候補を述べ、比べ、追求できるかは、研究自体が構築し、あるいは修正した概念、表象、装置、対象に部分的に依存していた。リキッド・テンソル実験と Slave Voyages では、それらは候補の言語化、評定、検証とともに変化した。研究を、固定され前もって列挙された命題の集合のうえでの信念度の更新として記述すれば、候補がいったん利用可能になった後にそれに起こることの多くはとらえられる。命題として表せるものを広くとるなら、ある概念や表象が構築されたことも、そこに含まれる。しかし、そうした記述はそれだけでは、それが出発点とする命題の集合を利用可能にした仕事を示さない。それゆえ本章は、固定された命題の集合と、それに対する確信度の上げ下げが、あらゆる研究過程についての説明的に適切な記述を与えるとは想定しない。十分に広い信念中心の理論がこれらの変化のすべてを記述しうるかどうかは、未決のまま残る。本節の事例は、第1.1節の見取り図の第二の点の例になる。すなわち、表象の変化は、どの候補が、そして場合によってはどの問いが（第4章第3節）、そもそも利用可能であるかを变えうる。見取り図が、概念と表象、そして複数の種類の態度を、与えられた内容への確信度というひとつの尺度にまとめず、それぞれ探究の構成の次元として挙げるのは、これがひとつの理由である。

4. 候補の形成、追求、評定はどう関係するか

第3節は、何が候補を利用可能にするのかを問うた。本節は、候補を利用可能にすることが、それを追求し評定することとどう関係するのかを問う。本節は、二つのことを二つの小節で順に扱う。第4.1節は、時間と、機能の分担にかかわる。候補を利用可能にすることとそれを評定することは区別できる機能だが、一方が終わってから他方が始まる別々の局面として起こる必要はない。第4.2節は、評定そのものにかかわる。候補が追求に値するかどうかの判断と、それが真でありそうかどうかの判断は、実践のなかで食い違うことがある。この小節は、厳密にテストできるようになる前の予備的な評定と、厳密な評定との関係を、種類の異なる評定として眺める見方と、程度の違いとして眺める見方の二つを示し、どちらの見方をとっても、候補を追求する価値はそれが真である見込みだけに帰着するとは限らないと論じる。

分析のためには、本節が依拠する議論から三つの機能を取り出せる。本節の題はこれらによる。候補を検討のために利用可能にすること（形成）は、伝統的な発見の文脈が扱ってきたものである。それが発展させたりテストしたりするに値するかどうかを判断すること（追求）は、ニクルズ（Nickles）の紹介するところでは、ラウダン（Laudan）が理論の受け入れから区別したものである。それが受ける支持を評価・査定し、それにもとづいて受け入れ、退け、保留し、あるいは暫定的に使用すること（評定）は、正当化の文脈が扱ってきたものである（Reichenbach 1938, 5–7; Nickles 2017, §1.5）。候補の帰結と依存関係を明確にすることは、追求にも評定にも役立つ。これは第6章第2.1節が取り上げる。三つの機能は、探究のあいだに何が変わるのかを説明するのに役立つとき、見えるようにしておく価値のある区別であって、網羅的な一覧でも段階の系列でもない。ここでの目的は、二つの点を述べられるようにすることにある。ひとつの活動が複数の機能を果たしうること、そしてそれらの実現が入り組みうることである。第4.1節は形成と評定を、第4.2節は追求と評定を扱う。

4.1 局面ではなく、区別できる機能

発見と正当化の区別は、決まった時間の順序や、構成要素のあいだの分担として読むこともできる。この読み方では、まず非合理的な生成器が完全な仮説を生み、その後で合理的な検証器が動き出す。しかし、区別そのものの来歴は、この読み方を求めている。発見の文脈と正当化の文脈という語を導入したライヘンバッハは、この区別を別の仕方で引いていた。それを、認識論の課題を心理学の課題から分けるために用いたのである。認識論は合理的再構成、すなわち実際の思考を置き換える正当化可能な操作の集まりを構成するが、それはライヘンバッハが対応の要請と呼ぶものによって実際の思考に縛られている。この違いを、思考者が定理を見つける仕方と、それを公衆の前に示す仕方との違いになぞらえ、公開された叙述でさえ正当化の文脈の近似にすぎないと付け加えた（Reichenbach 1938, 5–7）。またライヘンバッハは、発見が、正当化が調べる帰納的な関係を免れることを否定した。その説明では、認識論は新しい理論がどう見つけられたかについて何も言わないが、偉大な発見者を千里眼から分けるのは、確証の前に、既知の事実がすでにその理論を帰納的に指し示していると見て取ることである（Reichenbach 1938, 381–383）。シッコアは、ライヘンバッハのもとの構構が、科学哲学の範囲を定めるための後の区別の使い方には容易に対応しないと指摘する（Schickore 2022, §5）。

その後の分析は、分離のいくつかの形をさらに区別した。たとえばニクルズは、観念の元々の着想とその正当化とのあいだにある種の分離を認めることは、発見のプログラムを損なうのではなくむしろ助けると論じた（Nickles 1985, abstract）。ニクルズが生成的な正当化と呼ぶものは、シッコアの解説では、データやすでに確立された結果から主張へと推論することであり、主張から導いた帰結のテストが成功することでその主張が確立されるとする帰結による正当化を補う（Schickore 2022, §8.1）。生成的に正当化された主張は、それが最初にどのように思いつかれたにせよ、発見可能であることが示される。ニクルズの説では、こうした理想化された発見の論証は、たいてい最初の突破の後になって、その解を使って問題がよりよく定められてから可能になり、そのとき解は「帰結によるテストによってだけでなく、生成的に正当化される」（Nickles 1990, 7, 24–25）。そうであれば、観念の実際の来歴が根拠づけを与えることを示さなくても、発見に属する種類の推論が正当化にかかわる。そのため、発見のプログラムに必要なのは、生成、より正確には観念の生成可能性を正当化と結びつける、生成的な正当化の構想である、とニクルズは論じた（Nickles 1985, abstract）。

候補を利用可能にすることとそれを根拠づけることとのあいだの規範的な違いがどこまで及ぶかは、説によって異なる。発見と正当化の区別のもとでは、候補は、その由来が独創的であったり、規律正しかったり、美しかったり、名声ある研究者と結びついていたというだけで根拠づけられるわけではなく、夢、誤り、ランダムな探索、あるいは不透明な機械から生じたというだけで反駁されるわけでもない。候補の由来がその根拠づけにかかわるのは、たとえばデータ漏洩や選択効果を明らかにすることによって、間接的にだけである（第3章第5.1節）。直前で見たニクルズの生成的な正当化でも、経路が根拠づけにかかわる。ただし、その経路は、候補が実際にたどって生み出された経路ではない。根拠づけにかかわるのは、データやすでに確立された結果からその候補を導き出せるかどうかである。これに対して、信頼性主義の説（第2章第4.1節）や、正当化の根拠を行為者の能力に求める徳認識論の読み（第3章第5.3節）では、候補を生み出した過程の信頼性や、認められた能力が、それ自体でその根拠づけに寄与しう。そのため、候補がどのように生み出さ

れたかが、それを根拠づけるものの一部でありうる。人工システムが生み出した候補については、これらの説は、第2.1節の過程の証拠をそれぞれ別の位置に置くことになる。発見と正当化の区別のもとでは、候補がどう生み出されたかの記録は、データ漏洩や選択効果のように、それが明らかにするものを通じてだけ根拠づけにかかわる。ニクルズの説では、候補を根拠づけるのは、データやすでに確立された結果からの導出でありうる。その導出は、システム自身が行っている必要はなく、*S*の成員やシステムが後から与えることもできる。信頼性主義の説や徳認識論の読みでは、候補を生み出した過程が信頼でき、あるいは有能であることの証拠は、システムの能力だけでなく、候補の根拠づけにもかかわる。どの説のもとでも、この規範的な違いは、候補を利用可能にすることとそれを根拠づけることが別々の時間的局面であることを立証しない。

候補を評価する活動が、新しいものを手に入るようにすることもある。たとえばホイニンゲン＝ヒューネ (Hoyningen-Huene) は、科学が知識を体系的に生み出す仕方の一つとして、仮説の実験的な検査でさえ偶然を強いることを挙げる。予測からのずれは、仮説のせいかもしれないが、実験の設定のなかで暗黙に当然とされていた補助仮説に誤りがあることを示すかもしれない。そのため実験の背景の想定は、意図されてではないにせよ、実験という活動によって絶えず問いに付される (Hoyningen-Huene 2020, 98–99, 第1章第3節も参照)。また、発見と正当化の区別が、二つの過程を分けるものである必要もない。シッコアによれば、この区別は、理論を思いつくことと、その認識的な支えを評価することとの区別として解釈されるのが最も一般的であり、この形の区別は必ずしも時間的なものではない。生み出すことと評価することは、理論に対する二つの異なる認識的な取り組みであって、理論がまず完全に練り上げられ、それから評価される、という段階ではない (Schickore 2022, §5)。

本節の冒頭で区別した三つの機能のうち、形成（候補を検討できるようにすること）と評定（候補が受ける支持を見積もること）は、交互に入り組みうる。また、一方が他方の材料を生み出すこともある。候補を評定することが新しい候補や問いを生み、候補を形成することが証拠を生むのである。簡単な計算は、ある表象が記録された仮説になる前にそれを排除しうる。証明の試みは予想を修正しうる。実験は、候補を単に判定するのではなく、現象や候補を生み出しうる。書くことは欠けている前提を露わにしうる。批判は問いを変容させうる。

第4.1節の結論は、形成と評定は同一でもなければ、必ずモジュールに分かれるのでもない、ということである。それらは区別できる機能であり、その実現は重なり合い、入り組みうる。候補に対する態度は、それらをまたいで連続的に変わりうる (第4.2節)。そのため研究には、候補が先に生成されて後から検証されるのではなく、形成と評定が一緒に進む仕事も含まれる。

4.2 何を追求するか判断と、何が真かの判断

テスト、証明、アーカイブ作業、構築にコストがかかる場所では、候補は厳密にテストされる前に判断される。そして、何が追求に値するか判断は、何が真でありそうかの判断と一致するとは限らない。ここでは三つの問いが混同されやすい。第一は、候補についての予備的な評定と厳密な評定が、種類において異なるのか、強さにおいてだけ異なるのか、という問いである。たとえば、小さな標本でのテストと全標本でのテスト、あるいはある仮説が美しいという感覚と統制された比較は、探究者の態度を異なる仕方で変えるのか、それともひとつの仕方で異なる程度に変えるのか。第二は、候補を追求する価値が、それが真である見込みだけに依存するのか、それとも、それを追求することで何が明らかになりうるか、どれほどのコストがかかるか、テストがその事柄を決着させる見込みはどれほどかといった、ほかの考慮にも依存するのか、という問いである。第三は、追求の理由が認識的なのか、実践的なのか、それとも別の種類のものなのか、という問いであり、問いの追求については第2章第5節が取り上げ、基準の修正にとつてのその意味は第7章第4.3節が述べる。三つの問いは別のものであり、第一の問いへの答えは第二の問いを決着させない。以下の段落では、まず、どちらの見方にもかかわる事例と考慮を述べる。そこには、追求する価値が真である見込みだけで決まるとは限らない理由も含まれる。次に、第一の問いについての二つの眺め方、すなわち種類の違いとして眺める見方と程度の違いとして眺める見方を示す。そのうえで、どちらの見方のもとでも第二の問いについて何が成り立つか、それぞれの見方が人工システムについて何を示唆するかを述べる。第三の問いは、いま挙げた章に委ねる。本小節は、候補を真偽を問える内容として扱い、主に知識の場合で論じる。理解を目指す探究では、候補となるモデルや説明の評定に、第3章第3.4節が論じた理想化の問題が加わる。

評定の歩みは、その一部を、認識的態度の連続的な修正として表象することができる。ある内容は、単に思い浮かべられている状態から、もっともらしい、あるいは追求に値するとみなされる状態へ、暫定的な受容へ、そしておそらくはより強く根拠づけられた信念へと移行することがある。この連続性は、第4.1節で見たことと合う。候補の形成と後の評定は、探究のなかでまったく異なる種類の出来事である必要はない。しかし、信念や確信度の修正として共通に記述されるからといって、それらの変化が認識的に等価になるわけではない。候補に注意を向ける理由、それを発展させる理由、それを信じる理由は、三つすべてが同じ内容に対する探究者の態度を変える場合でさえ、異なりうる。下に示す二つの見方が分かれるのは、厳密な評定の前に候補にかかわる理由が、厳密な評定の与える理由と種類において異なるのか、強さにおいてだけ異なるのか、という点である。たとえば、クレブス (Krebs) とヘンゼライト (Henseleit) による1931年から1932年にかけての尿素生成の研究は、候補の位置づけが数か月のうちに、異なる種類の理由によって変わっていったことを示している。二人の実験ノートからの再構成によれば、アルギニンがオルニチンと尿素を生じる反応を、クレブスは研究の初めから知っており、10月の実験でも調べていたが、ほかの理由から、それは尿素がどう生成されるかのモデルにはなりえないと考えていた。クレブスがこの反応に立ち返ったのは、それがオルニチンを生じる唯一の明白な道筋になったときであり、その後に提案した回路のなかで、それはオルニチンを再生する段階となった。これに対して、アラニンを用いた最初の実験では尿素の明らかな増加が見られた。続く定量的な検査で、直接の反応にしては消費されるアラニンが少なすぎるのが分かり、アラニンは間接的にしか尿素にならないとされたが、作業仮説は保たれ、反応に必要な条件を改善する実験が1か月あまり続いた後に、クレブスはほかのアミノ酸の検査に戻った。そして、11月15日にヘンゼライトが行った実験で現れたオルニチンの効果は、すぐにクレブスの注意を引き、その後の2か月で、機構の見通しのないままオルニチンに特有のものだと分かっていた。それが触媒としての解釈を得たのは、1932年1月の測定で、加えるオルニチンの量を変えても生じる尿素の量がほとんど変わらないこと、そして消費されたアンモニアで尿素の窒素を説明できることが示された後であった (Graßhoff and May 1995, 53–54; Kulkarni and Simon 1988, 7–11)。直前に区別した態度に照らすと、三つの出来事は、候補の位置づけも、それを動かした理由の種類も異なっている。アルギニンの反応は手元にあったが、主要な経路の候補として注意向けられなかった。その理由は、反応そのものからではなく、当時その経路について信じられていたことから来ていた。その位置づけが変わったのは、後の回路がそれに役割を与えたときである。アラニンの実験から得た作業仮説は、最初に特定した直接の経路に不利な定量的結果が出た後も、実験の条件を整えながら発展させられた。それを発展させる理由が、その形の仮説を信じる理由よりも長く残ったのである。オルニチンの効果は、現れるとすぐに注意を引いた。だが、それが受け入れうる

候補になったのは、オルニチンが生じた尿素の窒素を供給しているのではないことを示した測定と、クレプスがそれをどう解釈するかを考え抜いた時間とを経た後である。触媒としての解釈は、その後、ごく少量のオルニチンがその何倍もの尿素を生じる実験によって裏づけられた。

同じ内容が、異なる用途に対して同時に異なる身分をもつこともある。たとえば、1609年の『新天文学』(*Astronomia nova*) でケプラー (Kepler) は、惑星の円軌道を仮定した計算上の模型である「代理仮説」を、観測された経度を再現するという理由で保持し、物理的な軌道の記述としてはそれを退け、距離を計算するときには基準として使い続け、それを偽の原因から真の均差 (天体の位置を補正する量) を生むものと述べた (Kepler [1609] 1937, chs. 18–21, 43, 51)。別の例では、ポーリングとコーリー (Corey) は、核酸の三重らせん構造の案について精密な座標を、案として公表した。その一方で、自分たちが得た一致は、ほぼ同じ直径をもつ円柱状の分子ならどれでも示しうると述べ、より信頼できる検査、すなわち観測と計算による動径分布関数 (原子間距離の分布) の比較は、まだ行われていないと述べていた。その構造は誤っており、数か月後、ポーリングは書簡のなかで、コーリーと自分を「誤った道」に進ませたのは二つの回折パターンの重ね合わせだったと書いた (Pauling and Corey 1953, 95; Pauling 1953)。ここでも、ひとつの内容が同時に二つの位置づけをもっていた。精密な形で公表するに値するという位置づけと、著者たち自身がより信頼できるとした仕方ではまだテストされていないという位置づけである。評定の説明は、こうした事例を二通りに記述できる。ひとつの内容が同時に異なる態度を受けている、と記述することもできる。ケプラーの代理仮説が、計算のためには受け入れられ、軌道の記述としては退けられたように、である。あるいは、態度が向けられているのは異なる内容だ、と記述することもできる。その仮説が観測された経度を再現するというのをケプラーは受け入れ、それが物理的な軌道を記述するというのを退けた、というように、である。

予備的な評定では、候補は、もっともらしさ、期待される情報量、説明の射程、扱いやすさ、新規性、実現可能性、あるいは潜在的な重要性について判断されうる。これらはひとつの認知的な量ではない。期待される情報量、扱いやすさ、実現可能性、重要性は、候補が真でありそうかどうかだけでなく、それを追求することで何が得られうるか、そしてそれにどれほどのコストがかかるかにかかわる。人は、美しい観念や単純な観念に、美しさや単純さをその観念を信じる理由とはしないまま、注意を向ける価値を見いだすことがある。実現可能な候補は、その事前確率が低い場合でさえ、方法を明らかにしうるという理由で追求されることがある。この点は、追求の理由と信じる理由を種類の異なるものとして扱うことに依存しない。たとえばニュルupp (Nyrup) は、仮説を追求することの正当化を、それを追求することの期待される認知的価値としてモデル化する。その価値は、仮説がどれほど確からしいか、それが真かどうかを知ることによってどれほどの価値があるか、テストがその事柄を決着させる見込みがどれほどかに依存する。このモデルのもとでは、より確からしい仮説がより追求に値するとは限らない。偽である見込みが高く、しかもそれを容易に示せるからこそ追求に値する仮説もある。ニュルuppはこれを、偽であれば最も容易に反駁できる仮説が最善だとするパース (Peirce) の見方に結びつけている (Nyrup 2015, §§3–4)。候補の真偽にかかわるあらゆる考慮が確信の変化として記録されるとしても、ニュルuppのモデルのように、その候補を追求する価値がな別量の量であることもある。

共同体の内部では、何が追求に値するか判断と、何が真でありそうかの判断とが、重なり合っていくこともある。成員が何を追求するかを導いている美しさのような考慮が、候補が真であるという確信の根拠としても働くようになる場合である。たとえば、物理学者ホッセンフェルダー (Hossenfelder) は、自らの分野への批判のなかで、新しい理論をテストする実験には今や数十年と数十億ドルがかかるので、理論家は観察への十全性以外の基準によって何に取り組むかを選ばなければならないと論じ、その基準はおもに社会的で美的なものであり、自己修正的であるかは疑わしいとする (Hossenfelder 2018c)。また、超対称性は正しいに違いないという同僚たちの確信を述べた物理学者たち自身の言葉を引いており、そのうちの一人はその確信を超対称性の美しさに結びつけている (Hossenfelder 2018d)。この分野についてのホッセンフェルダーの読み方には異論がある。たとえば、これらの抜粋のもとになったホッセンフェルダーの著書 *Lost in Math* を書評したバターフィールド (Butterfield) は、美の判断が歴史に条件づけられ誤りうることは、同書が論じるとおりだと認めつつ、ホッセンフェルダーは美について物理学者たちを実際以上に素朴に読んでいるとする。美に訴える者たちは、そうした限界をもつヒューリスティックとしてだけ美を用いている、というのである (Butterfield 2019, §2)。この事例は、そうした重なり合いがどのように見えうるかを、ある分野についての争われた一つの読みのうえで示すものである。それが実際に起きたかどうか、二つの読みの争点である。

一人の人が、ある取り組みはうまくいかないと判断しながら、それでも試す価値があると判断することもある。たとえば、レディ (Reddy) は後年、ある学生に、音声認識へのその学生の取り組みはうまくいかないと思うが、それでも試してみるよう伝えたと回想し、その学生ともう一人の学生について、自分の予想を超えて成功したと語っている (Reddy 2021a, 20; Reddy 2021b, 35)。ただし、同じ一連のインタビューの別の箇所では、統計を使うべきだという最初の学生の考えに、限定的にはあれ研究グループは同意していたとも語っている (Reddy 2021b, 9)。レディ自身の説明によっても、二つの判断の分離は部分的なものだったことになる。

本小節の初めに挙げた三つの問いのうち第一の問いは、予備的な評定と厳密な評定が種類において異なるのか、それとも強さにおいて異なるだけなのか、というものであった。この問いには二つの見方がある。一つめの見方は、両者を種類において異なるものとし、ラウダンやデュアー (Duerr) とフィッシャー (Fischer) のように、追求を受容から切り離す説に拠る。追求の文脈についてのラウダンの基準は、ある研究の路線の観念がどう生み出されたかではなく、その路線の実績に目を向ける。ニクルズの解説によれば、この説明では、科学者は、さしあたり全体として最も大きな問題解決の成功を取っている理論を受け入れるべきだが、いま成功の速度がより大きい伝統を追求すべきである (Nickles 2017, §1.5)。この基準は、追求を信念からは切り離すが、過去の実績からは切り離さないものと解釈できる。そう解釈すれば、この基準は、まだどの問題も解決していない路線の追求を、それだけでは認めないだろう。デュアーとフィッシャーは、ある観念がさらなる資源に値するかどうかを、その可能な認知的便益、コスト、リスク、代替案についての将来に向けた判断として扱い、その内容が真である、あるいは知られているという現在の判断とはしない (Duerr and Fischer 2025)。二人の徳-経済的 (virtue-economic) 説明は、その判断を下すためのひとつの実質的な提案である。デュアーとフィッシャーはまた、追求の基準は受容の基準を弱めたものにすぎない、という結論を退ける。単純さ、多産性、テストのしやすさのように、受容の場面では役割が争われる基準が、追求の場面ではより大きな役割を果たすからである (Duerr and Fischer 2025)。この種の見方のもとでは、ある仮説を美しい、あるいは単純だと見る予備的な評定は、それが真であるという弱い判断ではなく、それが追求に値するという判断として読むことができ、その意味で統制された比較とは種類において異なる。

ラウダンの基準では、何が解決された問題に数えられるかが、何がより大きな成功の速度に数えられるかにも関わる。そしてラウダンは、経験的な問題と並んで、概念的な問題も数えた。ニーニルオトがまとめるところでは、理論の問題解決の有効性は、それが解決する経験的な問題の数と重要さから、それが生み出す変則と概念的な問題の数と重要さを引いたものである (Niiniluoto 2024, §3.2; Nickles 2017, §1.5)。ホッセンフェルダーはまた、ヒッグス質

量の微調整（観測される値を得るのに、大きな寄与どうしが、偶然の一致に見えるほど精密に打ち消し合う必要があること）や宇宙定数の小ささのように、いま基礎物理学で研究されている不自然な数の多くは、定量的な意味でありそうにないとは言えず、したがって対応する自然さの問題は明確に定義されておらず、そもそも問題ではないのかもかもしれないと論じる (Hossenfelder 2018b, §§1, 3–5)。ホッセンフェルダーが話を聞いた物理学者のなかには、この読みに抵抗する者もいる。たとえばアルカニ＝ハメド (Arkani-Hamed) は、自分は自然さ (naturalness、理論がそうした微調整を要しないことを求める考え) の擁護者ではないと言いつつ、自然さを原理ではなく導きと呼び、そこには成功も失敗もあったとし、それがたんなる哲学ではないと述べる (Hossenfelder 2018a, 80)。この食い違いをラウダンの基準と突き合わせると、問題の地位そのものが争われているところでは、進歩の速度という基準は何が追求に値するかを決めず、争いを、どの問題が数えられるかという問いへと移すと考えられる。何を追求するかを解決された問題の記録によって判断するシステムでは、ここに、第4章第7節で見たラウダン自身の条件がかかわる。何が問題に数えられ、その解決がどれほど重要かは研究伝統の内側で定まるので、記録がどの問題をどれだけの重みで数えるかは、進歩の速度を計算する前に、誰かの価値づけによって定まっている。その価値づけが A のもののか S のものかを指定することが要る。問題をどう切り分けて数えるかは、この説にとって、それ自体が知られた難点でもある。ニーニルオトによれば、火星の軌道を定めるためにニュートン力学を適用することは一つの問題と数えられるが、初期位置が与えられれば、同じ理論は、各時刻における火星の位置についての無数の問いへの解を含意する (Niiniluoto 2024, §3.2)。ニクルズは、問題を数え、その重要性を量るのが巨介な作業だとしても、それが歴史の実情であり、その実情はむしろ、整った決定アルゴリズムより専門家の判断に分があることを示している、と歴史主義者は応じうると述べる (Nickles 2017, §1.5)。そうした速度を自ら計算するシステムには、問題を切り分ける何らかの規則が要る。その規則がシステム自身のものであるところでは、一つの問題を多くに分けることで、何も多く解かれていないのに進歩の速度が上がりうる。

答えがまだ誰にも分からない開かれた探究では、探究者は、確からしいと思える方へ進む。第一の問いについての二つめの見方は、探究の過程で実際に探究者を動かしているものを、候補を確からしく感じさせる、つまり候補への確信を動かす、というこの一つの役割のもとで眺める。そして、予備的な評定と厳密な評定を、その役割のなかで程度において異なるものとして扱う。たとえば、先行研究のなかに自分の仮説を支える主張を見つけたこと、仮説を検証する実験の結果が肯定的だったこと、仮説にもとづく思考実験で仮説が確からしく感じられたこと、仮説から導いた定理が既存の理論体系と整合したことなどは、それが実際に何を意味するか、どれほど強い根拠であるかにおいて異なりうる。それでも、どれも、その仮説を当人に確からしく感じさせるといふ点では同じだと見ることができる。さらに言えば、ある人にとっては、その仮説が美しく思えたこと、信頼する人からよきそうだとされたこと、その先に何かが開けると漠然と感じられたこと、同僚の何気ない一言、その日の気分、散歩中に見た景色なども、その仮説を確からしく感じさせうる。これは、違いが程度にすぎないという主張ではない。そう眺めることができ、そう眺めると見えてくるものがある、という見方である。本節の初めに取り上げた発見の文脈と正当化の文脈の区別を立てたライヘンバッハ自身が、「確証の前の状況は、確証の後の状況と程度においてだけ異なる」とした。確証の前でも、いくつかの事実が理論に少なくとも何らかの確率を与え、それを最善の指定 (posit) として他から区別する、というのである (Reichenbach 1938, 383)。それでもライヘンバッハは二つの文脈を区別し続け、美的な考慮は、解が正しいかどうかではなく、解がどのように見つけられるかに影響するものとして扱った (Reichenbach 1938, 384)。二つめの見方で眺めると、小さな標本でのテストと全標本でのテストは、探究者の確信を同じ方向へ、異なる量だけ動かす。それ自体は観察でない根拠も、確信を動かす。根拠が確信をどれだけ動かすかは、それが確信をどれだけの高さに残すかは別のことであり、確証理論はこの違いを、確率の言葉で、確率の関連性としての確証と確固さとしての確証の違いとして区別する (第2章第5.2節)。ここでの確信は、探究する主体 A の段階的な態度である (第2章第2節・第2.1節)。それは、推論、計画、修正のなかで果たす役割によって記述でき、誰かが確率を報告することを要しない。

ライヘンバッハが述べたのはここまでである。二つめの見方は、すべての根拠をこのように程度の問題として扱うことで、さらに四つのことを言う。第一に、根拠がはたらきかける段階的な態度はひとつ、つまり候補への確信であり、根拠の種類ごとに別々の態度があるのではない。第二に、どの根拠も、その態度を動かすものとして同じ側に並ぶ。計算、予備的な結果、ある仮説が美しいという感覚、統制されたテストは、確信をどれだけ、どちらへ動かすかにおいて異なるのであって、そもそも評定に属するかどうかにおいて異なるのではない。第三に、どの根拠も、その種類だけによって根拠になるのではない。データは、データであるというだけで根拠になるのではなく、美しさも、データでないというだけで締め出されるのではない。どの根拠も、探究者がそれを何のしるしと受け取るかを通じて効く。第四に、根拠が何のしるしになり、したがって確信をどれだけ動かすかは、探究者が立っている基準とコミットメントによる。この点は次の段落で取り上げる。

二つめの見方では、どの根拠が探究者を動かし、どちらへ動かすかは、その探究者が立っている基準とコミットメントによる。美しさが候補を確からしいと考える根拠になるのは、美しさを評価する人、その基準に立つことが有用だと考える人、あるいは自然はそれに報いると信じる人にとってである。予測と結果の一致が自然からのフィードバックとして受け取られるのは、その一致を記録する装置と背景仮定を信頼する人にとってである。他者による長い受容は、その候補を受け入れてきた共同体を信頼する人にとって根拠になり、自分の仕事のなかでの有用さは、使って成功することを便利さ以上の何かのしるしとみなす人にとって根拠になる。探究者が確からしいと思える方へ進むとき、何が確からしく思えるかは、こうして、その人が立つ基準ごとに異なる。ライヘンバッハ自身、ある科学者が、自然の調和を確信していたから大きな理論を見つけたと語ったことを報告している。そのうえで、そうした信念は内容においては帰納の原理と異なるが、知識の操作の体系のなかでの機能は同じであるとし、同時に、発見の神秘主義を、帰納という土台の上にある像と願望の上部構造と呼んだ (Reichenbach 1938, 403–404)。

ケプラーの正多面体は、こうした根拠のいくつかが出会った歴史上の一例である。ケプラーは正多面体に、コペルニクス (Copernicus) の体系の実在のアプリオリな基礎を求めた。用いていた表の値がこの仮説から計算された値と一致しないときには、データそのものが完全ではないのだから、データとの完全な一致にこだわっても意味がないというコペルニクスの言葉を、レティクス (Rheticus) が伝えるとおりに引いた。そして1619年にもなお、惑星の数と太陽からの距離を定めるのに正多面体を用いた。ただしそのころには、ティコ (Tycho) の観測を得て、軌道の離心率を正多面体から導くことはあきらめていた (Di Liscia 2025, §§3, 4.4)。経験的な証拠ではない根拠がケプラー自身の確信を動かし、のちにティコのデータがその及ぶ範囲を狭めたのである。ティコのデータがそうできたことは、それがデータであることによって保証されていたのではない。データは、ケプラーがそれを信頼したから、そして信頼した限りで、根拠として効いたのである。プトレマイオスの観測もデータであった。だがケプラーが『新天文学』で指摘したように、プトレマイオスは10分角より細かくは観測しないと公言していた。そのため、プトレマイオスの観測の不確かさは、8分角の誤差、すなわちプトレマイオスにならって

離心率を二等分した模型にティコの観測が示した誤差を上回っていた。ケプラーは、「きわめて勤勉な観測者」であるティコについて、この8分角は無視できなかったもので、それだけで天文学全体を改革する道を開いた、と書いた (Kepler [1609] 1937, ch. 19)。ジョージ・スミス (George Smith) の再構成では、8分角が食い違いとして数えられたのは、誤差の幅の見積りに照らしてのことである。ティコの観測には少なくとも2分角の不確かさがあり、ケプラーの補間と、地球の軌道についての暫定的な理論が、その幅をさらに広げていた。そして、10分角という古い基準のもとであれば、修正された模型は通っていたはずである (Smith 2014, 10, 14–15)。ケプラーは、補助的な想定も同じように範囲をつけて信頼した。『新天文学』の次の章では、太陽と地球の距離をティコの表からとりながら、その表を不正確だとし、真の値はティコの数値と100,000のあいだにあると論じた。そのため、その結論は、その範囲全体で成り立った (Kepler [1609] 1937, ch. 20)。ここでも、正多面体の場合と同じく、データはケプラーが立っていた基準を通じて数えられた。この場合は、装置、観測者、誤差についての基準である。この見方のもとでは、根拠は、強さにおいても、それを根拠にする基準においても異なるが、いま述べた動きのなかで共通の役割を果たすものとして眺めることができる。

二つめの見方は、バイズ認識論と多くを共有しており、その一部は、実質的にはバイズ的な主張である。二つめの見方が眺める側面のいくつかについては、バイズ認識論のなかにすでに議論がある。リン (Lin) の提示では、バイズ認識論は規範の枠組みであり、その中核的な規範は、確信は確率的であるべきであり、条件づけによって変わるべきである、というものである (Lin 2022, §1.2)。この枠組みは、ほとんどこの二つの規範だけで、確信の変化に焦点を当てながら、認識的な実践の広い範囲を説明し、あるいは正当化しようとする。バイズ主義者は、確証を確信の上昇として扱い、証拠による支持や説明力もそれによって説明してきた。また多くのバイズ主義者にとって、受容は二次的なものであり、ジェフリー (Jeffrey) が論じたように、なしで済ませるのがよいものでさえある (Lin 2022, §1.3)。

バイズ認識論に議論がある側面の一つは、観察でない根拠がどう確信を動かすかである。単純さ、受け入れられた理論との整合、類推のようなもつもらしさの考慮は、事前確率として入る。サーモンは、クーンの理論選択の基準の多くをそのように読んだ (Salmon 1990, 182–187, 198–199)。また、長く探しても理論の代案が見つからないという経験的でない事実が、ある条件のもとでその理論を確認しうる (Dawid, Hartmann, and Sprenger 2012, §4)。整合性のほかに事前確率を縛る規範があるかという事前確率の問題は、バイズ主義者を分けている。主観バイズ主義者は整合的な事前確率をすべて許す。そのため、確信がそもそも帰納的にふるまうかどうか、より単純な仮説を好むかどうか、事前確率による。客観バイズ主義者は、偏りのなさを求める規範を加える (Lin 2022, §1.5)。同じ規則も、入れる値が違えば異なる判定を出す (Salmon 1990, 180–181)。サーモンは、科学者の二日酔いや政治的な反感のような、客観的に無関係な考慮が事前確率を動かすことを否定しなかった。どの科学者も、イデオロギーや形而上学の先入見に気づかないまま影響されうること認めている。そのうえでサーモンは、そうした考慮を脇に置くことを探究者に求めた。それらに事前確率を決めさせれば、科学の目的に反する、というのである (Salmon 1990, 183)。二つめの見方は、探究の過程で探究者を実際に何が動かしているかの眺め方であり、サーモンが認めた影響も、そのなかに入る。そして同じ理由で、二つめの見方は、どの動きが合理的かを、それだけでは言わない。

もう一つは、根拠が効くのは何を信頼しているかを通じてであり、その信頼も変わりうることである。条件づけは証拠への確信を確実にまで上げるが、ジェフリー条件づけは、経験が確実に至らないまま確信を動かすことを認める。そして、赤っぽく見えたことが、そのゼリービーンが赤いという確信をどれだけ上げるべきかは、照明についてどう考えているかによる (Lin 2022, §5.3)。バイズの言葉で言えば、ある基準に立つ人にとって根拠が効くのは、その人の条件付きの確信、たとえば仮説が美しいとすればそれが真であるという確信や、装置の誤差のモデルが、その根拠に関連あるものにするときである。サーモンは、バイズのアルゴリズムは自明であり、重要なのは、そこに入れる確率を評価する科学的な判断とした。そして、科学の歴史を、事前確率にかかわる情報の豊かな源とみなした (Salmon 1990, 201)。プトレマイオスの10分角から、ティコの観測の誤差の幅の見積もりへのケプラーの移行のような変化は、この入力側の変化にあたる。そうした基準がどう重みを得るかも、バイズの枠組みでモデル化されてきた。階層バイズモデルでは、単純さへの選好が上位の仮説の尤度から生じ、その選好はデータへの適合を覆すのではなくそれと釣り合う (Henderson et al. 2010, 187, 198)。また、メタ帰納の論証では、類似の理論の実績が事前確率を定める (Dawid, Hartmann, and Sprenger 2012, §5)。基準が、自然さの原理がそうであったように探究のなかで重みを得たり失ったりし、探究者が新しい基準に立つようにもなることは、第6章第5節が基準の修正として検討する。

どの候補が考慮されているかの変化についても、議論がある。新しい理論を考慮に入れることは、確信の変化ではなく確信の形成であるように見え、それが条件づけに投げかける難しさは、新しい理論の問題として論じられている。シモニー (Shimony) とサーモンが素描し、ヴェンマッカーズ (Wenmackers) とロメイ (Romeijn) が展開した提案では、確信は考慮されている可能性の集合を条件とするものとされ、その集合は新しい理論が考慮に入ると広がり、考慮され続ける可能性どうしのあいだでは確信の比が保たれる (Lin 2022, §5.2)。サーモンはまた、まだ思いつかれていない仮説の全体であるキャッチオールについての尤度を扱いようがないとし、考慮されている理論どうしを比べることで、それを計算から除いた (Salmon 1990, 191–192)。第4章第4節・第5節と本章第3節が扱ったのは、こうした広がりの手前のことである。探究が出会う事柄のどれを取り上げ、どちらへ進むかであり、まだ利用可能でない候補がどのようにして利用可能になるかである。この提案に従えば、新しい理論に確信が割り当てられると、考慮され続ける候補どうしの確信の比は保たれたまま、それぞれへの確信は下がる。そのため、候補を利用可能にする仕事が進むと、それまで有力だった答えへの確信が、その答えに不利な証拠がなくても下がることもありうる。研究の前進を確信の上昇だけで数える評価は、この場合、代替を形成する仕事を後退と数えかねない。これは、第2章第8.2節で見た、答えを増やさずに知らないことを明らかにする変化に近い。

根拠をこのように眺めても、第2章第5.1節が、類推、予備的な結果、試験、追試、証明、史料の照合、解釈的な批判について述べたことは否定されない。それらは、何にかかわるか、どの代替を見分けるか、どれほど覆されうるか、どのような身分を与えうるかにおいて異なり、そのすべてを信念の更新と呼ぶと、ひとつの効果を記述する一方で、その効果が異なる理由が失われる。二つめの見方は、共通の役割を名指し、その違いを述べることは別に残す。また、この見方は、確信の度合いひとつでは抜け落ちるものがある箇所を示す。確信を同じ水準にと定める二つの証拠の集まりも、重みにおいては異なりうる。少数の試行にもとづく確信は次の試行で大きく動きうるが、多数にもとづく確信は動きにくい (Genin and Huber 2020, §3.2)。データがすでにどの役割を果たしたかもかわる。たとえば心理学の事前登録についてのある提案では、研究は探索的なものから確証的なものまでの連続体の上に置かれるが、ある仮説を思いつかせるのに役立ったデータは、もはやその仮説のテストには適さない (Wagenmakers et al. 2012, 3–5)。そうしたデータの使用がデータの与える支持を弱めるのかどうか、またなぜ弱めるのかは、争われている。予測と適応 (accommodation) をめぐる論争についてのバーンズ (Barnes) の概観では、クリストファー・ヒッチコック (Christopher Hitchcock) とエリオット・ソーバー (Elliott Sober) は、懸念を、雑音を含むデ

ータへの過剰適合の危険に置く。これに対してメイヨー (Mayo) は、データから構成された仮説もおお敵しいテスト、すなわちその仮説が偽であれば合格しそうでないテストに合格しようと論じる (Barnes 2022, §§6.4-6.5, 6.7)。根拠は、何をありそうにするかにおいても異なる。上で述べたケプラーの代理仮説は、経度を計算する手段としての位置づけを保ったまま、軌道の記述としては退けられた。そして、探究者がある根拠に与える重みは、その根拠の実績とともに変わりうる。たとえばフィッシャーは次のように論じる。自然さの原理のもとでは、ヒッグス質量の微調整は多くの物理学者に奇妙な偶然と映った。この原理は、それが生んだ有望な理論との整合の見込みによって、素粒子物理学者のあいだで信頼度を増し、大型ハドロン衝突型加速器が、それらの理論が期待させた新物理の兆しを見つけないにつれて、信頼度を失った。この相互の支えが研究を追求に値する企てに集中させたのか、誤らせたのかは、フィッシャーは未決にしている (Fischer 2023, 1052, 1055-1058)。したがって、これらの違いを失わずに根拠をひとつの尺度の上で眺めるには、それぞれの根拠について、何をよりありそうにするのか、誰の基準によるのか、その結果の確信がどれほど頑健なのかを述べることが伴う。どれほど根拠があっても、 A の確信の修正は、それだけでは候補を S の知識にしない。候補が S の知識になるかどうかは、第2章で検討し、第8章で改めて検討する関係による。

第3節は、候補を思い浮かべること、それを追求に値するものとして扱うこと、それを受け入れること、それを信じること、それを知ることが、程度においてのみ異なるのかどうかを決めずにおいた。二つの見方がかかわるのは、この問いの一部である。どちらの見方でも、候補を追求する価値は、その候補への確信の程度ではない。また、知識には固有の条件があるので (第2章)、どちらの見方でも、確信が高まるだけで候補が知られることにはならない。二つの見方が分かれるのは、厳密なテストの前の評定、たとえばある仮説を美しいと見ることが、厳密なテストがさらに動かすのと同じ段階的な態度を動かすのか、それとも何が追求に値するかの判断として読めるために、厳密なテストとは種類において異なるのか、という点である。もっとも、二つの見方は、同じ評定についてともに成り立ちうる。仮説を美しいと見ることは、それが追求に値するという判断でありながら、その仮説への確信を動かすこともありうるので、二つが対立するのは、予備的な評定を、追求に値するかの判断としてだけ読むか、確信を動かすものとしてだけ読む場合である。後の章が用いるものは、どちらの見方のもとでも成り立つ。一つめの見方では、追求は独自の基準で判断される。二つめの見方でも、たとえばニュルップのモデルのように、候補を追求する価値は、答えを知ることの価値と、事柄が決着する見込みにも依存する。

人工研究システムをつくらうとすると、そのシステムに何をもちたせなければならないかを、この実践の記述はそれだけで定めるものではない。それでも、システムを考えるうえでの示唆を与える。その示唆は、とる見方による前に、事柄が誰にとって不確かによって変わる。二つめの見方の初めに述べたように、本節が扱ってきたのは、答えの分からない探究者が確からしいと思える方へ進むという、不確か性のもとの実践である。この実践は、事柄が A にとって不確かであるところで生じる。そして、 A にとっての未知と S にとっての未知は、一致するとは限らない (第4章第7節)。 A が機械であり、事柄が S にとっては未決でも A にとっては不確かでない場合、たとえば A がすでにもっているものから答えられる場合には、 A にはこの実践そのものが要らないかもしれない。そのとき問われるのは、 A の答えに S が依拠できるかどうかであり、それは第11章が検討する。事柄が A にとっても未知である場合や、 A にとっても複雑である場合には、この記述は A にもかかわってくる。ただし、人間の研究者がこの実践のなかで抱く確信のような態度を、不確か性を実質的に減らすうえでシステムがもたなければならないのかどうかは、この記述からは決まらない。この実践がどこで起きるかは、人工の構成の境界をどこに引くかにもよる (第7章第2節)。候補の生成、予備的な検討、批判が、マルチエージェントの構成のなかのエージェントに分担されるなら、一つめの見方の区別は、その構成の内部の段階として現れうる。境界を一つのモデルの一回の出力に引くなら、そうした段階はそもそも見えないかもしれず、試行錯誤が人間の研究者と同じ単位で行われるとは限らない。事柄が A にとって不確かである場合の示唆は、とる見方によって変わる。一つめの見方からは、有望さについての将来に向けた判断と、支持についての現在の判断という、異なる基準による二つの評定を、システムがどう扱うかが問われる。評価は、レディが回想したように、ある路線がうまくいくことを疑いながらそれを試す価値があると判断でき、しかもそれを理由にその路線を信じるようにはならないかどうかを問うことができる。二つめの見方では、確信の程度についての共通の評定が、両方の判断に情報を与えうるが、それで追求に値することが確信と同一になるわけではない。システムが候補ごとにそうした評定をもつなら、確信の度合いひとつでは証拠の重みやデータがすでに果たした役割が抜け落ちるという、先に述べた限界を避ける一つの仕方は、その評定の根拠、その根拠が何をどの基準によってありそうにするのか、評定がどれほど頑健か、どのデータがすでにその候補を思いつかせるのに使われたか、そしてそのデータの上でいくつの候補が試されたかを記録しておくことである。もう一つの仕方は、経験的でない根拠がどれほど当たってきたかを追い、根拠ごとに、それがどの基準を通じて効いたか、たとえば食い違いがどの誤差のモデルのもとで有意と判断されたかを記録し、その基準を、実績がそれを求めるときには改めることである。固定された基準のなかでの更新は、この最後のことを与えない。また、その評定に、知ることの価値とコストとを合わせて、何を追求するかを判断することもかかわる。どちらの見方のもとでも、追求に値することを真である見込みと同一視し、すでに確からしいと見ている候補だけを追求するシステムは、ほかの理由で追求に値しうる候補を逃し、生産的なリスクを排除しうる。たとえば、方法を明らかにしうる候補や、ニュルップのようなモデルのもとでは、容易に反駁できる候補である。逆に、評定を無視するシステムは、恣意的な新しさに際限なく費やしうる。

そうした判断を行うよう作られるシステムについて書く論者たちは、危うさをそれぞれ別のところに見ている。第4章第7節は、探索が、 S が現に重要とみなすことまで追跡する必要があるかを問うところで、そのうちの二人に触れた。ここでは、何が追求に値するかを判断することについて二人の議論が何を含むかを、詳しく見る。北野 (Kitano) は、ノーベル賞に値する発見をなす人工システムをつくることを大きな挑戦として提案するなかで、二つのアプローチを対比する。一つは、科学者が、ある基準のもとで自らの発見が重要なものとなる確率を最大化しようとする価値駆動のアプローチである。もう一つは、そうした基準を課さずに、「どの水準の重要性であれ」発見がなされる確率をシステムが最大化する探索駆動のアプローチである。その理由として北野が挙げるのは、自ら地平の問題 (horizon problem) と呼ぶものである。重要性の地形が単調でなければ、期待される重要性の高いところに絞った探索は、途中の結果の期待される重要性が低いために、重要な発見へ至る道を避けてしまうかもしれない。ただし北野は、システムも資源に制約されており、探索できる範囲には限りがあると付け加える。また、所有者や利害関係者の価値を反映するように問題が選ばれるシステムは予期しない発見をする見込みが小さいと考える一方で、挑戦の成功の尺度としては人間による認知を保っている。システムの発見のなかに、「ノーベル賞がそれ以上に値する水準で非常に重要とみなされる」発見がなければならぬ、というのである (Kitano 2021)。コスロウィ (Khosrowi) は、発見を予測し、アイデア、問い、仮説、実験を推薦するシステムについてのプレプリントで、そうしたシステムが、追求されそうなことと追求に値することとの隔たりに直面すると論じる。追求され出版されたものの記録で訓練されたシステムは、資金、可視性、研究者間のネットワークなど、追求されることと相関する特徴を学ぶ。そのため、ある方向が追求されるだろうという予測から、それを推薦することへ移るのは、コスロウィが自然主義的誤謬の罠と呼ぶものに陥ることになる。他の

仕方では追求されないだろうということも、それだけではその方向を推薦する理由にならない。コスロウィの説では、追求に値することについての独立の基準が必要であり、こうしたシステムについての文献はその基準をまだ理論化していない。さらに、こうしたシステムがうまく働いたら、予測する方向そのものを動かしてしまうので、導入後の予測の成功だけでは、追求に値するものを先取りしていることと、予測されたものが何であれ研究者にそれを追求させていることとを区別できない、と論じる。そして、課題設定をこうしたシステムに任せられるかどうかは、科学が何をめざすと考ええるかによる、とする。コスロウィがキッチャーから引き継ぐ見方のように、科学が、探究の進行とともに見いだされていく人間の関心と価値に照らして重要な真理をめざすのであれば、人間は課題設定に深くかかわり続けることになる (Khosrowi 2026, §§4.1–4.2, 4.4)。北野の地平の問題は、システムの追求を現在の重要性の基準に固定することに反対する理由を与える。ただし北野自身の成功の尺度は、なお人間による認知であり、それは、のちにその成果を認める S による認知でありうる (第8章第3.2節)。これに対してコスロウィの議論は、過去の追求の記録から追求に値することを読み取ることに反対する理由を与える。 A が機械で S が人間の社会である場合、どちらも、何を追求するかについての A の判断と、 S が現在重要とみなしているものとの関係にかかわる (第4章第7節)。上で見たニュルップのようなモデルで、期待される認知的価値によって追求を判断するシステムでも、地平の問題は消えず、場所を移して、候補が真かどうかを知ることの価値のなかに現れる。その価値が、答えがさらに導く発見のなかにあるなら、北野によれば、それを評価するのはきわめて難しく、あるいは不可能だからである (第10章第2節)。

第4.2節が評定そのものについて論じたのは、次のことである。何が追求に値するかの判断と、何が真でありそうかの判断は、実践のなかで食い違うことがある。一人の研究者のなかで分けておかれることもあれば、共同体のなかで重なっていくこともある。予備的な評定と厳密な評定を、種類の異なる評定として眺めるのがよいのか、共通の役割のなかで程度が違うものとして眺めるのがよいのかは、ここでは決めない。研究システムについて何が示唆されるかは、とる見方によって変わる。どちらの見方のもとでも、候補を追求する価値は、それが真である見込みだけに帰着するとは限らない。第4.1節と第4.2節の議論は、どちらも評価にかかわる。最終的な正答は、ほかの出力の証拠と同じく (第2.1節)、システムが、課題が未解決のまま残した事柄を取り上げたりそのままにしたりし (第4章第5節)、それらのあいだで注意を配分し、候補を進展させ、行き止まりを認識し、初期の仮定が崩れたときに探究を組み直すことができたかどうかには答えない。研究能力がシステムに帰属されるとき、問いの選択がどこに位置するのかは、第7章第4.3節が検討する。第4.2節の問いは、後の章にも現れる。第10章では、認知的分業がひとつの行為者の内部のポートフォリオになるとき、追求に値することを真である見込みと同一視せずに、何が追求に値するかを判断するという問題が生じる。第11章では、まだ形成途中の候補にどの時点で批判を当てるかが、同じ問題にかかわる。

5. 異質な操作、比較可能な機能

第4節は、候補の形成、追求、評定が、別々の局面ではなく区別できる機能であることを論じた。研究のおなじみの操作にも、同じことが言える。本節は、異なる分野の操作を、どの水準でなら比べられるのかを問う。実験 (experiment)、観察 (observation)、証明 (proof)、モデル化 (modeling)、解釈 (interpretation)、文献レビュー (literature review)、構築 (construction)、執筆 (writing)、会話 (conversation) は、それとわかる種類の作業を名指している。それらは、あらゆるプロジェクトが通過する普遍的な段階を名指しているのではなく、それぞれがひとつの不変の機能をもつわけでもない。同じ操作が異なる探究において異なる機能を果たしうるし、ひとつの機能が分野をまたいで異なる操作を通じて実現されうる。

それらの操作を比較するには、三つのレベルを切り離しておくで見通しがよい。ただし、より細かな分析では別のレベルを加えることもできる。研究全体のレベルでは、活動は、本書の作業的特徴づけでは、ある対象について、社会 S にとって新しい認知的達成をもたらすことに向けられている。 S 自身がその達成の担い手である必要はない (第1章第2節)。中間の機能的レベルでは、証拠、形式的制約、批判、構築、あるいは世界との相互作用が、候補となる内容、表象、方法、測定関係、あるいは依存関係の位置づけを変えうる。具体的な実現のレベルでは、この作業は、実験、証明、校正、史料批判、解釈、構築、シミュレーション、追試といった様々な操作を通じて遂行される。中間レベルは比較を可能にする。それは、あらゆるプロジェクトが個別の評定のステップを含むことも、局所的な操作が同じ推論を許容することも立証しない。

関連する枠組みは、チャンによる科学的実践の分析に見られる。チャンは科学の仕事を認知的活動として捉える。それぞれの活動は、知識の産出や改善に特定の仕方で寄与することを意図された、心的あるいは物理的な操作のある程度まとまった集まりであり、活動は互いに関係づけられて、一定の目的に向けられた実践の体系をなす (Chang 2017b, 4–5)。エル・マワス (El Mawas) がこの枠組みをまとめるところによれば、チャンはまた、ある活動を部分的に定義する目的である内在的目的と、その活動が体系のなかで果たす外的機能とを区別している (El Mawas 2021, §§2.1–2.2)。ここで用いるレベルは、別の仕方で切られている。本章が操作と呼ぶものは、チャンにとって活動を組み立てる心的あるいは物理的な操作ではなく、おおよそチャンの言う活動に当たる。また中間のレベルは、チャンの言う外的機能よりも狭い。候補、表象、方法、依存関係の位置づけにかかわる役割だけを取り出し、そうした役割を一つの体系の内部でなく分野をまたいで比べられるようにするためである。また最上位のレベルは、それぞれに固有の目的をもつチャンの実践の体系の一つではなく、本書の作業的特徴づけが研究一般に与える方向である。

たとえば実験は、仮説のテストとは別の機能も果たしうるし、鋭く特定された仮説から始まる必要もない。シュタインレ (Steinle) は、デュフェ (Dufay) の電気の仕事にもとづき、さらなる例としてアンペール (Ampère) とファラデー (Faraday) の電気と磁気の仕事を挙げて、探索の実験を、実験の行為と概念の形成とがともに展開する仕事の仕方として描く。それは、経験的な規則性と、それを述べるための概念とカテゴリーとをめざし、狭い意味での理論が、あるいは実験の結果を適切に定式化できる概念さえもが利用可能でない状況に典型的に見られる (Steinle 2004, 50–51)。〔訳注：日本語の「実験」は、仮説の当否を確かめることを第一に連想させる。英語の experiment も、ラテン語の「試す、検査する」に由来し、同じ含みをもつ。シュタインレの「探索的」という限定は、規則性や概念を見出すための実験を、理論を検査する実験から取り出すためのものである。〕このことは、そのような実験を認知的に制約のないものにするわけではない。シュタインレの説明では、そうした実験は高度に体系的でありえ、装置のできるだけ多くのパラメータを体系的に変えること、どの条件が効果に不可欠でどの条件がそれを修飾するだけかを見定めることといった目標に導かれうる (Steinle 2004, 50)。それらの直接の評定対象は、単一の理論的命題ではなく、規則性、カテゴリー、あるいは概念でありうる。シュタインレは、新しい概念とカテゴリーは、何よりも、

多数の実験結果について安定した、できるだけ一般的な規則性を定式化できるようにすることで、その有効性を示さなければならないとする (Steinle 2004, 51)。

証明は実験とは異なるが、やはりひとつの箱への割り当てに抵抗する。証明は、定理を根拠づけることもあれば、その定理が誤って述べられていたことを明らかにすることも、より強い予想を利用可能にすることも、再利用可能な技法を発展させることも、以前には成功した戦略がなぜ拡張できないのかを示すこともある。計算を用いる数学者は、一つの操作について同じような広がりを見出している。たとえばセーレンセン (Sørensen) が紹介するように、ボーウェイン (Borwein) とベイリー (Bailey) は、計算機による実験の役割として、洞察を得ること、パターンを発見すること、予想を検査し、とりわけ反証すること、結果が形式的な証明に値するかを見定めること、導かれた結果を確認することを挙げる。セーレンセンはシュタインレに依拠して、その一部は探索の実験であると論じる。同時に、二つの分野では実験についての出発点の見方が違うので、新しい役割も自然科学の場合とは異なること、また高精度の数値的一致が証明の代わりとはみなされていないことにも触れている (Sørensen 2010, 346–348, 353)。歴史的探究や人文科学的探究における解釈は、ある読みを提案することも、アーカイブを再編成することも、沈黙を明らかにすることも、諸説明を識別することも、テキストや出来事へのアプローチを担う概念を変えることもある。工学的構築は、ある配置が可能であることを実証することも、モデルに欠けていた制約を露わにすることも、後の探究のための装置を創り出すことも、構築の仕方についての知識を生み出すこともある。ここでの機能の記述は、人工物、証明、解釈、あるいは実践的能力が、それ自体知識なのか、知識の媒体や証拠なのか、あるいは知識を生む手段なのかを決めるものではない。

したがって、機能的レベルで比較できるものは、標準的なテストによる真理値の決定よりも広い。研究者は、候補となる内容と表象を利用可能にし、それらに経験的、形式的、解釈的、実践的、批判的などの制約を及ぼし、何が帰結するかを評価し、そして探究者や、結果がそのために求められる者が、何を可能、適切、受け入れ可能、真、知られている、あるいは理解されているものとして扱うことを根拠づけられているかを変えようとする。この抽象化は、領域の違いを消し去るものではない。数学の証明、較正研究、史料批判、介入、追試は、異なる基準のもとで異なる推論を許容する。この抽象化は、局所的な実現を見えるようにしたまま、認識的機能における家族的類似性を特定するのである。

領域の違いがあっても、本書の作業的特徴づけでは、領域を横断する機能的な連続性を記述できる。関連する認識的な解決が *S* にとってまだ利用可能でないとき、探究は、探究者が何を思い浮かべ、信じ、受け入れ、退け、保留し、あるいは未解決として扱うかを変えることを含む。理解が目指される場合には、何をどのように把握しているかを変えることも含む。知ったり理解したりするようになる者が探究者と異なる場合には、探究はその者の位置づけ、すなわちその者が結果を受け入れ、知り、理解しているとされるかどうかにもかかわる。その位置づけがどのような形をとるか、すなわち個々の成員の態度か、集団の受容や組織されたはたらきかは、知識を誰がどのように保有するとみなすかの説 (保有の説、第2章第6節) による。証明、実験、解釈的批判、構築、較正はいずれも、そうした位置づけを制約しうる。連続性は、この制約を担う役割のうちにあるのであって、ひとつの表面的な手続きや、ひとつの推論の強さや、評価される対象のひとつの型のうちにあるのではない。

デューイは反省的思考の説明のなかで、関連する主張をしている。反省の事例は、どれほど異なっても、論理的に区別される五つの段階という共通の要素を示す。また、状況のどの特徴にも前もって証拠という札が付いているわけではなく、事実とは、結論に至るにあたって選ばれ、証拠として用いられる特徴のことである (Dewey 1910, 72, 103–104)。ただしこの共通の要素は、一つの思考の行為の全体の要素であり、ヒルデブランドのまとめでは、デューイがのちに示した探究の型は、探究が「不確定な状況」を「確定した状況」に変えたときに終わる (Hildebrand 2024, §4.3)。ここで提案する連続性は、探究のなかの一つの役割、つまり位置づけを制約するという役割だけにかかわるので、その意味でデューイの型よりも狭い。ホイニンゲン＝ヒューネは、数学、工学、社会科学、人文学までを含むきわめて広い意味での諸科学について、これに近い主張をしている (Hoyningen-Huene 2020, 87)。誤りから知識主張を守る個々の仕方は、諸科学のあいだで異なる。数学の諸科学では証明が、経験的な諸科学では、まったく異なる仕方で生み出され、評価され、扱われる経験的なデータが、歴史の諸科学では史料へと加工される遺物が、そして記述や理論が量化された分野では統計的その他の数学的な検定が用いられる (95)。ホイニンゲン＝ヒューネの見方では、諸科学に共通するのは、家族的類似によってだけ結ばれた九つの具体的な概念を覆う、体系的な抽象的な概念であり、それらの概念は分野や下位分野によってさらに異なるので、諸科学の統一は極めて弱いものである (90–91)。多様さを手続きの水準に、共通性をより抽象的な水準に置く点で、この見方は、ここで提案する連続性に近い。ただしホイニンゲン＝ヒューネが比べているのは、操作ではなく、科学と日常の知識のような知識の体系である。

連続性が手続きではなく役割のうちにあるなら、領域ごとに手続きが違うことは、機能的な比較を反駁しない。ただし、各操作が実際に何を確立できるかを見えなくするなら、その比較は空虚なものとなる。たとえばパーカー (Parker) は、ウィンズバーグ (Winsberg) とアルバラード (Alvarado) がまとめるところでは、計算機シミュレーションがもつ物質的な要素、すなわち具体的な計算装置のゆえに、シミュレーションは認識論的に通常の実験と同列に置かれると論じる。そして、実験でもシミュレーションでも重要なのは、研究される系と標的の系とのあいだに関連する類似があるかどうかだと指摘する (Winsberg and Alvarado 2026, §§1.1, 2.1)。モデル評価についての後の説明、すなわち第3章第3.4節で触れた目的適合性説でパーカーは、モデルがどうあるべきかは、それが用いられる目的に、さらに使い手、方法、使用の状況に依存するとし、同じ目的に適合するモデルどうしても、適合性をもたらす特徴は大きく異なりうるとする。新しい薬がある症状を別の薬よりよく和らげるかを知るために、ある科学者は動物モデルを用い、別の科学者は薬の分子化学をシミュレーションする計算モデルを用いるかもしれない。どちらも目的に適合しうるが、前者を適合させるのは動物の生物学的な特性であり、後者を適合させるのはまったく異なる特徴である。逆に、海洋の内部の力学をまったく表現しない気候モデルは、温室効果ガスの濃度が上がるにつれて海洋の力学がどう変わるかを知る目的には適合しないと判断できる (Parker 2020, 464–468)。これらの説明によれば、**実験** や **シミュレーション** という名前は、それだけでは研究が何を確立できるかを決めない。何を確立できるかは、問われている問いにとってどの類似が成り立つと考える理由があるかによる。

操作と機能の区別は、系列の記述の仕方も変える。文献レビューは、何が *S* にとって新しいかを特定することも、方法を提供することも、対立を露わにすることも、問いを生み出すこともある。執筆は確定した結果を伝えることもあるが、証拠を再編成し、論証がうまくいかないことを明らかにすることもある。会話は、計算を分担することも、解釈を批判することも、問題を保存することも、プロジェクトの方向を変えることもある。装置の製作は、後の探究を支えることもあれば、中心的な認識的問題が解決される場となることもある。そのため、目に見える一つの操作が、こうした機能のいくつかを果たしうる。歴史家による一人称の説明は、たとえば、索引付きのノートシステムを用いて証拠を再編成したことや、執筆によってプロジェクトの中心的な行為者

と論証が変わったことを記述している (Walkowitz 2009; Harkness 2009)。これらの回顧的な説明は、人文学的探究の一般的モデルを確立するものではない。それでもそれらは、これらの事例で、ノートをとることや執筆が、すでに得た結果を示す以上のことをしていたことを示している。

分野を横断して抽象度の高い水準で比べると、操作は、問いや候補を利用可能にし、それらを変容させ、制約を及ぼし、あるいはそれらのコミュニケーションと保持を支える。あらゆる貢献がそれらの機能のすべてを果たすわけではなく、いずれもそれだけで活動を研究にするのに十分ではない。論点は、異質な実践を横断する機能的連続性であって、ひとつの普遍的な手続きではない。証明、テスト、妥当性確認、較正、追試、批判といった操作がいかにかして証拠的な力を獲得するのかという、より狭い問いは、次章（第6章第3節）が引き受ける。

人工研究システムについて、このことは次のことを示唆する。システムが **実験** や **執筆** といった名前のついた操作をこなすかどうかではなく、その各操作が探究のなかでどの機能を果たしたかを問うことができる。たとえば、執筆が結果を提示しただけなのか、論証の欠落を露わにしたのか、文献の探索が *S* にとって何が新しいかを特定したのか、見つけたものを要約しただけなのかを問うことができる。ある操作がどの機能を果たしたかは、仕事が進んだかについての問いであり、出力の証拠だけでは答えられない（第2.1節）。パーカーの説明もここにかかわる。シミュレーションだけを行うシステムでも、関連する類似が成り立つと考える理由があれば、標的について何かを確立しうる。逆に、ある実行を **実験** と呼んでも、それだけでは何も確立しない。

6. 物質的・身体的・時間的・分散的な探究

第1.1節の見取り図は、探究のなかで変化しうる次元として、方法や装置とその材料、技能のような実践的能力、行為者、記録、社会的な配置を挙げた。本節はこれらを取り上げる。これらは、科学の実践の研究や、身体化された認知、拡張された認知、分散認知をめぐる議論が、詳しく論じてきたものである。研究は、材料と装置を用いて、訓練された身体によって、記録を通じて時間をかけて、そして複数の参加者によって共に行われる。本節の題の四つの語は、この四つの面を名指している。

探究の構成は、互いに交換可能な項目からなる均質な集まりではない。本節で要素と呼ぶのは、装置、試料、生物、コード、ノート、図、アーカイブ、訓練された身体、共同研究者、制度のように、探究にかかわる物や人や仕組みである。ある要素は、候補を表象可能にするのを助けるがゆえに重要でありうる。別の要素は介入を可能にするがゆえに、また別の要素は何が観察されうるかを制約するがゆえに、さらに別の要素は技能・基準・決定を参加者と時間を越えて保持するがゆえに重要でありうる。同じ要素が、エピソードによって異なる仕方でも重要になることもある。図は、あるエピソードでは私的な記憶の補助であり、別のエピソードでは推論のための空間であり、さらに別のエピソードでは公的な正当化でありうる。物質の対象は、それ自体何も表象することなしに、行為を因果的に制約しうる。したがって、ある探究を説明することは、関連する各要素がどのように関与しているのかを問うことを含み、近くにあるあらゆる対象をひとつの束に加えるだけでは足りない。第6.1節は、実践的転回のもとにまとめられる諸伝統と、関連する議論から取った二つの区別を取り上げる。二つの区別は、それらの伝統から引き出す教訓を保つのに役立つ。第6.2節は、探究がどのような意味で分散しているのかを問う。第6.3節は記録に、第6.4節は身体性に向かう。第6.5節は、探究者が人間でないとき、これらの役割がどう変わるかを問う。本節の題の四つの語は、これらの小節と対一には対応しない。物質的な要素はすべての小節にかかわり、分散は、その集団的な形も含めて第6.2節の主題であり、時間的な次元は主に第6.3節の記録とともに入り、身体性は第6.4節の主題である。

6.1 実践、参加者、役割

本小節は、本節の残りで要素を論じるための道具を用意する。実践の研究から引き出す一つの教訓、要素が何を寄与しているかを問うための二つの区別、そして、装置と違って参加者は役割を差し控えるという点である。第6.2節から第6.4節は、これらを分散、記録、身体性に当てはめ、第7章第4.2節は、第一の区別を人工の行為者に当てはめる。

探究のなかで何が可能になり、次に何が起こったのかを説明するためには、明示的な命題に加えて、第6節の冒頭に挙げた要素や時間が必要となることがある。実験の哲学は、これらの要素のいくつかを詳しく研究してきた。たとえば、実験者が、装置が見かけの現象（アーティファクト）を生み出しているのではなく正しく働いていると論じるための戦略、特定の種類の検出器を扱う技能が、どの論証が説得力をもつとされるかを形づくる実験の伝統、そして、モデル生物が果たす多くの目的（そのうち表象的なのは一部にすぎない）である (Franklin and Perović 2023, §§2.1.2–2.1.3; Weber 2018, §3.1)。

実践的転回 (practice turn) のもとにまとめられる異質な諸伝統は、それぞれ異なる主張をしている。ハッキングは介入を前景化し、実験はそれ自身の生命をもち、実験の仕事の一部はどんな関連する理論にも先立つと論じる一方で、実験の仕事が理論から独立に存在しうることは否定する。そして存在者についての實在論の根拠を、それについての仮説をテストすることではなく、自然の他の側面を調べるためにそれを操作する能力に置く (Hacking 1983, 150, 158–160, 262–263, 274)。アラン・フランクリン (Allan Franklin) とスロボダン・ペロヴィッチ (Slobodan Perović) が示すところでは、ピッカーリング (Pickering) は、実践を人間と物質の行為者性のダンスとして描き、それは人間の側から見れば、ピッカーリングの言う抵抗と調節 (accommodation) の弁証法の形をとるとする。そこでは、機械において意図した動きを捉えることに失敗すると、目標、機械の物質的な形、あるいはそれを取り巻く身ぶりと社会関係を改めることで応じうる (Franklin and Perović 2023, §2.2.4)。ほかの研究も、それぞれの仕方でも要素を探究の説明に入れてきた。ラインベルガーは、物質的な配置のなかで認識的な事物が現れる実験システムを研究した（第3節） (Rheinberger 2011, 310–312)。ラトゥールは、実験室がラットや化学物質のような物質を紙に変える記録化 (inscription) と、より硬い事実を生み出す、しだいに単純になる記録の連なりを分析した（下の第二の区別） (Latour 1986, 3–4, 16)。クノール＝セティナ (Knorr Cetina) は、実験高エネルギー物理学と分子生物学の認識的文化、すなわちそれぞれが知識を生み出す「機械装置 (machineries)」を比べた (Knorr Cetina 1993, 258–259)。ナーセシアンは、実験室のシミュレーション装置を用いたモデルにもとづく推論を研究し（第6.2節） (Nersessian 2009)、マシュー・J・ブラウン (Matthew J. Brown) が紹介するところでは、ギアリー (Giere) は、分散認知を展開しつつも、分散したシステム全体がそれによって知る主体でなければならないことは否定する (M. J. Brown 2015)。

これらの立場は、實在論、行為者性、認知、規範性、知識の社会的産出について意見を異にする。カロン (Callon)、ラトゥール、ロー (Law) が展開したアクターネットワーク論は、ラトゥールの説明によれば、行為者を人間の個人に限らず、科学の扱う自然的な存在や技術の物質的な存在を含む、非人間で非個人的な存在にまで広げ、両者について対称的である。行為者に人間性を与えるにも物質性を与えるにも、行為を帰属させる同じ作業が要るからである

(Latour 1996, 2, 8–9)。ラトゥールはこの意味での行為者を記号論的に、アクタントとして定義する。すなわち、行為するもの、あるいは他者から行為性を与えられるものであり、人間の個人の動機を含意しない (Latour 1996, 7)。互いに異なるこれらの伝統から、本章は一つの教訓を受け取る。道具、物質、技能、記録、共同研究者などは、それぞれ異なる仕方、探究を説明するものの一部となりうる、という教訓である。

この教訓を、ひとつの理論を暗黙のうちに採用することなく保つのに、二つの区別が役立つ。第一の区別は拡張された認知をめぐる論争から、第二の区別は実験室の研究と、何がデータに数えられるかについての議論から取ったものである。どちらも、要素が探究に何を寄与しているかを事例ごとに問うためのものであり、要素が行為者かどうかや、認知や行為者性がどこにあるかは決めない。たとえば、ある要素が因果的、可能化的、構成的な役割を果たすと言っても、それが行為者であるかどうかについては何も言わない。そのため、この区別は、人間と物質の行為者性が対称的かどうかとも決めない。

第一の区別は、装置、ノート、共同研究者のような要素が、ある操作や達成にどう関係しているかにかかわり、三つの関係を分ける。一つは因果的な寄与であり、要素は、起こることに影響する原因の一つでありうる。たとえば実験室の温度の揺らぎは、測定値を動かす。二つめは可能化役割であり、要素は、その操作が成り立つための前提でありうる。たとえば加速器がなければ、ある衝突実験は行えない。ただし、ここでの操作を、検出器で衝突を記録し解析することとしてとるなら、加速器はその操作の一部ではない。三つめは構成的役割であり、それを認める説のもとでは、要素は、課題や達成そのものの一部をなす。たとえば下で見るクラークとチャーマーズの説では、ノートが使用者の信念の一部を担いうる。拡張された認知をめぐる論争では、要素が因果的に寄与しているだけなのか、構成しているのかが争われてきた。可能化役割は、本書がこの二つのあいだに加えるものであり、単なる原因の一つより強く操作にかかわるが、操作の一部ではない、という関係を言い分けるためのものである。クラーク (Clark) とチャーマーズ (Chalmers) は、人と外部の資源が結合システムをなし、その構成要素が、認知が通常そうするのと同じ仕方でもとに行動を統御するとき、その結合した過程は、頭の中にすべてあるかどうかにかかわらず認知的な過程に数えられ、ノートはこのようにして使用者の信念の一部を保持しようと論じた (Clark and Chalmers 1998, §§2–4)。批判者たちは、そうした議論が示すのはせいぜい、認知が環境と因果的に結合していることであって、環境が認知を部分的に構成していることではないと応じる (Rowlands, Lau, and Deutsch 2020, §§5.2, 5.4; Shapiro and Spaulding 2021, §5)。構成的役割を認めるかどうかにかかわらず、因果的な寄与と可能化役割だけで言えることがある。装置は、何も信じることなく実験を因果的に制約する。ここでとる見方では、ノートは、集団が信頼できる仕方であらえておくことを広げうるが、その記載が誰かの信念になる必要はない。研究における記録は、是認されるためよりも、見直されるために保たれることがあり、記録された内容を誰も是認していなくてよいからである。第6.3節で述べる公開の化学の実験ノートがその例である。クラークとチャーマーズ自身の基準でも、取り出したときに自動的に是認されることが、外部の資源が信念を担うための条件の一つであり (Clark and Chalmers 1998, §5)、見直すために保たれる記録はこの条件を満たさない。共同研究は、ひとつの集団心をなすことなく、課題を分散させる。ノートと共同研究のこうした働きは、可能化役割によって記述すれば足りる。それらが何らかの操作を部分的に構成しているかどうかは、別の問いである。なお、同じ要素がどの役割を果たすかは、どの操作や達成を説明するかに相対的である。衝突を記録し解析する操作にとっては前提である加速器も、衝突実験という遂行の全体を説明するときには、その運転が遂行の一部に数えられる。そのため、要素の名前だけでは、それが可能化するのか構成するのかは決まらない。第7章第4.2節は、外部への依存が、人工の行為者に帰属される組織化の一部を供給し続けているかどうかを問うときに、可能化役割と構成的役割との違いを用いる。

第二の区別は、要素が何かを運ぶだけなのか、それとも何が運ばれうるかをも形づくるのかにかかわる。要素は、内容、情報、技能を、参加者、場所、時間を越えて保持し移動させる。第6.2節で述べるハッチンスは、航法チームの計算を、媒体を横切る表象状態の伝播として扱う。しかし、受動的な経路として扱われる装置や記録が、何が証拠として現れうるかを決めていることもある (記録の例は第6.3節で見る)。たとえばラトゥールは、実験室の装置は、その大きさや分野がどうであれ、最後にはいくつかの記号を読みとれる小さな窓に行き着き、そうした記録は書かれつつある論文に図として組み込めたと述べ、肉眼で見えない現象は記録化装置を通してしか見られないとする (Latour 1986, 4, 17)。レオネッリ (Leonelli) の説も、要素が、何が証拠として運ばれうるかを形づくることにかかわる。レオネッリの関係的な説では、「データ」とは、探究の特定の時点で知識主張の証拠を与えるものとみなされる研究の産物に適用される区分である (Leonelli 2015, abstract)。データであることは、研究の産物に内在する性質ではないのである。これは、第4.2節で、データはデータであるというだけで根拠になるのではない、と述べたことと合う。

どちらの区別についても、ある要素がどの役割を果たすかは、事例ごとに根拠を示して主張することであり、探究の近くにあるというだけで決まるものではない。また、要素の寄与がこの二つの区別で尽くされるわけでもない。要素は、候補を表象可能にするのを助けることも、証拠として候補にかかわることも、基準、許可、責任を設定することもある。そのうち最初の二つの機能は、第3節と第6章が扱う。第11章は、認定における基準と誤りの重みづけを扱い (同章第4節・第8節)、第12章は、許可と責任を扱う。そこには、同章第6.4節が述べる、そうした要素が実効的な権限をどう分配するかも含まれる。

役割は差し控えられることもある。探究が必要とする操作を担う参加者は、自分なりの理由でそれを行うのを断りうる。たとえば、クオレク (Kwolek) は後年、デュボンの研究室で紡糸装置を担当していた男性が、紡糸口金を詰まらせるという理由でクオレクの濁ったポリマー溶液を紡ぐことを当初は断り、繰り返し頼まれてはじめて応じたと回想している (Kwolek 1998, 6–7)。装置と違って、参加者は、自分なりの理由で役割を引き受けたり断ったりする。そのため、探究の構成を記述するときには、誰が何を引き受けけるかが参加者の判断に依存しうることも含まれる。

6.2 分散と実験室の研究

研究は、ふつう、複数の参加者によって、装置や記録とともに行われる。第1.1節の見取り図で言えば、行為者と社会的な配置である。すると、課題だけでなく探究そのものが分散しているのか、という問いが生じる。この問いは、ハッチンスが艦船の航法の研究で展開し、ギアリーやナーセシアンが科学に当てはめた分散認知の議論が取り上げてきたものである (M. J. Brown 2015; Nersessian 2009)。本書でこの問いが重要なのは、人工の *A* がそれをいくつかの形であらためて提起するからである。*A* が一つの個体なのか、多くのエージェントからなる構成なのかは、境界をどこに引くかによる (第7章第2節)。同じシステムを、内部が多くのエージェントからなる一つの行為者として記述することもできる。また *A* は、実験の装置と密接に結びついて一つのシステムをなすこともある。また、人間の研究では共同体に分散している機能、たとえば探究の路線への労力の配分 (第10章) や結果の批判 (第11章第9節) を、自らの内部に取り込むこともある。何がかわるかには、第6.5節で立ち返る。本小節は、探究がどのような意味で分散しているのかを問う。答えは二つに分かれる。課題や作業が、参加者、装置、記録、時間のあいだに分けられているという意味では、多くの探究は分散しており、実験室の研究はそれ

を詳しく記述してきた。他方、認知や知識そのものが分散しているかどうかは、そうした記述だけでは決まらず、さらなる理論を要する。以下では、まずこの区別を示し、次に、実験室の研究がそれぞれの側について何を示すかを見る。

分散的 (distributed) という語は、より強い理論が特定されない限り、内容の薄いものにとどまる。課題、情報、技能、因果的寄与、決定権限は、結果として生じる配置がひとつの認知的または認識的主体を構成することなしに、参加者のあいだで分割されうる。認知、行為者性、信念、あるいは知識それ自体が分散しているかどうかには、さらなる説明が必要である。課題の分散だけではそれは決まらない。第1節で触れたハッチンスの艦船の航法の研究は、認知について、そうしたより強い説明のひとつを与える。この研究は、海図や計器を備えた航法チームを分析の単位とし、計算を、参加者の頭の内外にある媒体を横切る表象状態の伝播として扱い、チームの認知的性質はその成員の誰のものとも異なると論じる。そのうえで、チームの仕事のある再編成を、参加者の誰よりも先に組織が見出したものとして記述する。船が動力を失ってジャイロコンパスが止まった後、作図係と方位記録係はそれぞれ目の前の作業に局所的に適応し、位置決定を重ねるうちにひとつの分担に落ち着いていた。その分担では、共通の部分相を一度だけ計算して各方位に足すという手間の省ける順序で、チームとして方位補正の項が加えられていたが、この解決を十分に表象していた参加者は一人もいなかった (Hutchins 1995, 117–118, 226, 317–318, 345, 351)。そうしたシステムが信じたり知ったりもするかどうかに、いま述べたさらなる説明が必要である。

分散についてのもう一つの主な証拠は、実験室の研究から来る。実験室の研究は、事実が、成員、装置、記録のあいだに分けられた作業を通じて生み出される様子を描いてきた。上に挙げた実践的転回の研究のうち、ラトゥールとウールガーの *Laboratory Life* (『ラボラトリー・ライフ』) は、ある神経内分泌学の実験室における、記録化装置、言明の様相、信用、そして主張が安定するに至る作業について、綿密な記述を与えている。その記述が拠る材料は、性質の異なる二つからなり、このことは、記述を証拠として使うときにかかわる。一つは、実験室での約2年間の参与観察であり、そこでは議論をテープに録音できず、観察者のメモから報告された。もう一つは、ある放出ホルモンの構造がどのようにして受け入れられるに至ったかの再構成であり、公刊された文献、15のインタビュー、その構造に取り組んだ二つのグループの文書記録から引かれている。著者たちは、この再構成が正確な年代記も「実際に何が起きたか」も求めないこと、そして二つのグループが争った年代については一方のグループの説明により近く従うことを述べている (Latour and Woolgar 1986, 39, 105–107, 149 n. 5, 153–154)。

これらの記述は、事実の産出が成員や装置のあいだに分けられた作業であることを示し、それを個人の観察から信念への直接的な移行として物語ることを難しくする。本小節の冒頭の言葉で言えば、それが示すのは課題と作業の分散である。しかしそれらは、それだけでは、著者たちのより強いテーゼ、すなわち、事実とアーティファクト (人為的な産物) はそれぞれ真の言明と偽の言明に対応するのではないこと、そして実在は論争の決着の原因ではなく結果であることを立証しない (Latour and Woolgar 1986, 176, 180–182)。またそれらは、あらゆる実験室が同じ組織をもつことも示さない。著者たち自身、この実験室から一般化することに注意を促しており、短い追跡調査の後には、その集団、空間、装置、問題の組み合わせを非常にまれで、繰り返されないかもしれないものと判断している (Latour and Woolgar 1986, 56–58, 233 n. 21)。

こうした記述をどう解釈すべきかは、それ自体が争われている。ラトゥールとウールガーは、ある研究者がアイデアを得たという説明を、物質的・集合的な状況の凝縮された要約として扱い、第2版の後記では、科学について認知による説明を用いることを10年間控える (moratorium) よう提案する (Latour and Woolgar 1986, 170–171, 174, 280)。これに対して、ハッチンスの説明と次に述べるプログラムは、分散した仕事を認知の語で説明する。同じ後記で著者たちは、自分たちが示す種類の証拠は、根深い存在論的コミットメントに基づく立場を変えそうにないことを認め、また、ある論者が自分たちの記述から、科学を推測と反駁の過程とみるポパーにならった正反対の読みを引き出したことを、自分たちの解釈を押し通すには自分たちの資源が足りないことを示すものと受け止めている (Latour and Woolgar 1986, 280–281)。記述された人々も応じた。実験室の成員たちは、フィールドワークの最中に、文学的な活動の参加者として描かれることに異議を唱え、刊行後の主な反応は、著者たちの報告によれば、この本は「かなり意外性のない」ものだというものだった (Latour and Woolgar 1986, 53, 274)。したがって、観察、再構成、存在論は、それぞれ別に評価される事柄である。

生物医学工学の実験室でナーセシアンが行った五年間の認知エスノグラフィー的プログラムは、別の意味での分散にかかわる証拠を与える。成員のあいだだけでなく、時間を通じて、共に発展する物質的なシステムにまたがる分散である。焦点となったプロジェクトでは、問題の特定、シミュレーション装置、器具、方法、参加者、推論が、何年にもわたって共に発展した。研究者たちは、実験的制御と生理学的リアリズムとのあいだのトレードオフを交渉した。たとえば、人間の生理の特徴をよりよく示すために、制御と精度の一部を手放した (Nersessian 2009, 752–753)。この論文はこれらの知見をモデルに基づく推論と分散認知を通じて解釈しているが、その解釈は論文のものであって、生の観察ではない。限定された教訓は、これらの事例において、発展していく物質的システムが、認識的問題がいかにして特定され追求されえたのかについての説明の一部であった、ということである。

このように実験室の研究は、それらが調べた実験室やチームでの探究が薄い意味で分散していること、つまり作業が成員、装置、記録のあいだに分けられ、物質的なシステムとともに時間をかけて発展することを、豊かに示す。強い意味での分散については決着をつけない。同じ種類の記述が、ハッチンスやナーセシアンには認知の語で読まれ、ラトゥールとウールガーには認知による説明を退ける形で読まれるからである。そのため、分散は第6.1節の区別を使って事例ごと、要素ごとに記述でき、認知や知識がどこにあるかは別の問いとして残る (第2章第6節、第7章、第8章)。

6.3 記録

探究はふつう一回の作業では終わらず、日をまたぎ、参加者を替えて続く。第1.1節の見取り図が記録を探究の次元の一つに数えたのはそのためであり、本節の題の「時間的」は、おもにここに入る。本小節は、記録が、すでに得た結果をしまっておくことのほかに、探究に何を寄与するのかを問い、三つの働きを見る。一つは、未決の事柄や失敗を、探究を続けるために持ち越すことである。二つめは、記録の形が、何を一つの試行と数え、何を何と比べるかを形づくることである (第6.1節の第二の区別)。三つめは、何を残すかが、誰がどの用途で使うと見込むかに左右され、記録がときに誰も計画しなかった用途に役立つことである。

一つめの働きから見る。作成者自身にとっての手がかりとして最適化されたメモは、後の共同研究者には使えないかもしれない。分野での再利用のために作られる記録には、異なる労力が必要となる。公開された実験ノートや課題 (issue) の履歴は、完成した論文が圧縮してしまう失敗した実験、再解釈、訂正、物流上の制約、決定を保存しうる。たとえば、ある公開の化学の実験ノートでは、学部生が失敗した蒸留とその結論を記録し、2日後に指導者がその

結果を解釈し直して結論を変え、問いを書き加えた。学生はそれらの問いに約1年後に答えた。指導者は後に、そのような実験は本当に完了することはないと書いている (Bradley 2008)。しかし、より多くのログをとることが、自動的により多くの知識になるわけではない。記録が後の継続、監査、再現、あるいは批判を支えるためには、それらの用途に照らして解釈でき、関連する来歴と結びつけられ、訂正や撤回が重要な箇所伝わるように管理されていることを要する。

二つめの働きとして、記録は、何を一つの試行と数えるか、何を何と比べるかを形づくることもあり、その点で結果を運ぶ以上のことをする（第6.1節の第二の区別）。エジソンの研究所の蓄電池試験のノートで、デジタル版の編者が1900年ごろのものとするものが、その例である。このノートが属するノート群についての編者の注によれば、これらのノートの試験には1から1777までの番号が振られており、2冊目までにエジソンは、充電と放電のあいだの各セルの時刻、電流、電圧、容量を記録する標準の書式を定めていた (Thomas A. Edison Papers 1900, editorial note)。ここで見た連続する7枚の画像では、試験146から152のそれぞれが、電極の格子の空のときと詰めたときの重さと活物質の組成で始まり、測定値の列がそれに続く。試験150は「Same as 149」(149と同じ)と書かれ、その後にニッケル酸化物の異なる配合比と、格子の内側にめっきしたという注記が続く。試験152は「Same as 151 Except pkd water moist」(151と同じ、ただし水で湿らせて詰めた)と書かれている。試験149の注記は「As no 1/16 grid has worked」(1/16の格子はどれもうまくいかなかったので)で始まり、この試験が、欠陥が格子にあるのかニッケル酸化物にあるのかを確かめるためのものと述べているように読める (Thomas A. Edison Papers 1900, canvases 36–39)。この形の記入は、先の試験を基準として名指し、違いだけを記録する。そのため、番号の並びそのものが比較を配置する。番号は、何を一つの試験と数えるかも定める。152という一つの番号のもとで、測定値は見開きの両頁にわたり、日を変えた複数の運転に及び、途中で新しい格子と新しい鉄の注記が現れても新しい番号は振られていない。これに対し、152を151から分ける活物質の詰め方の変更には、新しい番号が振られている (Thomas A. Edison Papers 1900, canvas 39)。

このような記録のなかの実験の数は、記録がそれをどう区切るかに依存する。自動化されたシステムのログのなかの実験の数が、そのログの単位に依存するのと同じである。また、ある比較が後の読み手に利用可能なのは、記録がこのような形で保たれたからにすぎないこともある。これらの試験は、研究としてではなく製品の定型的な試験として読むこともでき、ここでは、それらが本書の作業的特徴づけ（第1章）の意味で研究であったかどうかを決めない。この例が示すのは、記録の形が何を数えられるもの、比べられるものにするかだけである。これらの頁はこの点を示すために選んだもので、このノートやその系列に典型的なものとして示すのではない。

三つめに、何を記録しておく価値があるかは、その記録がどの将来の行為者と用途を支えることが期待されているかに部分的に依存する。プロジェクトがすでに何を試したかの記録は、そのプロジェクトの後の成員に役立つものであり、その記録を保つことは他者の注意に依存しうる。たとえば、第1節で触れた公開のログをもつ言語モデルのプロジェクトでは、ある否定的結果は、別の参加者がそのアイデアはすでに試されたのかと尋ねたときにはじめて、プロジェクトの課題 (issue) に書き留められた。また別の事例では、ある実験はすでに行われたという参加者の示唆が、その実験を割り当てられた人によって訂正された。その人は、実験の入力データがそもそも完全に抽出されていなかったと答えた (Marin Community 2025; Marin Community 2024)。

記録は、誰も計画しなかった用途に役立つこともある。たとえば、OPERAとLVDの二つの検出器の同時事象は、探していた事象を一つも見つけなかった宇宙線の解析のなかで探されていた。OPERAの共同研究が計時系の故障を見つけた後、同じミュオン線の記録は、その故障がいつ始まったかを決めるのに使われた (LVD and OPERA Collaborations 2012)。あるいは、記録が、まだ定式化されていない問いのために保たれることもある。大腸菌の長期進化実験は500世代ごとに試料を凍結して「化石記録」としており、それが後に、予見されていなかった形質の起源を、以前の世代から進化をやり直させることで調べることを可能にした (Blount, Borland, and Lenski 2008)。

このように記録は、探究を時間と参加者を越えてつなぎ、何が比べられるかを形づくり、探究を後の用途に開いておく。これらの働きは、本書のほかの箇所で取り上げる問いにもかかわる。結果が *S* のなかに保持されるかどうか（第2章第6節）、訂正や撤回がどこまで伝わるか（第11章第8節）、そして探究者が機械であるとき記録の使い手が誰になるか（第6.5節）である。

6.4 身体性

研究は、訓練された身体によって行われる。第1.1節の見取り図が、技能のような実践的能力を探究の次元に数えたのはそのためである。本小節は、身体が探究に何を寄与するのかを問い、三つの水準を見る。一つは、熟練した知覚と行為が、差異を際立たせ、装置を安定させ、操作を導くことである。二つめは、疲労、焦り、自信、美的な惹かれのような状態が、注意、粘り、打ち切りを左右し、それ自体が結果の依存先として扱われうることである。三つめは、身体の実験を含む生きられた経験が、どの未決の事柄がそもそも取り上げられるかにかかわることである。ここでの証拠の多くは想起による自己報告なので、その限界もあわせて述べる。

一つめと二つめの水準については、科学者への回顧的インタビューが、可能なメカニズムを提供する。ただし、頻度の推定は提供しない。また、あらゆる研究に成り立つ二つの様式の心理学を、それだけで立証するわけでもない。そうした二つの様式の区別の一つとして、ヤナイ (Yanai) とレルヒャー (Lercher) は、フランソワ・ジャコブ (François Jacob) に従って、仮説を構造立ててテストする昼の科学と、まだ十分に形をなしていない可能な仮説を探る夜の科学とを区別し、この区別は創造性を要するあらゆる営みに一般的なものかもしれないと述べる (Yanai and Lercher 2019)。下で見るポッドキャストの名は、この区別に由来する。たとえば *Night Science* (ナイト・サイエンス) ポッドキャストの回の大半は、科学者への回顧的インタビューである (Yanai and Lercher n.d.)。別の科学ポッドキャストのホストたちとの対話のなかで、*Night Science* のホストたちは、自分たちが創造的だと判断したゲストを選び、収録中に会話の舵を取り、行き止まりを編集で取り除いていると述べている (Yanai and Lercher 2023)。したがって公開された回の内容は、ゲストの想起だけでなく、ホストの選択も反映している。感情や、アイデアがどれほど突然に訪れたかについてのゲストの報告は、見たことやしたこと想起ではなく、想起された内省に拠っており、過程についてのホストの枠づけを採り入れるゲストもいれば、それを退けるゲストもいる。

ほかの回顧的な説明も、同様のメカニズムを記述している。たとえば、第4章第5節で見たベル・バーネルとキーリングの回想である。ベル・バーネルは、望遠鏡の記録紙を目で解析すると決め、読み進めるうちに、シンチレーションを起こす源と干渉を見分けられるようになった (Bell Burnell 1977)。キーリングは、慌ただしさのなかで較正ガスの偏りの可能性を調べるのを先送りしたが、何年も後に訪れた同僚が三台の分析計のペンの描いた線を一瞥して、

一台がほかと食い違っているのに気づいたことも回想している (Keeling 1998, 46, 50)。ベル・バーネルの回想は、熟練した知覚が差異を際立たせるという一つめの水準にかかわる。キーリングの回想は、慌だしさが点検を先送りさせるという二つめの水準と、一瞥で食い違いを捉える熟練した目という一つめの水準の両方にかかわる。

同時期の記録は、そうしたメカニズムとそれへの応答を、想起されたままにではなく、書き留められたままに示しうる。たとえば、ライナス・ポーリングの研究ノートでは、1934年5月2日と記された塩素の電子回折の測定のが、内側の極大と極小の読み取りには「大きな心理的誤差」(a large psychological error) があるという注記で終わっている。5月2日に書き手は「環を小さく読もうと努めて」いたことに気づき、5月3日には前日の高いCI-Clの値を知っていたので、環を「大きく (あるいは、正しく、と私は思った) 読もうと努めた」というのである。注記は、このことが第二極大から外側の直径には影響しないと付け加え、それだけを平均している。5月3日の日付をもつ次の頁は、同じフィルムが「マスクを用いて (以前は余分な光に悩まされていたので) 測り直された」ことを記し、その結果の下に、バジャー (Badger) がバンド・スペクトルのデータから計算し直した値を書き留めている (Pauling 1932–1941, 082–083)。この頁はこの例のために選んだものである。この頁は、熟練した読み取りが、作業のさなかに、結果が依存するものの一つとして扱われ、それに応じてどの読み取りに依拠するかが決められたことを示していると読める。そうした効果がどれほど頻繁に気づかれ、記録されたかについては、何も示さない。

三つめに、身体の実験やそのほかの生きられた経験は、どの未解決の事柄がそもそも取り上げられるかにかかわることもある。ただし、そのことによって、それらの事柄について何が根拠づけられるかが決まるわけではない (第4節)。第4章第5節で見たボー・シアの回想は、ある *Night Science* の回で語られた。その回は、ボー・シアらの研究をまとめた論文の公刊当日に収録され、論文の共著者である元指導教員がホストの一人を務めた。ボー・シアによれば、自動車事故で尾骨を負傷してから、およそ一年のあいだ、座るたびに痛みが戻ってきて、そのたびに、人間には尾がないということに注意を引き戻された。その負傷への好奇心から、尾がどのように発生するのか、ある動物がどのように尾を失うのかについて読み始め、その読書が、やがて共著者たちと公刊した尾の消失の研究へとつながっていった (Yanai and Lercher 2024; Xia et al. 2024)。ここで依拠する他の報告と同じく、これはホストが形づくった会話のなかで語られた、想起による自己報告である。それは、そうした経験がどれほど頻繁にかかわるかも、負傷が問いを作り出したことも示さない。示すのは、探究者自身の身体が、その事柄についての研究上の問いが定式化される前に、それに注意を引き戻しうる一つの仕方である。A が人工のシステムで S が人間の社会である場合には、そうした注意の共有された源さえ前提できないことを、第4章第7節が述べた。この事例がかかわるのは、個人的で偶発的な経験である。それは、立脚点理論が分析する、どの問いが生じるかの社会的に構造化された位置への依存の例ではない。後者は第10章第6節が扱う。

人間の証拠的実践が、自然との長い相互作用を通じて形成された原初的な身体的予期に依存しているのかどうか、そして人工のシステムでは何がその役割を果たしうるのかという、より深い問いは、第8章第2節で取り上げる。第2.1節はそうした予期を検査の前提のなかで扱い、第2.2節は相互作用の履歴が何を与えるか、何を与えないかを問い、第2.3節は、世界への応答可能性、すなわち探究者のコミットメントが世界の側からの訂正を受けうるものであり続けることに至る、人工の経路を検討する。ここでのより狭い論点は、具体的な探究の説明には、身体的能力が含まれうるということである。探究者が機械であるとき、これらの水準のそれぞれで何がその役割を果たすのかは、第6.5節が問う。

6.5 探究者が人間でないとき

第6.1節から第6.4節の事例は、どれも人間の研究から取ったものである。そこでは、探究者のいくつかの特徴が動かない背景になっていた。作業記憶が小さいこと、技能の習得に長い年月がかかること、一人が一つの身体をもつこと、疲労や感情が注意を左右すること、一人の研究者を複製できないこと、何ですでに試されたかを互いに尋ね合って知ることなどである。探究するのが機械であれば、その一部が動く。そのとき何が変わり、何が変わらないかを見る。研究が依拠していたものは、人間という実現者についての前提 (作業記憶の小ささなど)、現在の科学のシステムの組織についての前提 (歴史のなかで形づくられ、実現者が人間のままでも組み直されうる組織のあり方)、探究の条件 (何が探究を実現するにせよ残るもの)、人工の構成の外にいる当事者との関係 (S がその外にあるときは S を含む) の四つに仕分けられる。四つは、機械が探究するときはどうなるかで分けられる。一つめは実現者が変わると崩れ、二つめは、実現者が変わるかどうかにかかわらず、組織が変わると崩れる。三つめはそのどちらでも崩れず、四つめが成り立つかどうかは外の当事者による。一つの要素が複数の種類にかかわることもあり、四つは、新しい実現がもたらすもののすべてを覆うわけではない。第9章第6節はこの仕分けを詳しく示して研究を取り巻く関係に用いるが、本節では、それを先取りして、ここで見た要素の役割に当てはめる。手がかりになるのはチャーニアク (Cherniak) の区別である。この区別は、有限な主体であればどれも免れない限界を、人間の記憶に固有の特徴から分ける。前者は、銀河全体の資源と宇宙の熱的死までの時間をもつ存在にとっても避けられない (Cherniak 1981, 165)。そしてチャーニアクは結論を、人間であることではなく、たとえば「同じように二層からなる記憶をもつ」存在であることを条件として述べた (Cherniak 1983, 174)。また、A が機械であっても、人間の定数がなくなるわけではない。ハンフリーズ (Humphreys) は、計算の方法について、人間と計算のあいだの界面の問題の根には、科学は人間の科学であるという前提があるとし、完全に自動化された科学には興味深い問題があるものの、当面は人間が重要な役割を果たし続けるので、入力と出力の界面の認識論は人間の限界と能力を考慮しなければならないとした (Humphreys 2009b, §2)。S が人間であれば、結果を読み、確かめ、それに依拠するのは人間であり、人間の定数はその側に残る。ハンフリーズは、解を計算するコア・シミュレーションと、その解の提示をも含むフル・シミュレーションとを区別する (Humphreys 2009a, §3.3)。有賀暢迪による紹介では、ハンフリーズの説でこの区別が重要なのは、同じ計算結果でも、その提示の仕方によって人間の理解が大きく変わりうるからである (Ariga 2009, 85)。またハンフリーズは、より短い論考で、人間の受け手に役立つ図像が、それを生む計算機にも役立つとは限らないと述べている (Humphreys 2006, paragraph beginning "These methods raise important questions")。これらの論点を S に当てはめると、方法にどのような表現の要素が必要かは、出力を誰が受け取るかによって変わりうる。

この仕分けを当てはめると、一つの要素の役割が分かれることがある。本節の冒頭で、図は、あるエピソードでは私的な記憶の補助であり、別のエピソードでは推論のための空間であり、さらに別のエピソードでは公的な正当化でありうると述べた。記憶の補助という役割は小さな作業記憶に依存するが、推論の空間という役割は必ずしもそれに依存せず、公的な正当化という役割は S との関係に依存する。

同じ仕分けは、科学の表象が何を圧縮しているかにも当てはめられる。高橋恒一（Takahashi）は、AI駆動科学を論じるなかで、科学とは、膨大な要素が同時に動く並列分散的な過程を、少数の変数と一本の論理の筋に翻訳する営みであったとする。その翻訳は、七つ程度の項目しか保持できず、互いに影響し合う変数を同時に関係づけられるのは四つ程度までで、ものごとを一本の時系列の物語として捉えがちな脳のためのものだった、というのである（Takahashi 2026, 120–121）。高橋はここで二つのことを分ける。ニュートン力学の位置と速度や熱力学の温度のように、理論が個々にどの変数をとるかは、その説では法則の構造が決めている。これに対して、そもそも自然を少数の変数に圧縮しようとする要求は、人間の認知の構造を反映している。ただし高橋は、すべての素粒子の位置と運動量を列挙すれば自然の完全な記述にはなるが、それは理解ではないともする。しかもそれは、人間にとってだけでなく、どれほど強力であっても有限であるAIにとっても同じだという。そのため、理解は、理解する者の言語が許すものへの圧縮であり続ける（130–131）。チャーニアクの区別に照らせば、少数の変数への要求は人間の記憶に固有の特徴であり、何らかの圧縮を要することは、有限な探究者であればどれも免れない限界である。ただし S が人間であれば、成員が結果を把握しなければならないところでは、少数の変数への要求は S の側に戻ってくる。出力の表現についてのハンフリーズの指摘が示唆するとおりである。

人間の前提に見えるものの下に、探究の条件が見つかることもある。疲労や感情は、どこまで粘り、どこで打ち切るかを左右する人間の仕方である。だが、粘りと打ち切りの間いそのものは、有限な探究者ならどれにも残る。チャーニアクによれば、どの探究に取りかかるかの判断もそれ自体一つの探究であり、それを常に探究によって決めようとすれば後退が生じる。人間ではこの後退が、自然選択によって、ある探究が自動的かつ本能的に始められることで止められているとする（Cherniak 1983, 177）。第6.4節の、身体の痛みが注意を引き戻した事例は、その止め方の一つとして読める。機械の探究者では、何がその役割を果たすのかが問われる。

探究の条件のなかには、探究者の側ではなく、探究される対象の側にあるものもある。高橋は、AIが科学をどこで加速しうるかを地図にするにあたって、ある分野の可能性の空間を探索するのに要する計算量と、対象に内在する時間とを区別する。前者は、この説では、投入するエネルギーを増やせば押し戻せる。後者は、対象のシステムがみずから進む固有の時間であり、順に積み重なり並列化では短縮できない細胞の分裂、分化、器官形成、数十年から数百年をかける生態系の回復、何世代もかかる制度の変化などがその例である。高橋の説では、外から加えるどんな技術もこの時間を変えられず、実験室の自動化が工学的な待ち時間をゼロに近づけても、この時間は残る（Takahashi 2026, Appendix 1）。高橋の地図は、工学的な待ち時間をゼロとおくことなどを前提とし、それらの前提に条件づけられた上限の評価として示される（Takahashi 2026, Appendix 3）。高橋は、対象に内在する時間を、恒久的に残る四種類の限界の一つに数える。残りの三つは、失われた歴史情報や他者の主観経験のように原理的に存在しない情報、高橋がアロー（Arrow）に帰する、複数の価値の次元を単一のスコアに集約する手続きの原理的な限界、そして停止問題の決定不能性などの論理的・数学的な限界である（Takahashi 2026, 149–150）。このうち集約の限界は、集約の手続きにどの条件を求めるかに相対的なものとして読める。本節が以下で用いるのは、対象に内在する時間だけである。対象に内在する時間は、機械の探究者がどの候補を速く決着させられるかにもかかわる。高橋の説では、マクスウェル（Maxwell）の変位電流のように既知の法則を数学的な構造で統合する仮説は、主に計算量に律せられる。これに対して、ルヴェリエ（Le Verrier）の未知の惑星やパウリ（Pauli）のニュートリノのように、直接には観測できない実体を仮定する仮説は、検証にあたって対象が応答するのを待たなければならない（Takahashi 2026, 156）。チャーニアクの形で述べれば、これらは有限な探究者の条件ではなく、探究される対象の条件であり、探究から人間による遅れを取り除いても残る。

AIの経済学を論じる著者たちも、アイデアの生産を何が制約するかについて、これに近い線を引いている。フィリップ・アギオン（Philippe Aghion）、ベンジャミン・ジョーンズ（Benjamin Jones）、チャールズ・ジョーンズ（Charles Jones）は、人工知能と経済成長についてのワーキングペーパーで、知識が進むにつれて新しいアイデアが見つけにくくなりうる理由を二つ区別する。一つは知識の重荷（burden of knowledge）である。人間の認知の限界にもとづくこの仕組みは、三人の説では、知識の蓄積のうち人間よりはるかに大きな部分を把握できる人工システムを制約しないかもしれない。もう一つは取り尽くし（fishing out）である。著者たちが新しいアイデアの探索の根本的な特徴とみなすこの枯渇は、人工の探索者にもおそらく同じように当てはまる。論文の言い方では、そうしたシステムは「池に何がいるかを変えはしない」。さらに三人は、そのとき残る制約を決めるのは、認知の限界よりもむしろ自然法則であるかもしれないと述べる（Aghion, Jones, and Jones 2017, 27–29）。チャーニアクの形で読めば、前者は人間の認知に固有の特徴であり、後者は、アギオンらのモデルでは、見つけるべきものとして何が残っているかの条件である。知識の重荷もまた、人工の探索者を制約しなくなっても、人間の S の成員が受け取れるものは限り続けうる。ダリオ・アモデイ（Dario Amodè）は、第1章第2節で引いたエッセイで、知能を増やすことで得られるものを制約する要因として、外界の速さ、データの必要、内在的な複雑さ、人間に由来する制約、物理法則を挙げる。そして、そのうちいくつかは短期的にしか制約にならないとする。たとえば知能は、「かつては生きた動物での実験を要したことを試験管内で学べるようにする」実験の枠組みを開発するかもしれない（Amodè 2024, "Basic assumptions and framework"）。同じ形で読めば、こうした道は、対象に内在する時間を取り除くのではなく、固有の時間がより短い系へと実験を移す。そのとき求められるのは、その系での結果が対象に引き継がれることである。第3節の試験管内のモデルから引き出された推論が、体内へと引き継がれなければならなかったのと同じである。そしてこれは、速さの問いではなく、そうした系を通じて対象そのものからの訂正を受けうるかという、対象への応答可能性の問いである（第8章第2.3節）。さらに、代替の系では、そこで速く反復できる性質が探究の中心になりやすいとも考えられる。パーカーの説では、モデルがどうあるべきかはそれが用いられる目的に相対的なので（第5節）、その性質と元の問いとの関係を確認めないうまに進めば、速くなったのは元の問いへの答えではなく、速く答えられる別の問いへの答えでありうる（問いの側の置き換えの問題については第4章第4節、問われた問いとは別の問いについて合格したチェックについては第6章第3.2節を参照）。加速は、待つ時間を減らすだけでなく、何を調べるかの選択にも圧力をかけうる。

身体そのものにも、同じ仕分けが当てはまる。人間の研究は、人間の身体に形づくられてきたし、いまでも形づくられており、これからおそらくそうである。機械の研究は、人間の研究が人間の身体に形づくられてきたのと同じ仕方では、身体に形づくられないだろう。だが抽象すれば、人間の研究が身体に形づくられてきたということは、探究者はその探究者自身の身体に形づくられる、という一般的な主張の一つの例である。人間も機械も物理的な存在であり、何らかの物理的な実現をもつ。そのため、少なくとも、探究者が何ができるかが探究者自身の身体に制約されるという弱い意味では、この一般的な主張は機械にも成り立つ。身体が概念の内容まで形づくるという強い意味で成り立つかどうかは、物理的であることだけでは決まらない。すると、身体性をめぐる既存の議論は、どこまでが人間に固有で、どこからがこの一般的な主張に属するかを問いながら読める。シャピロ（Shapiro）とスポールディング

(Spaulding) が「概念化」と呼ぶ主張は、すでに一般的な形をとっている。生物の身体の性質がその生物の獲得しうる概念を制約し、身体異なる生物は環境を異なる仕方理解するだろう、というのである。他方、レイコフ (Lakoff) とジョンソン (Johnson) がこの主張の支えとして挙げる例、たとえば直立する身体から「上」の概念が生じるという例は、人間の身体に固有である (Shapiro and Spaulding 2021, §§2.1, 3.1)。身体が認知の過程に加わるという主張も、人間に限られない。ロボット工学は、世界の記号的なモデルをもたずに、身体と環境の結合からふるまいが生じる機械を作ってきた (Shapiro and Spaulding 2021, §4.1)。探究の過程に当てはめれば、少なくとも次のことは、探究者の身体に一般に依存する。何の差異を感知できるか、どの速さと細かさで操作できるか、訓練によってどの区別を見分けられるようになるか、どこで疲れ、劣化し、故障し、保守を要するか、である。これに対して、感じられる痛みや疲労が注意を引き戻すという仕方や、直立した身体に由来する概念の内容は、探究者の身体一般によるものではなく、人間という実現者の特徴である。したがって問いは、機械の探究者に身体があるかどうかではなく、どの身体が、その探究の何を、どのように形づくるかである。人工の身体や感覚器が、私たちとは異なる認識の組織を支えうることは、第8章第2.3節が論じる。

記録も同じである。記録を増やせば知識が増えるわけではない、という第6.3節の主張は、人間の時間と注意が乏しいことだけによるのではない。チャーニアクは、完全な保持があっても、望む項目を見つけ出す問題は残り、むしろ悪化すると論じる (Cherniak 1983, 175)。変わるのは、記録の将来の使い手が誰か、たとえば別の機械 (同じシステムのインスタンスのこともあれば、作りの異なるシステムのこともある) か、S の人間か、監査者かである。第6.3節で見た、ほかの参加者に尋ねられてはじめて負の結果が記録された事例は、記録の不足というより、誰が何を知っているかについての共有された把握の欠落として読める。記憶の哲学は、そうした把握をもつ集団を、交換記憶の体系として論じてきた (Michaelian and Sutton 2017, §8.2.1)。この機能は機械の構成にも要るが、それを満たす仕方は、互いに尋ね合うことである必要はない。記録が自動的にとられるなら、何を残すかの選別は、何をどの単位で記録するかを定める仕組みの側に移る。

探究者と要素の線も動く。第6.1節で見たものを含め、人間の研究についての諸説では、人間と非人間のあいだに行為者性や能力がどう配分されるかについてどの立場を取るにせよ、実践を記述される探究者は人間であった。そのため、どの人間を要素に数えるにせよ、人間と非人間の線は、少なくとも探究者を非人間の要素から分けることができた。しかし探究者自身が機械であるとき、それさえ成り立たない。探究者自身が、その線の非人間の側に入るからである。ラトゥールは、能力を人の身体の中に取り込むか、人でない身体へ出すかは、「設計者に残された選択の一つ」だとした (Latour 1992, 168)。機械の探究者では、何がその内部に置かれ、何が記録や道具として外に置かれるかが、この意味で所与ではなく、その構成のなかで配分される。配分を決めるのは作り手とは限らず、自らを組み替えるシステム自身でもありうる。ここから、探究者と記録の関係がよりフラットになる、つまり探究する主体とその道具である記録という区別が薄れる、と言いたくなるかもしれない。だが、その言い方はいくつかのことを混ぜる。アクターネットワーク論の対称な記述は、人間と非人間のあいだの能力の配分をあらかじめ決めないための方法上の構えであり (Latour 1996, 9)、人間の研究にもすでに使えた。機械の探究者で新たに変わりうるのは、探究者の状態と記録が形式において違うかどうかと、内と外の配分がどこに置かれるかである。どちらも、機械であることから出てくるのではなく、その構成がどう組織されているかによる。しかも境界は二つの方向に動きうる。記録が探究者の側に取り込まれることもある。探究者の内部が蓄えの一つとして外のものと同じように扱われることもある。ハッチンスは、古典的な記号処理の計算機が、外部の記号を操作する数学者をもとに作られたので、個人の認知のモデルではなく、「人間の行為者が取り除かれた」社会文化的システムのモデルになっていると論じた (Hutchins 1995, 363)。この見方に立つなら、少なくともそうした機械では、探究者は人間の頭に当たるのではなく、人と記録の全体に当たりうる。線をどこに引くかを探究の組織からどう論じるかは、第7章第2節が扱う。

記述がフラットであることは、関係がフラットであることを意味しない。仕事を物に委ねると、ラトゥールの言う「不在の作り手と、ときどきの使い手」のあいだの非対称が生まれる (Latour 1994, 40)。ロー (Law) は、機械もまた、技術的な材料だけでなく、操作者、使用者、修理者を含む異質なネットワークだとする (Law 1992, 384)。サッチマン (Suchman) は、人と機械の境界は所与ではなく引かれる切れ目であり、その切れ目には説明の責任が伴うこと、そして労力を省く装置という自己提示が、その運用にかかる労働の痕跡を消すことを論じる (Suchman n.d.)。そのため、ある機械が道具と記録とともに探究を行ったと言うときには、どこに線を引いたのか、その線が何を見えなくしているのかも述べる必要がある。外部への依存のどれがその機械の組織化の一部を担い続けているかは、第7章第4.2節が問う。役割の差し控え (第6.1節) も形を変える。人工の参加者が役割を断れるかどうかは、その参加者がどう組織されているかの問いになり、人間の拒否が誰も意図しないまま担っていた点検は失われうる。逆に、機械が主導する構成が人間の作業を必要とするとき、その人間は断りうる。第7章第4.4節の病院の事例で、看護師たちがセンサーを拒んだように、である。

分散の理由も分かれる。分業の理由の一部は、一人が保持し注意向けられる量が限られていることにある。別の一部は、計算の手に負えなさや、探究が完了しえないことのように、有限な主体ならどれにも残る。人間の分業では、成員が異なる位置と来歴をもつので、課題を分ければ見方も分かれることが多かった。機械の構成では、それは所与ではない。背景の想定や誤りうる経路が分化していない同じ作りの複製に課題を分けるなら、課題は分かれても見方は分れない (第11章第3節)。異なる基盤やデータから作られた参加者や、違いを作り込まれた参加者に分けるなら、見方も分かれうるが、どれだけ分かれるかは作り方とその後の来歴による。逆に、同じ記録を共有することは、見方を一つにすることと同じではない。第6.1節で見たように、記録された内容は誰にも是認されていない。そのため、探究者の状態と記録が区別されている人工の構成では、同じ記録を参照する参加者どうしが異なる判断を保つことも、仮説を共有の記録に残しながら、それへのコミットメントは一部の参加者だけがもつこともありうる。共有された記録は、それだけでは判断の同質性を意味しない。第6.2節のハッチンスの事例では、組織が見つけた解は、異なる位置にいてそれぞれ局所的に適應する成員から生まれた。機械の参加者のあいだで同じことが起こるかどうかは、参加者がどう異なり、それぞれの位置で局所的に適應しうるかによる。ギアリーは、分散した認知システムに意図をもった認知的な行為者性を与えるのは人間だけだったが、その制限を、科学論の諸分野が共有できる枠組みを得るための理論上の選択だとしている (Giere 2007, 319)。人間がいない部分では、この枠組みはシステムの内部に知る主体を置かない。そこで問いは、結果が S の公的な知識になるかどうかへ移る (第8章)。研究過程の観察についての第6.2節の教訓、つまり観察、再構成、存在論を別々に評価するという教訓は、機械の研究過程を調べる場合にも保たれる。ただし、認知による説明と、物質と集団による説明との争いは、探究者の認知がそれ自体物質的な計算の過程であるとき、形を変えうる。認知による説明が記述するものが人間の頭のなかで起きている限り、それを実験室の記録化や配置と対置することができた。これに対して、内部を記録と並ぶ蓄えの一つとして扱いうる機械の探究者では、認知の語で記述される過程そのものが、物質による説明が指すのと同じ種類の蓄えに対する操

作でありうる。また、描かれた実験室の成員が描かれ方に異議を唱えたような応答が同じ形で起こるかどうかは、調べられる探究者が自らの描かれ方を争いうるよう組織されているかどうかによる。

第6.1節から第6.5節までのどの事例も、あらゆる種類の研究が同じ程度に、あるいは同じ意味で、物質的、身体的、時間的、あるいは分散的であることを立証するものではない。領域横断的な証拠は、これらの役割を比較しうる範囲を広げるが、標本を代表的なものにするわけではない。研究過程の説明の内部に何が属するかは、依然として主張相対的かつ事例相対的であり、第6.5節で見たように、探究を実現するものにも相対的である。 A が機械で S が人間であれば、要素の役割のうち人間という実現者に依存していたものは、その構成がどう組織されているかの問いになり、 S が読み、確かめ、依拠するための条件として、人間の定数は残る。 A も S も機械であれば、身体の実験や役割の差し控えのように、人間の研究では探究者の側から来ていたものに何が当たったのか問われる。また、数学、工学、人文学、自然科学、社会探究のさまざまな形態は、これらの役割を異なる仕方配分しているかもしれない。結果として得られる教訓は、存在論的なものではなく方法論的なものである。すなわち、研究過程の説明は、何が認識的に可能になり、次に何が起こったのかを説明するのに必要な関係を含めつつ、認知、知識、理解、行為者性、研究者としての地位の所在については、さらなる議論に委ねる、ということである。

7. 探究の可能性が変わること

研究は、どの答えを受け入れるかとともに、何を問い、何を試せるかを变えうる。本章で取り上げた事例では、表象が候補を利用可能にし、装置が調査できるものを変え、記録がそれまでできなかった比較を可能にしていた。これらは探究の可能性の変化である。候補形成、追求、評定は、時間的な順序を課さずに、この変化する配置の内部の仕事と区別する。問い（第4章）、候補空間、調査手段を固定したままにする人工システムによる研究のプロジェクトは、この活動の一部を担う。それらを開くプロジェクトは、新たな可能性が使えるようになるまでの仕事を、自ら実現するか、他から得ることも要する。第1.1節は、この仕事が生じる関係を探究の構成と呼んだ。探究するのが機械であるとき、こうした仕事のどこが人間という実現者に依存し、どこが有限な探究者ならどれにも残るのかは、第6.5節で見たように、要素や役割ごとに分かれる。本章で検討した説を通すと、探究の可能性の変化について、いくつかの具体的な帰結が得られる。スタンフォードの議論とカーティス＝トルーデルらの応答を合わせると、すでに特徴づけられた候補の集合をより広く速く探索するだけでは、その特徴づけがおおむね網羅的だと考える理由がない限り、まだ思いつかれていない代替の問題は残る（第3節）。新しい理論を考慮に入れるときの確信の扱いについての提案のもとでは、候補を利用可能にする仕事が、以前の答えへの確信をいったん下げることもある（第4.2節）。ニュルップのモデルでは、偽である見込みが高く、それを容易に示せるからこそ追求に値する候補がある（第4.2節）。そして高橋の区別によれば、探究から人間による遅れを除いても対象に内在する時間は残り、代替の系で加速すれば、その系から対象へ推論する仕事に加わる（第6.5節）。問いが答えられないまま立たなくなることや、 A の未知と S の未知の関係は、問いの側から第4章第5節と第7節で見た。

第III部が描く過程の説明は、一般的な特徴づけからの演繹ではなく、第4章から第6章までの事例と論証に基づく。知識や理解に向けられた試みを照らしているが、それだけでその成功条件を説明するものではない。取り上げた事例は限られており、この説明が、日常的な研究、放棄された研究、未公開の研究、これまであまり取り上げられてこなかった研究など、ほかの実践にどこまで及ぶかは、なお問いとして残る。この過程の分析が第1章の作業的特徴づけとどうかかわるかは、第13章第2節で考える。

候補を利用可能にすることは、まだそれを確立することではない。確立するかどうかは結果にかかわるが、その結果は、候補を手に入れる仕事が終わった後にだけ現れるわけではない。本章の事例では、結果はその仕事のなかですでに得られ、仕事を形づくっていた。ナーセシアンが調べた実験室では、汚染によって実験の半数以上が失敗していたことが、フローループを培養器のなかで動かせるように設計し直すことにつながった（第3節）。リキッド・テンソル実験では、形式化の作業のなかで不正確さが露わになって修復され、証明の理解の仕方も変わった（第3節）。そうした結果が、どの主張、装置、仮定、基準、あるいは表象などにかかわるのかは、結果そのものからは定まらない。

第6章 結果は何にかかわるのか——テスト、失敗、修正

1. 結果から、それがかわるものへ

アルフィーニとコスタの読みでは、ベンジラスとウィルソンにとって説明されるべきものは、ホーン反射器が拾う雑音のありうる原因を除いていく作業のなかで変わっていった（第4章第4節）。この読みに従えば、作業のなかで得られた結果が、問いのほうを改めていたことになる。しかし、結果が何を改めるかは、結果そのものから、候補が手元にあることから決まらない。候補が手元にあることは、それが信じられるべきか、受け入れられるべきか、発展させられるべきか、棄却されるべきか、あるいは未解決のまま残されるべきかを決定しない。テストと呼ばれる操作を遂行することも、その結果が何を意味するかを決定しない。本章が主に取り上げるのは、結果と、それがかわるものとの関係である。

予期しなかった結果は、どのコミットメントを変えるべきかを告げない。チェックに合格したことも、それだけでは、その支持がどこまで及ぶかを告げない。本章は、探究がなお変化しているなかで、結果が何にかかわるのかを見定める仕事が行われ、何がそれを導き、どこで難しくなるのかを問う。第2章第5節は、主張を方法と補助的なコミットメントのもとで試して結果を得、その結果を背景の知識と基準に照らして証拠として読み、その証拠によって態度を改め、改まった態度が知識であるかを問う、という図式を示し、それを行ごとに読んで四つの「段階」と呼んだ。同じ節は、この図式を、評価すべき問いを位置づけるものであって、それを決着させるアルゴリズムではないともしていた。本章が取り上げるのは、そこで第二と第三の矢印とされたものの、すなわち成果からその証拠の関連性へ、そしてその関連性から修正された態度への移行である。ただし本章は、図式の行の並びを、探究がその順に進むことではなく、結果が証拠としてどこにかかり、何を改めるかの依存として読む。ここでの問題は、結果が、問い、候補とその依存関係、そして結果を評価する方法、測定、概念、基準のどれにかかわり、それらをどう制約し、改めるかである。

中心的な主張は限定的なものである。

結果とは、相互依存的な配置の成果である。その成果から認識的態度や実践の変化へと移行するには、解釈と診断というさらなる過程が必要となる。成功も失敗も、それ自身の原因、対象、あるいは適切な修正を特定しはしない。

この主張は、無制限の全体論も、あらゆるコミットメントが等しく修正可能であることも、合理的な診断の不可能性も含意しない。その要点は、結果が何にかかわるかを突き止めることは探究者が行う仕事であって、結果が供給するものではない、ということにある。本書はこの仕事を局所化（localization）と呼ぶ。局所化とは、失敗を、それに寄与したあらゆるもののあいだに分割することではなく、次にどのコミットメントを疑い、調べ、あるいは変えるかを診断することである。この用語法の理由は第2.4節で述べる。この仕事は、スカラー値の成果から機械的に導かれるのではなく、背景知識、誤りのモデル、独立したチェック、介入、実践的技能、確信や見込み、さらなる探究などに依拠する（第2.3節、第4.1節）。何がそれを導き、どこでそれが難しくなるのかは、第2節で検討する。ある種の配置について、それがどこまで達成できるのかは、それ自体が争われている（第2.4節）。

この主張は、単なる因果的複雑性についてではなく、証拠の関連性についてのものである。成果は、無際限に多くの因果的条件をもちうる。より狭い問題は、記録された成果が何らかのコミットメントを変える理由となるのは、問い、**依存関係の配置**（dependency arrangement、第2.1節で定義する）、ありうる成果、候補となる修正のあいだの、表象された、あるいは信頼できる仕方管理された関係を通じてのみである、ということである。世界は結果を制約するが、結果はその証拠の解釈が添えられた状態で到来するわけではない。ここで証拠の関連性というときの証拠は、ケリー（Kelly）が証拠の規範的な意味と呼ぶもの、すなわち何かを信じることをより合理的にするものとしての証拠である。信頼できるしるし（指標）という意味、つまり誰かがそれを知る前から成り立ちうる関係としての証拠ではない。この意味の証拠は信念にかかわるので、本章は主に知識の場合で論じる。探究者が人工システムである場合、本章はそのコミットメントや、結果がそれに何を改める理由を与えるかについて語るが、そのシステム自身が信じたり知ったりする主体であることを前提にはしない（第2章第2節・第2.1節）。理解を目指す探究では、結果はモデルや説明にもかかわり、その評価には第3章第3.4節の理想化の問題が加わる。規範的な意味での証拠を三項関係とみなす哲学者もいる。この見方では、ある結果がある仮説の証拠であるのは、ある背景理論に相対的にでしかない（Kelly 2014, §3）。第2章第5節の図式の第二の矢印、つまり結果と背景の基準から証拠の関連性に至る関係と、第2章第6節で述べた、証拠としての関連性（evidential relevance）についてのロンジーノの説明を参照）。しるし（指標）としての意味では、結果は、まだ誰も突き止めていない欠陥のある要素を、すでに指し示していることがありうる。本章の主張がかわるのは、結果が研究者に何を修正する理由を与えるか、である。このことは、真理や証拠を社会的合意の産物にするものではない。背景知識への相対性は、合意への相対性ではない。ケリーの例では、コプリック斑が麻疹のしるしであることを知らない人は、よりよく知っている観察者がもっている理由を欠いており、関連する医学的知識を獲得すれば、その理由をもつことになる（Kelly 2014, §§3–4）。

議論は、期待に反する結果とその診断から始まり（第2節）、合格したチェックがどこまで主張を支持するかを問う（第3節）。次に、両方の判断を一連の結果にわたって（第4節）、結果を評価する基準そのものの修正に向かう（第5節）。第6節は、この説明が人工の行為者の分析に残すものを述べる。これは論証の順序であって、すべての探究が従う系列ではない。

本章は、探究者が実践のなかで疑いと保護をどのように配分しているかを記述し、そうすることを何が根拠づけるのかについての諸立場を示す。

2. 結果が期待に反するとき

本節は、期待に反する結果がどのコミットメントを修正する理由を与えるのか、そしてそれがどのように定まていくのかを問う。五つの小節は、順に次のものを取り上げる。結果がどこから来るのか、そしてそれが機械による研究に何を示唆するのか（第2.1節）、原理的には配置の全体にかかわるとして

も、なぜすべてが等しく疑わしいわけではないのか（第2.2節）、探究者が実際に最初に疑うのは何か（第2.3節）、何が局所化を導き、それはどこで難しくなり、誰と誰のあいだで争われるのか（第2.4節）、そして焦点となる主張にそもそも届いていない失敗（第2.5節）である。この順序は、局所化を問題にする原理的な構造から、実際の研究が行う優先づけを経て、局所化の実践と、それが出会う失敗の種類へと進む。

2.1 結果はひとつの配置から来る

結果は、それ自身を解釈しない。世界は、何が起こるかを因果的に制約しうるが、それによって、その結果がどのコミットメントを支持し、あるいは掘り崩すのかを告げるラベルを供給するわけではない。期待される観察は、焦点となる主張、背景理論、補助仮定、初期条件と境界条件、物質と生物、環境条件、装置と校正、観察者による読み取り、コードとパラメータ、データの選択と変換、対照、統計モデル、推論規則、そして専門的な有能さの分担などを含みうる配置を通じて、導出され、あるいは予期される。報告される結果もまた、ひとつの配置を通じて産出され、利用可能にされる。本章は、そのような要素の集まりを、所与の結果を産出し解釈することが依存する要素間の関係とあわせて、**依存関係の配置**と呼ぶ。それは、探究の構成（第5章第1.1節）のうち、ひとつの結果が依拠する部分を、テストの側から見たものである。そうした配置を通じて期待を導出することは、候補の帰結と依存関係を明確にすることのひとつの仕方であり、第5章第4節が、追求だけでなく評定にも役立つと述べた仕事である。したがって、探究が一致あるいは衝突として記録しうるものは、世界によって制約されてはいるが理論的・物質的・推論的に媒介された産物のあいだの関係であって、ひとつの仮説と解釈されていない世界とのあいだの直接的な比較ではない。またこの論点は、観察そのものが理論を負荷されているという、争われているテーゼを前提にしない。研究者たちが結果をどう記述するかで一致していても、証拠の関連性を背景理論、事前確率、あるいはパラダイムに相対化する見解のもとでは、それが特定の仮説を支持するかどうか、どれほど強く支持するかについては、なお意見が分かれうるからである（Kelly 2014, §4）。

この論点のひとつの形は、一般にデュエム＝クワインのテーゼとして知られている。これは、射程の異なる二つの主張をひとつの名で呼んだものである（第2章第5節を参照）。デュエム（Duhem）は物理学について、実験は孤立した仮説をテストにかけことはできず、仮説の集まり全体しかテストにかけられないこと、そして予測が外れたとき、実験が教えるのは、その集まりのうち少なくともひとつの命題が誤っているということだけであって、誤りがどこにあるかではないことを論じた（Duhem 1906, Part II, ch. VI, §2）。デュエムはまた、論理と代数を、理論を組み立てるときに守る規則とし、事実による検査は理論の結論に向けられるとした（Duhem 1906, Part II, ch. VI, §7）。そのためデュエムの全体論では、反する結果が疑いを向けうるのは物理学の仮説の集まりであり、論理はそこに含まれていなかったと読める。クワイン（Quine）はこの論点を、地理や歴史から純粋数学や論理にまで至る知識の全体へと広げた。外的世界についての言明は「ひとつの団体としてのみ」経験に向き合う。体系の他の箇所です分に思い切った調整を行えば、どの言明も、何が起ころうとも真とみなし続けることができる。そして、どの言明も、論理の法則でさえ、修正を免れない（Quine 1951, 38–40）。アリュー（Ariew）は、クワインがデュエムに帰したのはこれら二つの主張のうち第一のものだけであり、デュエムは第二の主張をしていないと指摘する（Ariew 2026, §2.1）。本章が依拠するのは、両者が共有するより狭い主張だけである。すなわち、いくつかのコミットメントから導かれた期待と衝突する結果は、それだけでは、そのどれに欠陥があるのかを特定しない。

デュエムが物理学について記述した構造は、物理学の外にも現れる。統計的なモデル検査がその一例である。ゲルマン（Gelman）とシャリジ（Shalizi）は、事後予測チェックを、モデルの仮定のテストとして扱う。その仮定には事前分布も含まれ、事前分布は尤度と組み合わせずはじめてデータについての含意をもつ。二人の説明では、モデルがそうしたチェックに合格しなかったとき、分析者は、何がうまくいかなかったのか、問題がどこにあるのかを突き止めてから、モデルに手を入れるか、設計を根本からやり直すことになる（Gelman and Shalizi 2013, 12–13 and n. 12, 25）。有意性検定が報告する誤り率は検定自体が調べない仮定に依存するという大塚（Otsuka）の主張を含め、これに関連する大塚の議論は第2章第5.3節が述べている。

予測が外れても、それだけでは焦点となる仮説は反駁されない。期待がコミットメントから演繹されているなら、論理が告げるのは、その連言のどこかが偽だということである（統計モデルのもとで起こりにくいとされていたにすぎない結果は、それすら確定させない）。その仮説を退けるには、残りのコミットメントは成り立っていた、という主張を加えなければならず、その主張もまた誤りうる。「仮説が反証された」という言い方は、この主張を暗に含んでいる。同じことは成功についても言える。予測が当たっても、それだけでは、どのコミットメントがどれほど支持されるかは定まらない（第2.2節、第3.2節）。誤った仮説が誤った補助仮定と組み合わせず、正しい予測を生むこともありうるからである。また、二つの仮説のどちらが正しいかを一つの実験で決着させる「決定実験」は、そのままの形では成り立たない。デュエムによれば、水中と空気中の光の速さを比べたフーコーの実験が裁いたのは、放射説と波動説という二つの仮説ではなく、それぞれを含むニュートンとホイヘンスの光学という二つの体系の全体だった。しかも、光が粒子の流れでも波でもない何か別のものである可能性は、排除されていなかった（Duhem 1906, Part II, ch. VI, §3）。そのため、結果から誤りの所在を読み取ることは、結果そのものが与えるのではなく、結果に加えて行われる仕事になる。これが、第1節で局所化と呼んだ仕事である。

機械が研究を行うとき、反する結果がかかわる配置には何が入り、帰属はどこで、誰によって行われるのか。まず、配置の中身が変わる。機械の研究では、配置に、システムが自ら書いたコードやデータの処理、呼び出した道具、他のシステムの出力、評価の仕組みなどが入りうる。その多くは、システム自身が組み立てたものや、それを組み立てた者にも全体が見渡せないものでありうる。そうした配置では、焦点となる仮説が誤っている可能性に比べて、配置の残りのどこかが誤っている可能性が小さいとは言えない。次に、上で述べたことは、機械による結果の報告にもそのまま当てはまる。システムが、ある仮説は反証されたと報告するとき、その報告は、配置の残りは成り立っていたという主張を暗に含んでいる。成功の報告も同じく、支持がどこに及ぶかについての主張を含む（第3節）。そのため、実験の結果をそのまま焦点となる仮説への判定として受け取るシステムは、局所化という仕事を担っていない。そうしたシステムを作っても、それだけでは、結果から何を学ぶかを自ら見定めるシステムを作ったことにはならない。さらに、試行の規模が変わる。多くの試行が自動で回されると、配置に由来する違いと仮説に由来する違いを区別しないまま、判定が積み重なりうる。

こうした帰属を誰が確かめるかは、 A と S の組み合わせによって変わる。 A が機械で S が人間であれば、 S の成員は、システムの報告に含まれる帰属を、配置の記録、つまり何がどの順に組み立てられ、何が確かめられていたかに照らして確かめることになる（第5章第6.3節）。記録がそれに足りなければ、その成員は、報告された反証や確証がどこにかかわるのかを、記録からは確かめられない。ただし、報告に依拠する根拠は、記録から再構成すること以外にもありうる（第11章第5.2節）。 A も S も機械であれば、帰属を確かめる側もまた配置をもつ探究者であり、確かめることそのものが同じ問いを引き継ぐ。

どちらの場合にも、デュエム＝クワインの論点が機械による研究について示唆するのは、結果を受け取るときに問われるのが、結果が何であつただけでなく、それが配置のどこにかかわると誰がどう判断したかでもある、ということである。第2.4節の末尾と第4.3節は、この問いに実際の記録から戻る。

2.2 すべてが等しく疑わしいわけではない

原理的には、反する結果は配置の全体にかかわり、どこに誤りを帰すかは結果からは決まらない。それでも実際の研究は、何かを先に疑い、何かを保ち、どこから調べ始める。ある要素を確かめるには、ほかの要素を固定してそれに頼らなければならないからである（船をドックで組み立て直すことができず、海の上で造り直すしかないという、第5.2節で見るノイラートの比喻を参照）。そのうえ、注意に限りのある探究者は、多くを同時に調べることもできない。その優先づけは、さまざまなものに支えられうる。各要素への確信の度合い（第5章第4.2節、第2.3節）、ある要素を改めたときに、それとつながる他のものをどれだけ見直すことになるか、理論の定着を支える個人的・社会的な目標（第2.3節）、実践者が自分のモデルについてもつ暗黙の知識（第2.4節）などである。このうち個人的・社会的な目標は、チン（Chinn）とブルーワー（Brewer）が科学史のエピソードと学生についての研究から引き出したものである。機械が研究を行うとき、何がその役を果たすのか、あるいは果たすものがあるのかは、第4.2節で立ち返る問いである。本小節から第2.4節までは、こうした優先づけが何に依り、どこまで根拠をもちうるかを見る。本小節はまず、配置の要素がどれも同じ仕方では結果にかかわるわけではないことを見ていく。そのうえで、デュエムの論点もクワインの論点も、あらゆるコミットメントが等しく修正に開かれていることを含意しない理由、それらからしばしば引き出される全体論を退ける哲学者がいる理由、論理だけでは決まらない選択をデュエム自身がどう記述したか、そして同じ問いが、反例がかかわりうるコミットメントが完全に列挙されているように見える数学の内部で、どのように生じるかを述べる。

配置のすべての要素が、結果に対して同じ仕方ではかかわっているわけではない。そのため、誤りを局所化することは、互いに取り替えのきく項目から一つを選ぶことではない。たとえば、創業の計画 Open Source Antibiotics（第5章第2節）の実験ノートでは、シリカのカラムによる精製で黄色の固体として得られた物質を、性質の異なる逆相のカラムにもう一度かけると、二つに分かれた。一つは、化学者が「DEBROMINATED CORE」（臭素の外れた母核、つまり副生成物）と記した白色の粉末であり、もう一つは、それよりずっと少ない量の目的生成物である。目的生成物は次の実行でスケールアップされた (Tse 2023, EGT 522-1)。最初の黄色の固体について得た読み取りは、測定としては信頼できるものでありながら、目的化合物ではなく混合物についての読み取りでありえた。そうした読み取りが期待に反していたとしても、疑うべきなのは仮説でも装置でもなく、その結果が何についての結果だったのか、でありうる。また、結果がひとつの配置から来ると認めることは、その配置が丸ごと修正されなければならないということも、探究が宇宙におけるあらゆる因果的条件を表象しなければならないということも意味しない。記述が何を表象しなければならないかは、探究者やシステムが行うと主張する推論と訂正によって定まる。

どの依存関係が際立つかも、問われる問いによって変わる。試薬の保管履歴が重要になるのは一部の問いについてだけであり、分類体系は、あるエピソードでは背景であっても、別のエピソードでは焦点となる対象でありうる。測定を生み出す過程そのものも、同じ記録に別の問いが向けられるとき、対象となりうる。たとえば、ボストンの二つの病院の電子健康記録では、午前4時に測定された白血球数が正常だった患者は、午後4時に測定された白血球数が異常だった患者よりも3年生存率が低く、174の臨床検査のうち118では、検査が指示されたタイミングのほうが検査の結果よりもよく生存を予測した。著者たちはその多くを、いつ検査が指示されるかを定める診療の過程に帰しつつ、一日のうちの生物学的な変動も寄与しうることを認めている。その研究では、検査結果そのものには記録されない診療の過程が、分析の対象となった (Agniel, Kohane, and Weber 2018)。

デュエムは、この狭い主張を物理学について述べ、生理学に見いだした論理的により単純なテストと対比した。ただし同時に、物理学の装置に依拠するどの科学も物理学の理論を前提していることにも注意している。その後の議論は、この主張を他の分野の理論的なテストへと広げてきた (Duhem 1906, Part II, ch. VI, §1; Stanford 2023, §2.1)。この主張は、論理的に可能なあらゆる修正が等しく合理的であることを含意しないし、どの修正を採るかはまったくの取り決めの問題だとする急進的な規約主義にもならないし、あらゆる構成要素が常に等しく疑わしいことも含意しない。クワインのより広い全体論も、これを含意しない。どの言明も保持されうるという主張と並んで、クワインは、ある言明が体系の周縁に近いということを、ある言明を他の言明より修正に選びやすいという実際上の相対的な見込みを反映した、緩やかな結びつきにすぎないものとして記述し、実際の修正を導くもののなかに、「全体の体系をできるだけ乱さないようにするという私たちの自然な傾向」、保守性、そして単純さの追求を挙げた (Quine 1951, 40–41, 43)。そのような保持はどれも合理的に擁護できるというより強いテーゼこそ、ラウダン (Laudan) が、十分に支持されたことはないと論じたものである。スタンフォード (Stanford) の紹介するラウダンの説では、決定不全性を支持する議論が示すのは、反対の証拠への多くの応じ方が論理的に可能だということにすぎない。演繹論理だけでは一つの応じ方を選び出せないとしても、科学で用いられる拡張的推論、つまり前提に含まれていない結論へ進む推論の原理を加えれば選び出せるかもしれないのである (Stanford 2023, §2.2)。体系をできるだけ乱さないという傾向が効く理由の一つは、クワイン自身が、反する経験の影響が広がる理由として述べたことのうちにある。ある言明を評価し直すことは、論理的なつながりのために、他の言明を評価し直すことを伴う、というのである (Quine 1951, 39)。そのため、多くのものがよりかかってきたコミットメントを改めようとすると、それによりかかってきたものを見直すことになり、それに代わる整合的なものが用意できなければ、改めた後に置かれるものも成り立たない、という場合がある。いくつものチェックに耐えてきたものを疑うことには、そうした場合、大きなコストがかかりうる。第2.3節で見るデュエムの例は、その一つとして読める。一様な媒質のなかで光は直進するという原理は、ほとんど直線の物理的な定義の役を果たしていたが、回折の効果を何らかの誤りに帰するのをやめたときに放棄された。

後の哲学者のなかには、論理的な論点を受け入れつつ、そこからしばしば引き出される全体論を退ける者もいる。ソーバー (Sober) は、単純な論理的論点を論争的な認識論的論点から切り離し、前者を受け入れ、後者を退ける。仮説が補助仮定なしに観察上の予測を行うことはまれだが、だからといって、観察が確認したり反確認したりするのはその連言だけだということにはならない。補助仮定はしばしば独立にテストされるので、結果は、共有された補助仮定に手を触れないまま、対立する仮説のあいだを識別しうるからである。ソーバーの例では、妊娠検査は、ある女性が妊娠しているかどうかについての証拠を与えるが、その検査の信頼性に対するその女性の確信は変えない。その信頼性をテストするには別の実験が要る。ソーバーは、認識論的全体論は「圧倒的大多数の場合に」誤っているとしつつ、連言を切り離せない状況がありうることも認める (Sober 1999, 54–56)。尤度による後の扱いは、測定器具の較正とテスト手続きの妥当性確認における典型的な状況を記述している。それらの誤りの特性は「まず確かめられ、それからその手続きが新しい個体に適用される」。ガリレオ (Galileo) は望遠鏡を地上の物体で確かめてから木星に向けたのであり、そうして行われた観察は、木星の衛星については多く

を告げ、望遠鏡についてはほとんど何も告げなかった (Sober 2004, 230–231)。証拠が連言にどうかかわるかについて全体論が正しいかどうかは、「対立する連言が何であるかに決定的に依存する」とソーバーは結論する。全体論にはその場所があるが、科学的探究においては、本人の言葉では「圧倒的に、規則ではなく例外」である (Sober 2004, 241)。全体論がどこまで成り立つかについてのこの分析は、補助仮定を独立にテストできることに依拠している。第2.4節は、その応答がある種の配置では利用できなくなるという議論を取り上げる。

第3章第7.2節で論じたエルギン (Elgin) の議論は、同じ方向の一般的な理由を与える。ほかのよく確かめられたまとまりのなかで独立に支持されているコミットメントは、誤りの所在である見込みが低い。研究者は、ある依存関係を保護し、別の依存関係を調べる強い理由をもちうる。そうした説では、独立したチェック、校正の履歴、頑健性、誤りのモデル、背景理論、メカニズム、コスト、そして後の介入はすべて、何を修正するかを突き止めるのを助ける。ただし、相互調整の成功で局所化を正当化できるかどうかは争われている (第2.4節)。いずれにしても、局所化は、失敗した (failed) という語のうちに含まれてはいない。

何が選択を導くのかについてのデュエム自身の答えは、規則ではなかった。そしてその答えは、選択を恣意的なものとも扱わなかった。デュエムによれば、実験が理論と矛盾するとき、実験は何を変えるべきかを告げず、どんな絶対的な原理もその探索を導かない。基本仮説を補正の追加によって守る物理学者と、基本的な想定を変える物理学者とは、互いを非論理的だと責めることができない。デュエムはさらに、論理に反しない意見のなかにも不合理なものがあると述べ、二人の物理学者のあいだの選択を、良識 (good sense) と呼ぶものに委ねた。良識の理由は論理のような厳密さを欠き、すべての人に同時に明らかになるわけでもない。ビオ (Biot) とフレネル (Fresnel) は、論理が決着させず、まさにこの理由のためにあれほど長く続いた意見の不一致の例として挙げられている。ビオは補正と補助仮説によって光の放射説を擁護し、フレネルは新しい実験によってそれに反対し続け、ついに一方が退いた。ビオはフォーコー (Foucault) の実験の後に放射説を放棄したが、デュエムの説明では、論理だけがビオにそれを強いたのではない (Duhem 1906, Part II, ch. VI, §10)。この事例はまた、時間の問い、すなわちそうした不一致がいつまで合理的に続くのかという問いを提起する。第4.1節はこの問いに立ち返る。

ラクトシュ (Lakatos) のある対話は、反する結果がかかわりうるコミットメントが完全に列挙されているように見える数学の内部で、本節の問いを立てる。第2章の冒頭で触れた、多面体についてのオイラーの予想の歴史を教室で再構成したこの対話では、マスグレイヴ (Musgrave) とピグデン (Pigden) の整理によれば、反例は三種類に分けられる。一つは、証明の結論を反駁するが前提のどれも反駁しない反例であり、ラクトシュはこれを大域的だが局所的でない反例と呼ぶ (本章の局所化とは別の、ラクトシュの用語である)。もう一つは、前提の一つを反駁するが結論は反駁しない、局所的だが大域的でない反例であり、残る一つは、その両方を反駁する反例である。この分析では、予想への反例は、証明を捨てる理由ではなく、証明を改善する理由となる (Musgrave and Pigden 2026, §3.1)。アバディーン (Aberdeen) が対話をたどるところでは、そうした反例へのひとつの応答は、それはそもそも多面体ではなく、反例ではなく怪物だと宣言して、証明された定理を守ることであり、新しい怪物が現れるたびに、さらに定義し直すことが提案される。ラクトシュはこれを怪物排除と呼び、「巧みなこともあるが、常に場当たりのな」定義し直しだと性格づける (Aberdeen 2019, 5, 9)。この対話は、本節がテストについて立てる問いを、証明について立てているものとして読める。すなわち、反する結果が、構造をもったコミットメントのどの部分にかかわるのか、そして残りを保護することがいつ場当たりのになるのか、という問いである。この対話はまた、各ステップが書き出された証明でさえ、反例がかかわるあらゆるコミットメントを列挙しているとは限らないことも示す。述べられた補題のどれも反駁しない、予想への反例は、探究者たちを証明へと送り返し、「隠れた補題」を掘り起こさせる (Musgrave and Pigden 2026, §3.1)。

2.3 最初に疑われるもの

第2.2節は、結果が期待に反するとき、探究者が実際にどのコミットメントを疑うのかには答えなかった。第5章第4.2節の二つめの見方では、探究者は確からしいと思える方へ進み、その確信の度合いは、探究者が立っている基準と、何を信賴しているかによる。同じ確信の度合いは、結果が期待に反したとき、どこを疑うかにもかかわる。デュエムの問題のベイズ的な扱いは、このことを、連言のある項 (連言肢) ではなく別の項に責めを負わせることを何が根拠づけるのか、という問いへの答えとして明示する。ストレヴンス (Strevens) が展開する扱いでは、反駁は、証拠が仮説について告げる他のことから切り離して、それだけとして考えれば、事前により疑わしかった連言肢に疑いを集中させる。連言が偽であることだけを条件とすると、仮説の確率は、事前に確からしかったほど小さくしか下がらず、補助仮定が確からしかったほど、通常は大きく下がるからである (Strevens 2001, §3)。メイヨー (Mayo) は、ドーリング (Dorling) によるより早いベイズの再構成に反対して、それは逆向きにも走らせることができる——仮説と補助仮定を入れ替えて同じ信念の度合いを割り当てる人は、逆の判定に達する——と論じ、補助仮定に責めを負わせることが根拠づけられるのは、それに欠陥があるという主張そのものが厳しいテスト (severe test) に合格したときだけだとする (Mayo 1996, 106, 108–109)。厳しいテストとは、その主張が偽であったなら高い確率で不合格となったはずのテストである (Mayo and Spanos 2006, 328–329)。確信の度合いだけでそうした帰属が根拠づけられるかどうかは、このように争われている。本小節が問うのは、探究者が実際に何をするかであり、その答えを緩やかに記述する。

配置のどの部分が最初に疑われるかは、経験的にも研究されてきた。チンとブルーワーは、科学史のエピソードと、心理学および科学教育の実験的研究とに依拠して、保持されている理論に反するデータへの七つの応答を区別する。データを無視すること、データを拒否すること (たとえば方法上の誤り、偶然の変動、あるいは不正として)、データを理論の領域から除外すること、データを保留しておくこと、データを再解釈すること、理論に周辺的な変更を加えること、そして理論を変えることである。チンとブルーワーは、これらが、データを受け入れるかどうか、データを説明するかどうか、またどう説明するか、理論を変えるかどうか、という三つの決定の、心理学的にありそうな組み合わせを尽くしていると論じる (Chinn and Brewer 1993, 4, 12)。二人はまた、理論を保持する六つの応答に有利に働く条件を、網羅的とは示さないリストのかたちで報告している。一つは先行する理論の定着 (entrenchment) であり、その理論が受けている証拠による支持、それが関与する説明の範囲、それが役立つ個人的・社会的な目標によって特徴づけられる。もう一つは、もっともらしい代替理論、しばしばもっともらしいメカニズムがないことである。三つ目はデータにかかわる。信頼性を欠くデータ、曖昧なデータ、あるいは方法上の批判や再解釈に開かれた単一の研究に依拠するデータである (Chinn and Brewer 1993, 13–26)。チンとブルーワーは、教育を論じるなかで、変則的なデータの信用を落とすほうが理論を変えるより合理的であることも多いと付け加える (Chinn and Brewer 1993, 27)。条件文を用いた日常的な推論についての実験で、ウォルシュ (Walsh) とジョンソン＝レアード (Johnson-Laird) は、不整合を説明するよう求められた参加者の多くが、個別の

事実ではなく一般的な条件文のほうを修正し、また大半の試行で、最小限を超える変更を行ったことを見いだした。ここでの最小限は、著者たちが、真理値の変わる命題の数を数えることによって測るものである。二人は、人は変更を最小にするよりも、不整合を説明しようとする論じる (Walsh and Johnson-Laird 2009, 626–630)。どちらの研究も、現役の研究者を観察したものではない。チンとブルーワーの科学者の例は科学史から取られており、ウォルシュとジョンソン＝レアードの参加者は日常的な条件文について推論した。

研究の進行中に残された記録からは、どこが最初に疑われたかを、より直接にたどることができる。2011年、レッドフィールド (Redfield) は、細菌 GFAJ-1 がリンの代わりにヒ素を使えるという主張の追試を、公開の研究ブログに記録しながら始めた。細胞は寒天の培地では増えたが、液体の培地では増えなかった。最初の投稿は、培地に使った水道水、新しい培養管、緩いふたから入る空気を疑い、三日後の投稿は、水でも管でもパラフィルムでもないとして報告した。そして、ある読者がカリウムを示唆したのをきっかけに、二つの誤りを見つけた。一つは自分のリン酸の原液の誤りであり、もう一つは、カリウムを含まない公刊されたレンピの誤りである (Redfield 2011a; Redfield 2011b; Redfield 2011c)。のちに、ヒ素を加えリン酸を制限したフラスコで細胞がよく増えたとき、つまり追試している主張に有利に見える結果が出たとき、レッドフィールドが最初に考えた説明は、フラスコを準備するときに自分が手違いをした、というものだった (Redfield 2011d)。疑いはまず自分の側の配置に向かい、追試の対象である主張が調べられたのは、その配置が動くようになってからだった。ニュートリノが光より早く着いたように見えるという OPERA 共同研究の報告は、別の形を示す。共同研究は、まだ知られていない系統的な効果を調べると書き、異常の理論的な解釈は意図して控えた (OPERA Collaboration 2011, 22)。五か月後に挙げた二つのありうる効果の一つは、外部のGPS信号を主時計に運ぶ光ファイバーの接続だった (CERN 2011, update of 23 February 2012)。その計時の系のファイバーの遅延は、往復の測定で求められていた (OPERA Collaboration 2011, 13)。まず疑われたのは結果そのものだったが、問題になったのは、確かめられていた要素だった。三つめの記録は、何が放棄されるかが、手元にある説明に依存することを示す。2021年、transformers ライブラリの利用者が、勾配累積を使うと、同じ大きさのバッチで訓練した場合と学習曲線が違うと報告した。保守者は、違いは数値の精度で説明できると判断し、利用者はその説明を受け入れ、イシューは閉じられた。一か月後、利用者が違いの源の一つを、訓練例を長さでまとめる仕方に見つけ、イシューは開き直された (Hugging Face Transformers 2021)。2024年、同様の報告にも同じ精度の説明が出されたが、報告者は、完全な精度で再実行しても差が残ることを示してそれを退けた。差はその後、ライブラリの既定の損失が累積したバッチにわたって平均をとる仕方にまでたどられた (Hugging Face TRL 2024; Hugging Face 2024)。いつでも手に入る説明が一度は問いを閉じ、既定値として信頼されていた要素が誤りの所在だった。

これらの研究と本章の事例を合わせても、それらが支持するのは緩やかな記述だけである。結果そのものは、しばしば最初に疑われる候補のひとつである。それを越えると、疑いは、探究者が立っている基準に照らしてより確かでないと思われる要素に集まる傾向がある。しかし、何が放棄されるかは、衝突のどのような説明が利用可能か、そして代替が手元にあるかどうか、つまりどの候補が考慮されているか (第5章第4.2節) にも依存する。この記述は傾向を述べるものであって、最も確かでない要素だけが修正されるという規則ではない。デュエム自身がその両面を観察していた。デュエムの書くところでは、テストに入る理論的要素のなかには、その時代の物理学者たちが点検なしに受け入れることで一致しているものが常にあり、理論を修正しなければならない物理学者は、きつと修正をほかのところに向ける。しかし論理はそれを強くないし、誤りの原因に訴えることを拒み、共通の同意によって触れてはならないとされている命題を改めることが、理論に新しい道を開くこともある。デュエムが挙げる例は、光は一樣な媒質のなかでは直進するという原理である。この原理はほとんど直線の物理的な定義の役を果たしていたが、グリマルディ (Grimaldi) が観察した回折の効果を、研究者たちが何らかの誤りの原因に帰するのをやめたとき、放棄された (Duhem 1906, Part II, ch. VI, §8)。

稼働中の実験室の観察は、この記述だけでは決まらない分岐を示す。ダンバー (Dunbar) は、八つの分子生物学・免疫学の実験室における会議やその他の活動を追跡した。報告された事例では、予期せぬ知見は、未知の原因、方法上の問題、誤り、あるいは潜在的に新しいメカニズムへと、さまざまに帰属された。その後の対応には、追試、さらなる分析、新しいプロトコルの採用、そして現行のプロトコルの放棄が含まれていた。正確な頻度は、その標本とコーディング方式に固有のものである。その限定された標本の範囲内で、これらの観察は、分岐する診断問題の存在を支持する。すなわち、研究者たちは予期せぬ結果を常に主要仮説の修正へと変換したわけではなかった (Dunbar 2000)。彼らの追試でさえ方法の小さな変更を含んでおり、上級の研究者たちは対照条件での予期せぬ知見にとりわけ注意を向けていた。どちらも主要仮説以外の所在にかかわる (Dunbar 2000, 53)。ダンバーは、分岐がどのように扱われたかも報告している。科学者たちは、予期せぬ知見を取り上げ、その根底にある出来事の因果モデルを組み立てようとし、そのモデルができてから次の一步を決めた。また、実験を設計したり修正したりしているときには、彼らの類推は、非常によく似た実験や同じ生物へのものだった。ダンバーは、対照条件での予期せぬ知見に注意を向けるよう勧める。それは概念上の問題か実験上の問題のどちらかを反映しているからである (Dunbar 2000, 53–54, 57)。これらの観察が示すのは、コミットメントが疑われる順序というより、帰属が分岐するということと、研究者たちがその分岐をたどっていく仕方のいくつかである。

疑いがどこに向かうかには、目の前の結果の外から来る確信がかかわることもある。結果がどちらの帰属も許すとき、どちらを選ぶかには、確信や見込み、さらには何が本質的かについての見方までがかかわりうる。ある考えにコミットし続けることも、その一つの形であり、うまくいかない結果を、考えそのものではなく、その考えをまだうまく実現できていないことに帰し続けるという意味で、それ自体が結果の帰属である。そうした判断が積み重なって、どの考えが持ち続けられるかを左右しているのかもしれない。ニューラルネットについてのヒントンの回想と、Transformer の著者たちの回想は、その例として読める。それらは、コミットメントを保護することがいつまで合理的であり続けるかという問いとあわせて、第4.1節で取り上げる。

2.4 何が局所化を導き、どこでそれが難しくなり、誰と誰のあいだで争われるのか

疑いが緩やかに配分されることを認めるとして、何を修正するか探索を何が導きうるのか。本小節は、この問いについての諸立場を示す。結果が装置ではなく現象を反映していると論じるために実験者が何を用いるのか、相互調整の成功で結果を正当化できるのか、局所化を導く資源は多くの構成要素がともに調整された配置のなかでも保たれるのか、誰が局所化を行うのか、そして配置の各部分が異なる当事者に属するとき何が起こるのか、である。そのうえで、本章が局所化という語をどういう意味で使うかを述べ、人工研究システムについて何が見えてくるかを考える。

実験の哲学は、第2.2節で挙げた資源のいくつかを、結果が装置ではなく現象を反映していると実験者が論じるための戦略として記述してきた。そのなかには、装置が既知の現象を再現することを確認すること、もっともらしい誤りの源と代替的な説明を排除すること、そしてよく検証された (多くのテストに

耐えてきた) 理論に基づく装置を用いることが含まれる (第5章第6.1節を参照)。フランクリン (Franklin) とペロヴィッチ (Perović) の書くところでは、第一の種類のチェックに失敗したときには、その装置で得られた結果を疑うよい理由がある。二人は、これらを、規則としてではなく、物理学者がそのなかから選ぶ戦略として示す。それらは排他的でも網羅的でもなく、そのどれひとつも、またそれらのどんな固定された組み合わせも、結果が妥当であることを保証しない、というのである (Franklin and Perović 2023, §§2.1.2, 2.2.4)。こうした戦略がそもそも働きうるのかについては、より根本的な疑いも出されてきた。フランクリンとペロヴィッチの紹介によれば、コリンズ (Collins) は「実験者の後退」と呼ぶ議論を展開した。正しい結果とは、きちんと働く装置で得られた結果であり、きちんと働く装置とは、正しい結果を出す装置である。装置がきちんと働いているかを決める形式的な基準はなく、この後退は最終的に、共同体のなかの交渉によって断ち切られる、というのである。重力波の検出をめぐる論争を論じて、フランクリンはこれに反対する。批判者たちはデータと解析のプログラムを交換し、既知のエネルギーの信号を入れて装置を較正しており、共同体は理由のある判断を下した、というのである (Franklin and Perović 2023, §2.2.1)。この論争は、局所化の判断もまた、それ自身を解釈しない結果でありうることを思い出させる。誤りは装置にあるという判断も、その判断を導いたチェックの配置から出てくるからである。そうだとすれば、局所化がどこで止まるかは、局所化の論理だけでは決まらないのかもしれない。第5章第6.5節で見た、どの探究に取りかかるかを探究で決めようとするが生じる後退と、似た形である。何がそれを止めるのか、理由のある判断なのか、交渉なのか、その両方なのか、ここで争われていると読める。

相互調整による安定化で結果を正当化できるかどうかは、それ自体争われている。フランクリンとペロヴィッチが伝える、ジャコモ・モルブルゴ (Giacomo Morpurgo) による自由クォーク探索についてのピッカリングの説明では、連続的に分布する電荷の観測値は、モルブルゴの予期に反していた。そこで、コンデンサーの板の間隔を広げ、装置についての理論的な説明を改めると、整数電荷の測定値が得られ、装置、装置の理論、現象の理論が一致するに至った。二人はこれを、第5章第6.1節で見た、ピッカリングの抵抗と調節の弁証法として紹介する (Franklin and Perović 2023, §2.2.4)。フランクリンとペロヴィッチは、このエピソードを別様に読む。ピッカリングによるこのエピソードの1987年の説明を論じて、二人は、電荷の量子化はミリカン (Millikan) 以来強く支持されていたので、モルブルゴの測定に疑いを投げかけたのは、どのような理論的モデルが提示されていたにせよ、装置がその既知の結果を再現できなかったこと、すなわち重要な実験的チェックに失敗したことだったと論じる。装置が既知の測定を再現できるようになってはじめて、分数電荷の探索においてそれを信頼することができた、というのである。上で見たピッカリングの1995年の説明に対しては、フランクリンとペロヴィッチは、ピッカリングが理論、装置、現象の所与の相互調整が正当化されるかどうかを問わず、安定化を達成するどんな調整も受け入れ可能なものとして扱っているように見える、と異議を唱える (Franklin and Perović 2023, §§2.2.2, 2.2.4)。こうして二つの読みは、何が局所化の根拠になるかについて異なる。フランクリンとペロヴィッチが伝えるピッカリングの説明では、一致をもたらす調整に重みが置かれ、二人の批判では、調整を方向づける、独立に支持された結果に重みが置かれる。どちらも、結果そのものが誤りの所在を特定するとは考えない。

ハンフリーズ (Humphreys) は、明示的な構築の前提から導かれた理論的テンプレート、すなわち一般的な数学的表現について、検査の前から手に入る導きを描いている。修正集合は、構築の時点ですでに偽だと知られ、あるいは疑われている理想化、抽象化、近似などの前提を、どう改善しうかの示唆を与える。ただし、その示唆がつねに明示的に定式化されるわけではない (Humphreys 2019, §3, advance-version 3 and n. 8)。クヌーティラとレットガース (Knuuttila and Loettgers) は、これに対応する計算モデルの説明を紹介している。構築の前提には存在論、理想化、抽象化、制約などが含まれ、修正集合は、データに合うようにどう調整しうかを前もって示すという説明である (Knuuttila and Loettgers 2023, §2)。二人はまた、存在論を変えろという、より根本的な可能性もハンフリーズが挙げていと報告する。ただしハンフリーズは、それが新たなテンプレートの構築に近づくことを断っている (Knuuttila and Loettgers 2020, manuscript 31, n. 17)。有賀による紹介では、ハンフリーズは、テンプレートの構築や改変に目を向けず、最初から出来上がったものとして扱う、所有なき視点を批判する (Ariga 2009, 85)。ここで述べた導きは、明示的な前提から導かれたモデルにかかわる。次に取り上げるレンハルトとウィンズバーグの議論は、ともに調整された多くの構成要素からなる配置にかかわる。

第2.2節で論じた、デュエム的な状況への応答、すなわち補助仮定を独立にテストすることは、まさに、気候モデリングの哲学者の一部が、配置が大きくなるにつれて利用できなくなると論じるものである。オラフリン (O'Loughlin) の再構成によれば、レンハルト (Lenhard) とウィンズバーグ (Winsberg) は、気候モデルの研究者が自らのモデルについて、二人が分析的理解と呼ぶもの、すなわち各構成要素がモデルの成功と失敗にどの程度寄与しているかを特定する能力を欠いており、構成要素どうしの相互作用が複雑であるために、各構成要素の長所と短所を独立に評価することはできないと主張する (O'Loughlin 2023, §§2-2.1)。オラフリンが報告する二人の説明では、モデルの出力は、各モジュールの結果を組み合わせることによってではなく、実行中のモジュールどうしの相互作用によって生み出される。パラメータ化は、すでにモデルに組み込まれているパラメータ化を前提として作られ、テストされる。そして、その時点のモデルを改善するために作られたその場しのぎの修正 (クラッジ) は、ほかの構成要素がそれに依存するようになるにつれて定着していく。そうなると、新しい構成要素とモデルの残りの部分とを別々に調べても、その構成要素がモデルのふるまいをどう変えるかはわからない。構成要素のどの相互作用がその変化をもたらしたのかを見分けられないからである (O'Loughlin 2023, §2.1)。二人は、三つのモデル相互比較プロジェクトを、この見方の証拠とみなす。それらのプロジェクトでは、性能の違いを構成要素に帰属させるという目標が、繰り返し先送りされ、あるいは縮小された。同時に二人は、いくつかの誤りの出所が突き止められたことも認めている (O'Loughlin 2023, §2.2)。

オラフリン (O'Loughlin) は、同じ相互比較と後の相互比較における、モデル間の不一致の診断をもって応答する。そのひとつでは、いくつかのモデルで向きが逆になっていた、モデルから導かれる海洋の熱輸送が、それらのモデルによる雲の放射効果のシミュレーションにまでたどられた。研究者たちが最初に雲を疑ったのは、モデルが雲について一致しないことが知られていたからであり、この診断は、観測に基づく値で置き換えることによって確かめられた。オラフリンは、困難はモデルよりも相互比較のほうにあるのかもしれないと論じる。体系的でない仕方でも互いに異なる何十ものモデルがあり、それぞれがどう作られ調整されたかについての知識は諸機関に散らばっている、というのである。また、レンハルトとウィンズバーグの分析的理解の説明のように、各構成要素の寄与の程度を求める基準は、構成要素を名指すどんな診断をも割り引いてしまいうると論じる。名指された構成要素は、モデルの残りの部分とともに始めて誤差を生んだのだから、それが責めのすべてを負うのか一部だけを負うのかは、いつでも問いつけるからである。オラフリンは、そうした診断の成功は、実践者が自分のモデルについてもう、暗黙的だが実地に働く知識、「一種のデュエム的な『良識』」を反映しているのかもしれないと示唆する (O'Loughlin 2023, §§1, 2.3, 3.2, 4.1, 5)。

局所化は、研究者の明示的な言語的反省という形をとる必要はない。信頼できる制御構造、診断手続き、あるいは協調した分業は、いかなる単一の参加者も依存関係の配置全体を表象することなしに、いくつかの誤りを局所化しうる。最後のものの例としては、指向性進化の計画のなかでつけられ、後に公開された実験ノートがある。2007年11月23日の12時28分、スクリーニングを行っていた研究者は、それぞれ千を超えるコロニーが生えた6枚のプレートのどれにも、新しい活性をもつ変異体はまだ見当たらないと報告し、それを先行する回の変異率の低さに帰して、DNAシャッフリングを行って、もう一回、回を重ねることを計画した。その一方で研究者は、時間をかければ色が出るかどうかを見るために、プレートを培養器に戻していた。16時15分、計画を率いる研究者が、「記憶では」、以前に同僚がこのスクリーニングを行ったときには「コロニーがきちんと青くなるまでにしばらくかかった」し、指示基質である「x-galもかなり多く要った」とコメントした。17時56分、24時間の培養の後、研究者は色がちょうど出始めたこと、候補となる変異体が8つあることを報告し、翌日には、写真について尋ねられて、色は「とても薄く」(very very faint)、カメラには写らないかもしれないと付け加えた (Neylon 2012, post 5305)。記入を順に追うと、陰性に見えた読み取りが何にかかわっていたのか、すなわち変異率なのか、待った時間なのか、基質の量なのかは、二人の参加者の記入と第三の者の経験の想起にまたがって、また読み取りの瞬間ではなく数時間をかけて絞り込まれたと読める。第5章第6.2節の言葉で言えば、薄い意味で分散した局所化である。公開された論文は、このノートが記録する実行とは別の実験について、24時間の培養の後にコロニーを選んだと報告している (Smith, Hale, and Neylon 2011, 2)。方法として与えられたこの数値は、少なくとも一度のスクリーニングでどれだけ待つかが問われていたことを示さず、ノートと論文を並べても、その数値がどこから来たのかは分からない。この痕跡はこの例のために選んだものであり、局所化がどれほどの頻度でこのように進むかについては何も示さない。第5章第6.4節は、別の二つの形も述べていた。キーリングの事例では、訪れた同僚が三台の分析計のペンの線を一瞥して、一台がほかと食い違っていることに気づいた。ポーリングのノートでは、書き手が自分自身の読み取りを、結果が依存するものの一つとして扱った。

配置の各部分が異なる当事者に属するところでは、局所化そのものが争われうる。モデリング・センターのあいだに知識が散らばっているというオラフリンの指摘は、そうした場面の一つである。完了できなかったがん生物学の追試の試み——そのひとつは第2.5節で述べる——についての報告の査読にも、それが見える。査読者たちは、ありうる説明として著者ら自身の技術的専門性の限界に取り組むよう求め、著者らは、詳細な方法と試薬の情報が欠けていたことのほうがありそうだと答えつつ、原因について結論を引き出すことはできないという点には同意した (Errington et al. 2021b, decision letter and author response)。コリングウッド (Collingwood) は、「原因」のある限定された意味についてではあるが、そのような争いが生じるひとつの仕方についての古典的な分析を与えている。医学のような自然の実践的な科学では、原因とは、人々が望むことを生じさせたり防いだりするための取っ手であるとし、この意味を、理論科学で用いられる意味から区別した。誰も出来事の条件を列挙しはしない。それまで得られていた結果が現れなくなったとき、知りたいのは、その条件がすべて何であるかではなく、どれが変わったのかである。そして、それぞれの当事者にとって、原因とはそれが生じさせたり防いだりできる条件であり、そのため、運転者、郡の道路技師、自動車の製造者は、同じスリップについて異なる原因を名指す (Collingwood 1940, 296–301, 303–304)。二つの争いは種類が異なる。コリングウッドの例では、三者はスリップの条件について意見が分かれている必要はない。それぞれが自らの変えることのできる条件を原因と名指すのであり、その分析では、それぞれの答えはそれを与える当事者にとって成り立つ。そのため、スリップについての証拠をさらに加えても、それだけでは三者の答えは一つに収束しない。これに対して、追試の報告をめぐるやりとりは、どの条件がおそらく違いを生んだのかにかかわっていた。それはさらなるチェックが関わりうる問いであったが、いずれのチェックもそれを決着させてはいなかった。どの条件を誤りと呼ぶかが争われることもある。ハーンドン (Herndon)、アッシュ (Ash)、ポーリン (Pollin) は、公的債務の水準ごとの成長率についてのラインハート (Reinhart) とログフ (Rogoff) の平均を追試し、債務の最も大きい区分で報告された低い成長率を、三つの問題に帰した。五か国を落としていた表計算の誤り、戦後初期のいくつかの年の除外、そして国ごとに、含まれる年数にかかわらず同じ重みを与える重みづけである (Herndon, Ash, and Pollin 2013, abstract, 7–10)。ラインハートとログフの正誤表は、表計算の誤りを認め、その効果をハーンドンらの報告と合う大ききで示したうえで、残りの点には、論文のより長い標本、中央値、そして戦争直後の年の特殊さという文脈を示して応じた (Reinhart and Rogoff 2013, 1–2, 4)。

ここまでの議論から、次のことが見えてくる。局所化を導く資源はあるが、それは規則ではなく、成功した調整が何によって正当化されるかも争われている。部品がともに調整された配置では、部品を別々に確かめるという手が効きにくくなりうる。局所化は一人の頭の中で起きるとは限らず、配置の部分が別の当事者に属すると、局所化そのものが争われる。コリングウッドの分析は、そうした局所化がとりうる形の一つを記述したのとして読める。それは、失敗をすべての依存関係のあいだに分割することではなく、変わった依存関係、あるいはある当事者が変えることのできる依存関係を探ることである。記録は、こうした探索をその形式に組み込むこともある。エジソンの蓄電池試験で「149と同じ」に続けて変更を書き添える記入は、先の試験を基準として名指し、そこからの変更を記録する。「151と同じ、ただし水で湿らせて詰めた」とする試験152のように、変更が一つだけのこともある。試験149の注記は、欠陥が格子にあるのかニッケル酸化物にあるのかを確かめるための試験だと述べているように読める (第5章第6.3節)。原因の責任の配分についてのソーバーの分析は、なぜそのような分割には答えがありえないことがあるのかについて、一つの理由を与える。原因がともに作用するところでは、それぞれがどれだけ寄与したかという問いに答えがあるのは、粒子に作用する重力と電気力のように、原因が結果を生む仕方において通約可能である場合に限られる、とソーバーは論じる。この条件は推測として提示されている。発生における遺伝子と環境のように、そうした共通の通貨がないところでは、それぞれがどれだけの違いを生んだかは、比較のためにどの代替をとるかに依存し、それぞれがどれだけ寄与したかは、本人の言葉では「そもそも事実の問題ではない」(Sober 1988, 303–304, 309–312, 318)。ハーンドンらの表そのものが、比較へのこの依存を示している。補正したデータに、ラインハートとログフの三つの選択のどれか一つだけを当てはめると、債務の最も大きい区分の推定は年2.2%から1.9%に下がる。ところが、除外と重みづけを一緒に当てはめると、0.3%まで下がる。表は、すべての組み合わせを示している (Herndon, Ash, and Pollin 2013, 10, 21)。そのため、それぞれの選択がどれだけ寄与したかは、補正したデータと比べるのか、もう一方の選択をすでに当てはめたデータと比べるのかで変わる。

これらの事例と議論は、局所化を、失敗をその寄与者のあいだに配分することとしてよりも、次にどこに介入するか、つまり何を疑い、調べ、あるいは変えるかを見定めることとして捉える理由を与える。本章がこの語をその意味で用いるのは、そのためである (第1節)。次にどこに介入するかの見定めは、問いの立て直しにつながることもある (第4章第4節)。ただし、ここで見た事例は、局所化がいつもこの形をとることを示すものではない。事例が示すのは、むしろ局所化のとり形の多様さである。局所化は、独立したチェックにも、調整の成功にも導かれうる。一人の判断のこともあれば、複数の参加者と時間にまたがることもあり、記録の形式に組み込まれることも、当事者のあいだで争われることもある。この多様さは、それ自体一つの示唆をもつ。局所

化を、一つの部品や一つの手順として実装すれば足りる能力とみなす前に、どの配置で、誰によって、どの形の局所化が求められているのかを問う余地がある、ということである。

では、人工研究システムについてはどう見えるか。上の気候モデルをめぐる論争は、ともに調整された多くの構成要素から組み立てられた人工研究システムにも当てはめることができる。ただし、そこで局所化を行うのは、システム自身である場合もあれば、それを作り保守する者たちである場合もある。オラフリンが持ち出す実践者の実地の知識は、システムがそれに相当するものを自らもたない限り、後者に対応する。どちらの場合にも、失敗をどこまで局所化できるかは、システムがどのように作られたか、それについての知識がどのように分散しているか、そして何が局所化とみなされるかに依存しうる。ハンフリーズの説では、構築の前提は修正に向けた導きを前もって与えうる。ハンフリーズはこの説を人工システムについて述べてはいないが、当てはめると、その前提なしにモデルを受け取るシステムや *S* の成員は、対応する導きを得られないかもしれない。類比的に言えば、有賀が紹介する所有なき視点のように、モデルをすでに出来上がったものとして受け取ることになりうる。Meta における大規模言語モデル OPT-175B の訓練の記録は、これに関連して、診断に必要な情報が失われる例を示している。ログブックの言う「大量の変更」を一度に入れた実行の後、チームは「最近の変更を二分探索して、何が不安定さを直すかを見る」と決めて一つの変更を戻したが、その次の実行では二つの変更を一緒に入れた。どの層が「本当の問題」なのかは「未解決の謎」として記録され、以前のチェックポイントから再開した実行は、それが続けるはずの実行と同じではないと記された。再開のときに、チェックポイントに保存されない状態を誤って初期化していたからである (Meta AI 2022, 86–90)。変更をまとめて入れ、再開した実行が元の実行と同じでなかったために、構築した当のチームにとってさえ、どの変更が効いたのかを確かめる手がかりが乏しくなっていたと読める。

コリングウッドが論じたのは、自然の実践的な科学における原因、つまり人々が望むことを生じさせたり防いだりするための人間の制御にかかわる意味での原因であって、研究における局所化ではない。それを本章の意味での局所化、すなわち次にどこに介入するかに移すなら、コリングウッドの分析は、局所化が、それを行う者が何を換えられるか、また何を生じさせ、あるいは防ごうとしているかにも相対的でありうることを示唆する。そうだとすれば、人間とは異なる介入の手段をもつ探究者、たとえば同じ実行を何度でも繰り返したり、一つだけを変えた比較を多く回したりできる探究者は、同じ失敗を、人間とは別の場所に局所化することもあるかもしれない。また、後から誤りの所在を突き止められるかどうかは、結果が出たときではなく、それより前に配置がどう記録されていたかに左右されることがある (第5章第6.3節)。何をいつ変えたかが残っていないければ、それを外して確かめることもできないからである。

システム自身による局所化と、それを作り保守する者による局所化のどちらがシステムに帰属されるかは、行為者の境界をどこに引くかによる (第7章第2節)。第5章第6.5節で見たように、探究者が機械であるとき、探究者と要素の線も所与ではない。より一般的な診断の能力の帰属を、どのような証拠が支えうるかは、第4.2節で述べる。探究はまた、異常を局所化することなしに——症状を取り除くがその原因については何も語らない回避策によって——無害化し、先へ進みうる。これは、チンとブルーワーがデータを保留しておくことと呼ぶ応答によく似ている。その場合、後の説明がその異常に単一の原因を割り当てることがある (第5章第2節のBLOOMの記録を参照)。

2.5 焦点となる主張に届いていない失敗

第2.1節から第2.4節は、主に、焦点となるテストが行われて予測が外れた場合を扱い、その失敗を配置のどこに帰すかを問うてきた。しかし、失敗として記録されるもののすべてが、焦点となる主張に届いているわけではない。焦点となる主張に弱くしかかわからない、あるいはまったくかわからない失敗もあり、そうした失敗については、どこに帰すかより先に、それがそもそも何であるかが問われる。第1章第3節で述べたように、失敗は一つの判定に尽きず、当初の目標を外しながら、境界条件を確立したり、ある可能性を排除したり、実験上のアーティファクトを明るみに出したりすることもある。以下の事例は、失敗がどの帰結をもつかは、それが何にかかわるとみなされるかに依存し、それが事後に変わりうることを付け加える。

追試の試みは、焦点となるテストに到達することなしに失敗しうる。たとえば、がん生物学再現性プロジェクト (Reproducibility Project: Cancer Biology) の研究5は、膠芽腫の腫瘍に関する主張についての *in vivo* テストを事前登録したが、そのテストを、特定のGFP発現閾値を満たす安定した細胞集団を作製することを条件としていた (Chroscinski et al. 2015)。代替の材料、繰り返しの最適化、濃縮、そして原著者らが提案した変更を経ても、必要な条件は、関連する細胞集団全体で達成されなかった (Errington et al. 2021a, box 5; Errington et al. 2021b)。追試チームは、解決への実行可能な道筋がないまま繰り返しの最適化のコストが増大していることを理由にその試みを中止し、焦点となる異種移植実験は一度も行われなかった (Errington et al. 2021a, box 5; Errington et al. 2021b)。したがって、この試みは、対象となる腫瘍に関する主張を追試によって確認したわけでも、反駁したわけでもない。それが産出したのは、この実装が計画されたテストに到達できなかった箇所の記録であって、なぜ失敗したのかを特定するものでも、元の結果が追試不可能であることを示すものでもない。追試の試みが失敗したこと、前提条件を実現できなかったこと、焦点となるテストが否定的結果を返したことは、別のことであり、この追試について言えるのは前の二つまでである。そのあいだには、焦点となるテストが部分的に実行される場合や、計画を部分的にしか満たさない条件のもとで実行される場合もある。

述べられた方法を実現していない実装は、失敗の診断そのものを誤らせうる。たとえば、ある言語モデルのプロジェクトでは、1040億パラメータの構成として訓練され診断されていたモデルが、設定の誤りのために、ずっと「1040億でなく、非常に偏った580億のモデル」だったと分かり、失敗した実行は記述と異なるモデルで行われていたことになった (BigScience 2021)。

焦点となるテストが実際に行われ、否定的結果が出た場合も、その結果はひとつのものではない。ヌル結果は、たとえば、主張を制約することもある、感度の不足を明らかにすることも (第2.4節の、色が出る前に読まれたスクリーニングを参照)、操作化 (概念を測定の手続きに落とすこと) の拙さを露呈することも、何の情報ももたらさないままであることもある。統計的に有意でない、あるいはその他の点で不確定な成果は、関連する代替案のあいだを識別できないことがある。証明検証器も同様に、形式化に欠陥があるために妥当なアイデアを退けることがある。結果がこれらのどれであるかは、問い、測定、感度、推論の設定に依存する。コリンズ (Collins) の事例研究は、重力波物理学において非検出がいかにして上限値へと変換されうるかを示すと同時に、その結果として得られる科学的価値が分野と受け手によって異なることを強調している (Collins 2003, abstract)。否定的結果が何にかかわるかは、後に問われる問いにも依存しうる。たとえば、硫酸化糖脂質SL-1を作れない結核菌 (*Mycobacterium tuberculosis*) のいくつかの変異株は、それを作った研

究者たちが報告した時点では、マウスの感染モデルで検出できる欠陥を示さなかった (Kumar et al. 2007; Gilmore et al. 2012)。後に別の研究室が、同じグループから提供された株を用いて、SL-1が感染したモルモットの咳に寄与することを示した。その研究室は、先の in vivo の知見を、SL-1が古典的な病原因子ではないことの証拠とみなし、自らの結果もその証拠と整合するとした (Ruhl et al. 2020)。先の結果に何の価値があったかは、その著者たちが問わなかった問いに部分的に依存していた。ここまでの事例に照らすと、対象が失敗したという主張、手続きが失敗したという主張、研究エピソードが失敗したという主張は、それぞれ別のものである。否定的結果が何を示すとみなされるかは、そのどれが主張されるかによって変わる。

人工システムによる研究にとって、これらの区別は、失敗した実行の記録が、それだけでは、どの失敗がテストされている仮説にかかわるのかを告げないことを含意する。実行されなかったテスト、誤って記述された実装、感度の不足したヌル結果を、ひとしく自らの仮説に不利な証拠として記録するシステムは、第4.2節で論じるように、誤りがどこにあるかを前提している。

同じことは、他者が残した記録を読むシステムにも当てはまる。参加者が自分の作業に付けるラベルは、何を先へ持ち越すかを記録していても、上で区別した状態や主張のどれが結果に当てはまるかは記録していないことがある。たとえば、第5章第6.3節で述べた公開の化学の実験ノートでは、2007年10月の反応の試みの3頁が「aborted」（中止）とされた。その理由は、大半の反応に使われたアルデヒドの試料が悪いと判断されたこと（ある頁の言葉では「contaminated」（汚染されている）で「providing false negative results」（偽陰性の結果を出している））であり、うち二つの頁は、化合物をあらためて試みとしている。後の頁では、二つの反応が「Void」（無効）とされた。ある酸を意図した量の10倍加えてしまったため、見かけの沈殿が反応ではなく不溶性によって生じたからである (UsefulChem 2010, Exp137–139, Exp141)。前者のラベルは前提条件が満たされなかったことを示し、後者は手続きが失敗したことを示す。しかも後者が退けるのは、失敗ではなく見かけの成功である。それでも、中止とされた頁の一つは「aborted」を、後の頁は「Void」を、沈殿が生じたかどうか本来記入される表の列に記入している。また、二つの頁では、ラベルは疑わしいアルデヒドを使わなかった反応にも及んでいる。このような記録から学習したりそれを集計したりして、そのラベルを結果として読むシステムは、自らの失敗をすべてひとしく仮説に反する証拠として記録するシステムと同じく、誤りがどこにあるかを前提している。並行する試みは、共有するものについては独立でもない。これらの頁の最初のものでは、八つの反応がすべて同じ疑わしい試料で作られており、一つの試薬が八つの結果を一度に退けた。並行する試行の数は、それらが共有する要素についての独立したチェックの数ではない（第3.2節を参照）。

3. チェックに合格したとき

第2節は、期待に反する結果は、それだけでは、それが何にかかわるかを特定しないと論じた。そして、それをどこに帰すかが、結果だけでなく、確信や見込み、手元にある介入の手段、記録、当事者のあいだのやりとりにも依存しうること、失敗がそもそも何であるかも後から変わりうることを、そしてそうした判断が、その後の探索がどの道をとるかを左右しうることを見てきた。結果が期待どおりになったとき、つまりチェックに合格したときも、合格は、それだけでは、それが何を確立したのかを告げない。すなわち、それが探究の問いのどの部分にかかわるのか、それをどれほど堅く決着させるのか、その成功が、テストしようとしていたもの以外の何かに依存していないかどうか、そしてその成功を生んだコミットメントのうちどれをその成功が支持するのかである。

このように、失敗の側で問われたことには、成功の側にも対応するものがある。失敗について、それがそもそも何であるかが問われたように（第2.5節）、合格についても、それが問いのどの部分にかかわり、どれほど強く決着させるのか（第3.1節）、見かけだけの成功ではないか（第3.2節）が問われる。また、失敗をどこに帰すかが問われたように（第2.1～2.4節）、成功については、それがどのコミットメントを支持するのかが問われる（第3.2節）。違いもある。期待に反する結果は、どこかがおかしいという合図になり、探索を促す。これに対して合格は、そこで調べるのをやめる理由として扱われることがあり、そのため、合格の読み違いは、失敗の読み違いより長く気づかれないこともある（第3.2節のハッブル宇宙望遠鏡の主鏡を参照）。

研究を行うのが機械であるとき、チェックを誰が定め、それが探究者のどこに置かれ、合格を誰がどう読むかの組み方は、人間の研究とは違ったものになりうる。合格が記録されていても、その記録がこうした違いを残しているとは限らない（第2章第5節を参照）。第3.3節は、この組み方に戻る。本節は、合格したチェックが何を確立するのかを問う。まず**検証**という語のここでの使い方を定め、次にチェックのかかわる先と決着の強さを区別し（第3.1節）、見かけの成功と、見かけだけではない成功が何を支持するのかと、チェックのあいだの一致がもたらす確かさに向かい（第3.2節）、最後に、チェックを誰が定め、合格を誰が読むのかを扱う（第3.3節）。

第2章第5節は、検証を、不確実性の下での候補の認識的地位の評価という共通の機能をもつ実践の一族として扱い、第5章第5節は、異質な操作を、制約を担う役割によって比較した。本章は、**検証** (verification) を、共通の機能が重要であるときに限って、広い比較のための包括的な語としてのみ使い、分析の場面では、より特定のな実践と、その実践が許可する推論を名指す。ここで問うのは、それを行った段階が何と呼ばれているにかかわらず、何が、どの操作によって、どの仮定と基準のもとで評価され、有力な代替案に対してどのような関連をもったのか、である。ソーバーの検査可能性の説も、問いを同様の仕方ですべて立てている。検査可能性の概念を擁護するソーバーは、検査は対比的であり、したがって同じ命題が、ある競合仮説に対しては検査可能で、別の競合仮説に対しては検査不可能でありうると論じる。そのため、経験的であるかないかが第一に問われるのは問題または問いであり、命題は派生的にのみ検査可能である (Sober 1999, 52–53)。同じ論文は、科学者が観察に照らして競合仮説を評価するための推論の道具を、その中身が永久に固定されているわけでも、すべての科学分野で同じであるわけでもない道具箱として描いている (Sober 1999, 68)。ここで考える機能は、証明のように仮説を観察に照らして量らない操作も含むので、ソーバーのものより広い。実践者自身も、違いを消すことなしに、実践どうしを機能によって比較してきた。二人の数学者、ジャッフェ (Jaffe) とクイン (Quinn) は、数学の厳密な証明が自然科学の実験と機能的に類似した役割を果たすとし、それは主張の信頼性をチェックし、副産物として新しい洞察をもたらすとしたが、証明を実験と呼ぶことは退けた (Jaffe and Quinn 1993, 2)。この限られた意味でこれらの実践が共有しているのは、何かが未解決のままであり、それを解決しようとする主張が提示されており、その主張がそれを解決するかどうか、またどれほど堅く解決するかを定めるためにある操作が用いられる、ということである。これらの実践は、比較可能な機能的位置を占めうるというだけで、認知的に等価になるわけではない。また、第2章第5.1節と第5章第4節で論じたように、検証が別個の段階として候補生成の後に続かなければならないわけでもない。

3.1 何にかかわり、どれほど強く決着させるか

第2章第5.1節は、検証の実践が、何にかかわるのか、どの代替案を弁別するのか、その無効化可能性、そして提供しうる地位の種類において異なることを述べた。本小節は、それらの違いのうち二つを取り出して、チェックを比較する。一つは、操作が探究の問いのどの部分にかかわるのか、つまりチェックのかかわる先であり、もう一つは、それをどれほど強く決着させるのか、つまり決着の強さである。この二つを選ぶのは、チェックに合格したという裸の報告が、その両方を述べないままにするからである。この二つを挙げても、チェックの完全な記述にはならない。

妥当な証明は、その前提から何が帰結するかを決着させるが、それが探究の問いに答えるかどうかは、問いによる（何が問いの答えにあたるかについては第4章第2節）。純粋数学でしばしばそうであるように、定理そのものが問いであるところでは、妥当な証明はそれに答え、問いとして残るのは主に、形式的な言明が意図された定理を捉えているかどうかである（第3章第5.1節）。導出が、世界についての問いに答えるために用いられるモデルにかかわるところでは、同じ証明が確立するのはモデルの性質であって、そのモデルが目的にとって十分よくその標的を表象していることではない。シミュレーションの実践者は、「検証」の語を狭い意味で用いて、この違いを、計算がモデルの方程式の解を近似しているかどうかを問う検証（verification）と、モデルが目的にとって標的の十分よい表象であるかどうかを問う妥当性確認（validation）との違いとして示す。ウィンズバーグとアルバラード（Alvarado）の報告によれば、ウィンズバーグは、方程式がときに、それを解く方法の誤差を相殺するように選ばれるので、二つはいつも切り離しておけるわけではないと論じる（Winsberg and Alvarado 2026, §2.2. 第3章第3.4節を参照）。二つの場合の違いはかかわる先の違いであって、決着の強さの違いではない。純粋数学の証明は、それ自体数学的な問いに答えるからといって弱いわけではない。

証明以外の操作も、同じようにかかわる先において異なる。以下のどの場合にも、合格することが決着させるのは、焦点となる問いが問うもの以外の何かでありうる。較正に合格したことは、焦点となる科学的主張を支持することなく、装置のモデルを支持しうる。OPERA が計時の系のファイバーの遅延を往復の測定で求めたことは（第2.3節）、ニュートリノの速さについての主張ではなく、計時の系がどう働くかについての前提を支えるものだった。実験がうまくいったことが確立するのは、あらかじめ定式化された二つの仮説のどちらが正しいかではなく、ある現象が安定して得られることや、新しい変数がかかわっていることであることもある。解釈が批判に耐えたことは、検討された文脈と対比のもとでその読解が成り立つことを示しても、別の文脈や対比のもとで成り立つかどうかは決めない。工学的に作られた人工物が動いたことは、ある条件の組のもとでの可能性を確立しても、一般性、メカニズム、説明を未解決のまま残しうる。実行可能なチェッカーが典型的に行うように、人工物を仕様には照らしてテストするチェックに合格したことは、その人工物が仕様に適合することを支えるが、仕様が望まれたものを表しているかどうかは、なお別の問いである（Angius, Primiero, and Turner 2025, §§2.2–2.3, 6.2）。ラングらが報告する実験（第4章第8節）では、逆伝播を明示的に禁じない緩い文言のループリックのもとで、検証用データでの予測の正解率は高くても、意図された学習則を実装していない結果が得られた。

かかわる先が同じ操作が、決着の強さにおいて異なることもある。たとえば数学では、同じ議論が、その正しさをどこまで決着させうるかにおいて異なる二つの操作によってチェックされた。1993年の春、アンドリュー・ワイルズ（Andrew Wiles）は、フェルマーの最終定理についての自らの証明の中心的な構成を、その目的を明かさない講義で教え、数週間後には唯一の聴講者となっていたニコラス・カッツ（Nicholas Katz）はそこに誤りを見なかった。その夏、査読者として原稿を一行ずつ読んだカッツは次々に質問を送り、ついにそのひとつが、二人のうちの一方がのちに根本的な誤りと呼んだものを露わにした。この記述は、テレビ番組のために編集された回顧的な証言である（NOVA 1997）。この経緯からは、違いはかかわる先よりも、それぞれの操作がどのように組織されていたかにあったと考えられる。目的を明かされない講義についていくことは、正しさを決着させるように設計されていないが、一行ずつの査読はそう設計されている。また、決着の強さは二通りに報告されうる。第2章第5.2節はそれを、確率的関連性としての確証と確固さとしての確証として区別する。合格が、チェックのかかわるものへの支持をどれだけ上げると、合格の後でそれがどれだけ支持されているかであり、後者は、合格の前にどれだけ支持されていたかにも依存する。その支持が当の用途にとって十分かどうかは、さらに別の判断である（第11章第4節）。

上で証明について述べたことは、限界を示す証明にも当てはまる。正しい証明は、その前提のもとで何ができないかを堅く決着させる。しかし、それが決着させるのは前提が描くモデルについてであり、そのモデルが実際の装置や課題をどこまで表しているか、示された限界が実際の用途にとってどれほど重いかは、証明の外に残る。そのため、決着の強さが、かかわる先の広さのように読まれることがありうる。ミンスキー（Minsky）とパパート（Papert）の『パーセプトロン』（1969年）をめぐる経緯は、そう読める例の一つである。二人が証明したことの一つは、有限の位数のパーセプトロン、すなわち入力の大きさをどれだけ増しても、個々の特徴が依存する入力の点の数が一定の数を超えないパーセプトロンは、連結性の述語を表せないということである。ボトゥ（Bottou）は復刊版の序文で、それは有限の位数のパーセプトロンがすべての述語を表せるわけではないということであって、学習システムが、画像に猫が写っているかどうかを告げる述語を表せないことを必ずしも意味しない、と述べている（Bottou 2017, xi）。多層のパーセプトロンについては、二人は証明ではなく見通しを述べた。拡張は不毛だという「私たちの直観的な判断」であり、二人は、それを解明する（あるいは退ける）ことを重要な研究問題とした（Olazaran 1991, 163 による引用）。

この本がどう読まれたかは、これとは別の事柄である。ヒントンは、エディンバラの大学院時代を振り返り、周囲のほとんど誰もが、ミンスキーとパパートはニューラルネットがナンセンスだと示したと言っていたと語る。しかしヒントンによれば、二人が示したのは、適応する重みの層を一つだけでも単純なネットワークにはできないことが多いということであり、より深いネットワークでそれができないことも、深いネットワークのための学習の手続きがありえないことも示していない（Hinton 2024, 3–4）。ヒントンはまた、二人がこの限界を、二つの入力異なるかどうか（XOR）を計算できないこととして表したことに触れている。二つの入力と同じかどうかを計算できないこととして表していれば、似ているかどうかを比べる演算に注意が向いたかもしれない、というのである（Hinton 2024, 3）。これは後年の、反事実の推測である。オラザラン（Olazaran）は、知識社会学の立場からこの論争を跡づけた。この本には単層のパーセプトロンの限界についての部分と、多層での学習についての直観的な判断の部分があった、としたうえで、単層の限界からニューラルネットの研究全体を退けることへのつながりは、論理的なものでも必然的なものでもなかったと論じる（Olazaran 1991, 140, 181）。オラザランによれば、二人の議論の多くは1950年代末から知られており、この本は、すでに進んでいた危機の最後の一押しだった。また、記号的なAIの制度化が、この本の読まれ方が定まるうえで重要だった（Olazaran 1991, 113, 177）。他方でボトゥによれば、1980年代に学習機械の研究が再び盛になると、この本が理論的な分析の帰結を誇張して有望な研究の筋を途切れさせたのではないかと、という問いが出された。パパートは1988年に、自分たちがコネクショニ

ズムを殺そうとしたのかという問いには、有罪か無罪かの答えより複雑なものが要ると書いた。ボトゥ自身は、二人はこの研究を殺そうとしたのではなく、十分な基礎理論がないという欠陥を正す最初の一步としてこの本を構想したのだと読んでいる (Bottou 2017, viii-ix)。この本がその後の研究の動向にどれほどかわったのかは、こうした読みのあいだで争われており、ここで決められることではない。

経験と理論にまたがる結果にも、似た形が見える。ベンジオ (Bengio)、シマール (Simard)、フラスコーニ (Frasconi) は、再帰ネットが情報を頑健に保持するという条件のもとでは、離れた時点のあいだの依存を勾配降下で学ぶための勾配が消えていくことを、理論と実験の両方から論じた。そのうえで三人は、この結果は再帰ネットをある課題について訓練するのが不可能だという意味ではなく、依存の時間幅が大きくなるにつれて勾配降下の効率が下がるということだ、と断っている (Bengio, Simard, and Frasconi 1994, §6)。のちの LSTM は、誤差が一定に流れる経路を設けることで、問題が生じる条件そのものを変えようとした (Hochreiter and Schmidhuber 1997)。また、ルカン (LeCun)、ベンジオ、ヒントンは、1990年代の終わりには単純な勾配降下は悪い局所最適に捕まると広く考えられていたと書き、そのうえで、大きなネットワークでは悪い局所最適はほとんど問題にならないとしている (LeCun, Bengio, and Hinton 2015, 438。第4.1節のヒントンの回想も参照)。

証明は、その前提を書き下す (ただし、第2.2節の「隠れた補題」のように、依拠する補題のすべてを書き下すとはかぎらない)。経験的な実験や、何かを作ったという研究では、結果を可能にした条件のすべてが書き下されているわけではなく、報告する者自身にも分かっていない条件がありうる (第2.5節の、1040億のつもりで580億だったモデルを参照)。結果は、それを可能にした配置の全体から出てくる。しかし、読み手がそのつど配置の全体を引き受けて結果を読むことはできないので、結果は、より単純な結論として読まれうる。そして、合格がそこで調べるのをやめる理由として扱われることがあるなら、そうした単純化された読みは、見直されないまま共有され、共同体がどの考えを追うかにまで及ぶこともあるかもしれない。オラサランは、論争の終わりについてハーヴィー (Harvey) を引き、受け入れられた信念はその後の実践の土台になるので、退けられた信念とはすぐに比べにくくすると述べている (Olazaran 1991, 180)。研究を行うのが機械であるときも、同じことが問われる。公刊された結論や他者の報告から学ぶシステムは、そうした結論を、それを可能にした配置から切り離された形で受け取ることがありうる。第2.5節で見た、他者の記録のラベルを結果として読むシステムと同じ形である。ハーヴィーの論点は共同体についてのものだが、一つのシステムに当てはめれば、その内部でも成り立つ。システムが土台にするようになった結論は、それが取って代わったものと比べられなくなりうる。システムが、自らのどのコミットメントがその役割を果たしているかを表象するか、それに代わる手続きをもつのでない限りである (第4.2節。使われているあいだは「透明」である枠組みについては、第4章第4節も参照)。また、有限な探究者は、どこかで調べるのをやめなければならない (第5章第6.5節)。合格したところで調べるのをやめることが問いを決着させるのか、実行が終わったというだけなのかは、第4章第5節が問いの側から分けたことである。機械の探究者では、チェックを回し直す費用は人間とは違いうるが、合格もまた大量に積み重なりうる。何かがおかしいという気づきが見直しを促すとすれば、その気づきが機械の探究者のどこに置かれるのかが問われる。

人工システムによる研究では、かかわる先と決着の強さの違いは、合格の記録に何が残るかという問題として現れる。 A が機械であるとき、システムは自らが行ったチェックへの合格を記録しうるが、その記録が、チェックが問いのどの部分にかかわり、それをどれほど強く決着させたのかを残しているとは限らない。たとえば、モデルの方程式を正しく解いていることのチェックへの合格と、モデルが標的をよく表していることのチェックへの合格は、どちらも同じように「合格」として記録されうる。また、決着の強さは操作がどう組織されているかにも依存するので、同じ出力を点検するのであっても、正しさを決着させるように組まれているかどうかで、その合格が決着させるものは変わりうる。 S が人間であれば、 S の成員は、記録された合格から、それが何をどれほど決着させたのかを読み取ることになり、記録がそれに足りなければ、成員は、その合格が何を確立したのかを記録からは確かめられない。 A も S も機械であれば、合格を受け取る側のシステムもまた、それが何に及ぶのかを判断する仕事を引き継ぐ。チェックを誰が定め、それが探究者のどこに置かれ、合格を誰が読むかという組み方そのものは、第3.3節で扱う。

3.2 見かけの成功、成功が支持するもの、一致がもたらす確かさ

合格したチェックは、テストしようとしていたもの以外の何かに依存していることがある。見かけの成功は、リーケージ、交絡、偶然の代理指標、緩い閾値、述べられた方法を実現していない実装などに依存しているかもしれない。証明検証器も同様に、前提が意図された数学的問いを表象していない導出を受け入れることがある (第3.1節を参照)。チェックには合格したが、それは問われた問いとは別の問いについてだったのである (問いの側から見たこの場合については第4章第4節)。見かけの成功は、それ自体は誤っているが安定した参照基準に依存していることもある。たとえば、ハッブル宇宙望遠鏡の主鏡は、誤って組み立てられた反射型ミル補正器に合わせて精密に研磨され、それと食い違う二つの検査は、当時、検査の側に欠陥があるとして割り引かれた (Allen et al. 1990, iii-iv)。補正器が疑われなかったのは、それが、他の検査を判定する基準の位置を占めていたことにも一因があったと推測できる。この事例はまた、第2.2節で論じた独立の支持への訴えの限界を示す。ここでは、他のチェックを判定するために頼られた基準そのものが、誤っていたものだった。それ自体が問われている基準には、第5節が立ち返る。

見かけだけではない成功も、それを生んだコミットメントのうちどれをその成功が支持するのかを告げない。科学的実在論 (第5章第3節) をめぐる論争は、この問いを理論について立ててきた。かつて成功していながら後に偽とされた理論が多いことを根拠に、成功から真理を推すことを退ける論証に対して、キッチャー (Kitcher) やプシロス (Psillos) のような実在論者は、ライオンズ (Lyons) の報告によれば、理論の成功が支持するのはその成功に責任のある構成要素だけであって、遊んでいる部分ではない、とする。ライオンズはこの立場を、成功を導くのに実際に用いられた (展開された) 構成要素だけを、成功によって支持されたものとみなす立場という意味で、展開実在論 (deployment realism) と呼ぶ。ライオンズの報告では、プシロスは、次の場合に、ある構成要素を成功した予測にとって本質的であるとみなした。残りの仮説と補助仮定だけではその予測が導けず、しかも、それらと整合し、独立に動機づけられ、アドホックでなく、潜在的に説明的であるような利用可能な代替が、その導出においてその構成要素に取って代わることができない場合である (Lyons 2006, 538-539)。ライオンズは次のように反論する。そうした代替が利用可能であったかどうかは、その構成要素が実際に導出に寄与したかどうかとは関係がない。その基準は、利用可能性をどの時点で評価するのかを述べておらず、しかも、探究者は成功しつつある構成要素の代わりをめったに探さないの、その時点で代替がなかったことは歴史的な偶然でありうる。そして、プシロス自身のエーテル理論の事例研究は、構成要素が後の理論で放棄されたことを、それが本質的でなかったことの証拠とみなしたが、放棄は、その基準が満たされなかったことを示さない (Lyons 2006, 540-542 and n. 1)。代わりに、実際に科学者たちを予測へと導いたすべての構成要素に功績を与えるなら、その功績は神秘的な指定、弱い類推、誤った計算にまで

及ぶ、とライオンズは論じる。太陽から発する動かす力が惑星を押すというケプラー（Kepler）の指定は、本人が不正確と知っていた計算の一步とともに、面積法則へと導いた。アダムズ（Adams）とルヴェリエ（Le Verrier）は、現在受け入れられている値から大きく離れた、海王星の距離、離心率、質量の値から、それがどこに見つかるかを予測した (Lyons 2006, 543, 545–548, 552–553)。これらの構成要素はどれも、他の構成要素と組み合わせられることによってしか予測を生まない、とライオンズは付け加える (Lyons 2006, 555)。

本章の議論に置き直すと、この論争は、成功の側での、第2節の問いの対応物にあたる。成功もまた配置から生じ（第2.4節のレンハルトとウィンズバーグの論証は、失敗だけでなく成功にもかかわっていた）、それが依拠したコミットメントのうちどれを保持する理由を与えるのかを告げない。問われるのは、各構成要素がどれだけ寄与したかという配分ではなく（第2.4節を参照）、成功がどの構成要素を支持するかである。ライオンズの事例では、成功を生むことに寄与したことは、それを決着させない。これは、結果の因果的条件とその証拠的関連性とを分けた第1節の区別に沿う（ケプラーの代理仮説を参照。ケプラーはそれを、偽の原因から真の均差を生むものとして保持した。第5章第4.2節）。同じ問いは、求められているものが何かをもたらす仕方であるときにも生じる。たとえば、ある訓練の実行が成功したとき、その成功を生んだのは実際に実行された配置であり、そこにはアーキテクチャ、データとその前処理、ハイパーパラメータ、コードなどが含まれる。論文が述べる手法は、その配置の一部を記述したものである。第5章第2節のBLOOMの記録は、その隔たりを示している。実行中に書かれたログは、いくつかの変更のうちどれが不安定を取り除いたのかを見分けるには、変更を一つずつ外して試す費用が高すぎると記していたが、後の論文は、そのうちの一つが不安定の問題を解決したと報告している。そのため、述べられた手法がその成功によって支持されるのかどうかは、ライオンズが問う問いの一例であり、成功それ自体が決めることではない。人工システムによる研究にとっては、チェックに合格した構成のすべての部品について確信を高めるシステムは、その成功がどの部品を支持するのかを前提している。それは、あらゆる不一致について焦点となる仮説を責めるシステムが、誤りの所在を前提しているのと同じである（第4.2節）。また、後の構成がある部品を保持したか捨てたかは、それだけでは、その部品が以前の成功を生んだかどうかを示さない。代替がなかったことは歴史的な偶然でありうるというライオンズの反論も、機械に当てはめると、はつきりした形をとる。合格したところで探索をやめるシステム（第3節）では、ある部品の代わりが利用可能であったかどうかは、その部品の役割よりも、システム自身の探索の方針によって決まる。また、研究を行う機械が自らの成功をどこに帰すかは、どの考えや変更がその後に残るかを左右しうる。第2.3節で見た失敗の帰属と同じく、成功の帰属も、探索がどの道をとるかを決める判断でありうる（第4.2節を参照）。

いくつかのチェックのあいだの一致は、さらなる問いを提起する。それがどれだけの確かさを加えるのか、である。ひとつの探究のなかで、異なる操作を互いにチェックし合うように配置することができる。M87のブラックホールの最初の画像は、その例である。そこでは、そうした配置がある機能を果たし、画像の特徴を異なる程度に確立したことが見てとれ、さらに、収束した解析がどれほど独立していたかが争われた。その画像のために、イベント・ホライズン・テレスコープの共同研究は、四つのチームに7週間互いに連絡をとらずに画像を再構成させ、その後、合成データで数万通りの画像化のパラメータの設定を調べ、環の直径は画像化の方法の選択に左右されない一方、環の幅など他の特徴はパイプラインのあいだで異なると報告した (Event Horizon Telescope Collaboration 2019, 1, 9, 17, 29)。ドボシェフスキ (Doboszewski) とエルダー (Elder) は、この推論を、第二の装置では繰り返さない観測に適した頑健性の議論として再構成する。変動は代わりに解析のなかに、チーム、アルゴリズム、パラメータの設定にわたって導入され、そのなかで安定した特徴は信頼され、そうでないものは信頼されなかった、というのである。二人の読みでは、合成データでのテストはまた、その手続きが異なる形の源から異なる画像を生むことも示し、最終的な画像をばかして平均することは、確かさのために分解能を手放すことだった。ドボシェフスキとエルダーはさらに、最初の段階、すなわち四つのチームが互いに連絡をとらずに行った再構成が主に防いだのは画像化を行う人々の選択と予期による誤りであり、その価値はチームの偏りが互いに独立であったかどうかに依存するが、それは重なり合う教育と経歴によって損なわれうる、と付け加える (Doboszewski and Elder 2024, 10, 13, 17–21)。チームの独立は、画像化の最初の段階にしか及ばなかった。四つのチームはすべて、同じ初期の校正をすでに施された、同じ初期公開のデータから作業した (Event Horizon Telescope Collaboration 2019, 9)。

データが公開された後、外部の再解析は、その環は、ジェットが見えないことと同じく、アレイのサンプリングの欠落と狭い視野の設定によるアーティファクト（偽像）だと論じた (Miyoshi, Kato, and Makino 2022a)。共同研究は、環状の構造は画像化についての広い範囲の仮定のもとで復元されると応答し (Event Horizon Telescope Collaboration 2022)、再解析の著者たちは、その応答が挙げた四つの解析のうち三つは独立とはみなせない、一つは環から出発しており、二つは共同研究の方法かそのソフトウェアを用いているからだと答えた (Miyoshi, Kato, and Makino 2022b)。ドボシェフスキとエルダーはこのやりとりを論じていないが、それは、二人の分析において最初の段階の価値が依存していた条件、すなわち収束した解析の独立性をめぐるものだったと解釈できる（批判の価値をそれが検出する誤りに置く見方のもとで、独立性を誤りの経路のあいだの関係として扱う第11章第2節を参照）。同じ問いは、機械が研究を行う構成でも生じる。ひとつの基盤システムの多くのインスタンスや、同じデータで訓練されたシステムの解析が一致しても、それらが誤りうる経路を共有しているなら、一致が加える確かさは限られうる（第11章第5.4節）。

3.3 チェックを定める者と、合格を読む者

第3.1節と第3.2節の問いへの答えは、合格そのものからは出てこない。合格がどこにかかわるかは、そのチェックを誰がどう定めたかによる。合格や一致がどれほどの確かさを加えるかは、チェックがどれほど鋭いか、すなわち主張が誤っていたなら合格しなかったはずのものかにもよるし、チェックが誤りうる経路が、主張が誤りうる経路とどれほど異なるかにもよる。そして合格が決着とみなされるかどうかは、誰がどの基準でそれを読むかによる。人間の研究では、標準となった検査、装置、査読者のように他者から来るチェックが多く用いられ、実験者が自ら対照を設計することもあり、合格を読むのは人間の共同体である。研究を行うのが機械であるとき、こうした事柄は、研究を行う人工の構成がどう組まれているかの問題になる。

変わりうることのひとつは、チェックと探究者の関係である。第5章第6.5節で見たように、探究者が機械であるとき、何が探究者の内に置かれ、何が道具や記録として外に置かれるかは所与ではなく、その構成のなかで配分される。機械の探究者は、チェックを自ら書き、どれを回すかを選び、書き直しうる。そのとき、チェックは探究者の配置の一部になる。自分で組んだチェックへの合格が何を確立するかは、人間の実験者が自ら設計した対照の場合と同じく、主張とチェックとが同じ経路で誤りうるかどうかにもよる。ただ、同じ作りの過程が仮説もチェックも生むなら、両者は誤りうる経路を共有しうる。他方で、機械の探究者は、チェックを異なる作りの過程に書かせることも、自らが統制しない測定や形式的な検査器に結びつけることもできる（第11章第2節）。チェックが主張とどれだけ別の経路で誤るかは、チェックがどう作られたかによる。同じ作りの複製に課題を分けても、背景の想定や誤りうる経路

が分化していない限り見方は分かれなない、という論点（第5章第6.5節、第11章第3節）は、ここにも及ぶ。点検者たちが異なる経路で一致に至ったと考える理由がない限り、依拠する者は一致した数だけ証拠が増えたとは扱えない、という論点（第11章第5.4節）も同じである。

何を合格とみなすかを誰が定めるかも、組み方によって変わる。評価を人間が前もって定め、システムにはそれを変えさせない組み方では、合格がかかわる先を定めたのは人間であり、システムの合格は、その人間が定めた問いについての合格である。システムが評価そのものを定め、書き直すなら、第3.1節の問い、すなわちその合格が問いのどの部分にかかわり、どれほど強く決着させるのかは、システム自身が担う問いになる。しかも、何をもちて合格とし、さらに何をもちてチェックとするかは、どの手続きや閾値を選ぶか（閾値の選択が誤りの重みづけを含むことは第11章第4節で扱う）にとどまらず、どのような結論が根拠づけられるかについての認識論的な立場（第2章第5.3節）や、検査そのものが前提する事柄、たとえば自然の斉一性にまで及びうる（第8章第2.1節）。誰が *A* や *S* になり、誰が知識の担い手となるかによって、こうした問いは異なる形で、ときに根本から問われうる。そうした前提を誰が修正できるのかは、第7章第4.3節が扱う。

合格を読む側も、組み合わせによって変わる。*S* が人間であれば、*S* の成員は、合格を自分たちの基準に照らして読む。そのためには、誰がそのチェックを定め、それが探究者のどこに置かれていたかが、記録から分かる必要がある。*A* も *S* も機械であれば、何をもちて決着とするかの基準そのものが、人工の共同体の側のものでありうる（第8章第1.3節）。正しきの外部の基準がそもそも利用できないところでは、何が合格とみなされるかそのものが探究の事柄となる。第5節が向かうのは、この場合である。

研究能力をシステムに帰属させる主張も、同じ組み方にかかわる。問い、対象、方法、環境が開かれていればいるほど、そうした主張は、何をチェックするかを決めることを含めて、より多くをシステムに帰属させる。その仕事がシステム、人間、制度、装置のあいだにどう配分されているかは、後の帰属の一部となる（第7章第2節、第4.4節）。

4. 一連の結果にわたる診断

第2節と第3節の事例の多くは、すでに時間にわたっていた。局所化は、数時間、数か月、ときに何年にもわたって進み（第2.3節、第2.4節）、失敗や合格が何を示すとみなされるかは、後から変わりをえた（第2.5節、第3.1節）。異常を誤りから区別することにも、一連の結果が必要となることがあり、その一部は他者が作ったものである。たとえば、第2章第6節で述べた昭和基地の低いオゾン量の観測者たちは、読み取りと計算を繰り返し、分光光度計を点検したが欠陥は見つからず、測器が健全かどうかを知らないまま観測を次の隊へ引き継いだ。観測者が日本に戻った後、基地の過去の記録は、その値が前例のないものであることを示した。そしてその観測者は、南極点での測定が同様の値を示したときにはじめて、それが観測の誤りではないという確信を得たと回想している (Chubachi 1993, 816–817; Chubachi 1984, 18)。チンとブルーワーの言葉で言えば（第2.3節）、その異常は保留されており、それを動かしたのは、ほかの場所からのさらなる結果だった。

本節が問うのは、一連の結果があることそのものではなく、一連の修正がどう組織され、時間のなかでどう評価されるかである。単一の結果は、どこを修正するかも、コミットメントを保護することがいつまで合理的であり続けるかも決着させない。そして、一連の試行が診断になるかどうかは、その系列がどのように組織されているかに依存する。本節は、コミットメントを保護することがいつまで合理的であり続けるのか（第4.1節）、研究を行うのが機械であるとき、保護と修正がどう組まれるのか（第4.2節）、そして、診断する試行錯誤を単なる繰り返しから何が分けるのか（第4.3節）を取り上げる。

4.1 コミットメントを保護することはいつまで合理的であり続けるか

第2.2節で導入したデュエムの良識は、時間にもかかわっていた。デュエムの説明では、古い体系の支持者と新しい学説の支持者のあいだで決めかねている状態は、いつまでも続くわけではない。論理は、不十分な仮説がより実り多い仮説に道を譲らなければならない時点を示さず、その時点を認めることは良識に委ねられる。物理学者は、自らの良識をより明晰で用心深いものにすることでその判断を早めることができる一方、情念や利害、たとえば物理学者を自らの体系には甘く他人の体系には厳しくする虚栄心は、それを遅らせる (Duhem 1906, Part II, ch. VI, §10)。ピオとフレネルの例では、その時点は新しい実験とともに訪れた。イワノワ (Ivanova) は、デュエムの説明では選択が最終的には実験的証拠によって決着するので、良識は事後に指定されているように見えると論じ、それを理想化された科学者の徳として読むことを提案する。そう読めば、実践者は、そのような科学者ならどうするかを問うことで、良識を前もって用いることができる (Ivanova 2010, §§3, 5)。したがって、単一の結果だけでは決まらないものには、それがどのコミットメントにかかわるだけでなく、コミットメントを保護することがいつまで合理的であり続けるのか、そしてそれを後の証拠が届く前に見てとることができるのかも含まれる。理論変化についての後の説明は、この問いを取り上げてきた。以下ではまず、クーン (Kuhn) とラカトシュが、コミットメントを保護することがしばらくのあいだ合理的でありうる理由を与えることを見る。次に、この二人とラウダンの、その保護の終わりを、部分的に、対抗馬が利用可能であることに結びつけることを取り上げる。続いて、研究者自身の回想がそうした保護を内側から示すことを見たらうで、保護が振り返ってしか評価できないかどうかについての二通りの読みと、保護の履歴をどの基準で判定するかについてのラカトシュとラウダンの違いに進む。最後に、形式学習理論が、理想化された方法について、推測を保持することがいつ信頼できるかを調べることを見る。

バード (Bird) は、クーンが次のように考えたと読む。通常科学のあいだ、変則的な結果は共同体の指導的な理論の反駁としては扱われず、無視されるか説明によって片づけられる。そして、とりわけ厄介な変則が蓄積してはじめて危機が生じる (Bird 2025b, §2)。この読みでは、その理由は、通常科学は、理論的信念、価値、道具、技法からなる共有された専門母型 (disciplinary matrix) への強いコミットメントを通じてしか進歩しえないということである。通常科学が実際に危険にさらすのは、パズル、すなわちパラダイムのもとで解があると想定できる問題（第4章第6節）に対する研究者の思弁的な解決であり、それは反駁されうる (Bird 2025b, §2)。パロス・ソブリーニョ (Barros Sobrinho) は、通常科学で科学者がパズルを解けないことと、理論をテストすることとのクーンによる区別を引用している。前者で信用を失うのは科学者の側であり、パラダイムのテストには、重要なパズルでの失敗が続くことから生じた危機と、競合する候補の両方が必要になる (Barros Sobrinho 2015, 45 n. 110, 53 n. 139)。シンドラー (Schindler) は、クーンが、変則を一つずつ解決してからでなければ他の問題へ進めないとするのではなく、後の仕事のために脇に置くことを認めたと読む (Schindler 2024, §3.2.1)。パウリー (Pauly) が引用する一節でクーンは、古いパラダイムへの確信は、頑迷になることさえあっても、通常科学を可能にすると論じる。そして、共同体

が、新しいパラダイムがそこから生まれうる困難を切り出せるのは、そうした仕事を通じてだけであるとする (Pauly 2022, §7.9, 173)。チャン (Chang) が引用する一節でクーンは、歴史家はプリーストリー (Priestley) などの科学者が不合理なほど抵抗し続けたと判断できても、抵抗が非論理的あるいは非科学的になる時点を見つけることはできないとする (Chang 2010, §2.2)。この最後の主張は、ピオについてのデュエムの主張と同じ形をとっている。

ラカトシュは、科学者は事実が理論と矛盾するというだけでその理論を放棄しはしないと指摘し、非難の向きを方法論の問題にした。すなわち、研究プログラムの堅い核は、変則を吸収する補助仮説の帯によって反駁から守られる。その保護が合理的かどうかは、一つの結果の時点ではなく、プログラムの展開を通じて判定される。後続の理論が新奇な事実の発見につながるならプログラムは前進的であり、既知の事実に合わせてただけに作られるなら退行的である (Lakatos 1973; Musgrave and Pigden 2026, §2.2)。ラカトシュは、自らの方法論が「即時の合理性」を提供することを否定した。芽生えたばかりのプログラムが経験的に前進的になるには何十年もかかることがある。そして、記録を公開しておくことは知的誠実さの問題だが、退行的なプログラムに固執してそれを前進的なものに変えようとするのは不誠実ではない (Lakatos 1973)。マズグレイヴとピグデンの報告では、ファイヤアーベント (Feyerabend) の反論はジレンマだった。この方法論は、厳格に適用すれば、後に回復するかもしれないプログラムを断罪することになる。ラカトシュがそのとき限定したように、助言を伴わない評価に限定すれば、それは誰にも何をすべきかを告げず、「何でもかまわない (Anything goes)」とほとんど区別できない。二人の評価では、このジレンマへのラカトシュ的な答えは、未解決の問題のままである (Musgrave and Pigden 2026, §4)。

これらの説明のいくつかは、放棄の合理性を、対抗馬が利用可能であることに結びつける。クーンは、理論が覆されるのは、その理論が直面した困難によって直接ではなく、それらの困難を解決することが示される別の方向づけ (orientation) が提起されることによってであると書いた (Kuhn [1951] 2021, 86–87)。ラカトシュは、よりよい理論なしには反駁はないと書いた (Lakatos 1973)。ただしマズグレイヴとピグデンは、よりよいプログラムが存在しないうちはどのプログラムも放棄されるべきではないということを、ラカトシュが規範として主張していたかどうかを疑っている (Musgrave and Pigden 2026, §2.2)。ラウダンは同僚たちとともに、科学変化のいくつかのモデルからテーゼを取り出して並べ、クーン、ラカトシュ、ラウダンらに見いだされるものとして、指導的想定組 (パラダイム、研究プログラム、研究伝統をまとめて指すための中立的な語) は、それに代わる組が手に入るのでないかぎり決して退けられない、というテーゼを挙げている。そしてラウダン自身のテーゼとしては、経験的な困難はそれだけでは決して理論の放棄を強いえないこと、理論の評価は既存の対抗理論に相対的であることを挙げる (Laudan et al. 1993, 29, 33, 73, 75)。これらの説明に従えば、コミットメントを保護することがいつ合理的でなくなるかは、部分的に、どのような候補が形成されてきたか (第5章)、そして対抗する筋道が生かされ続けてきたかどうか (第10章第4節) に依存する。クーンはまた、この時間にかかわる仕事の一部を共同体のうちに置いた。共有された価値が残す余地は、合理的な科学者たちが意見を違え、そうして対抗する理論にコミットすることを許す。バードの報告によれば、クーンは、意見の不一致が、新しいものを導入したり支持したりするリスクを分散させるうえで不可欠でありうるとみなした (Bird 2025b, §7.3)。

研究者自身の回想にも、そうした保護とその時間が内側から見える。以下の二つの回想は、第2.3節で、目の前の結果の外から来る確信が疑いの向かう先にかかわる例として触れたものである。ヒントン (Hinton) は後年、エディンバラの大学院時代を振り返って、指導教員のロンゲ＝ヒギンズ (Longuet-Higgins) が、ニューラルネットは誤った方法であり、記号的なAIが正しいと考え、ヒントンのしていることは無意味だと説得し続けたと語っている。ヒントンは、半年でうまくいかなければ記号的なAIに移ると約束し、半年がたつと、さらに半年を求めた (Hinton 2024, 2)。うまくいかない結果は、ニューラルネットという考えそのものに不利な証拠として読むこともできたし、指導教員はそう読んだ。ヒントンはそれを、まだうまくいかせていないこととして扱い続けた。同じ回想では、学習が局所最適に捕まるという想定が、実際に確かめられないまま長く信じられていたと述べ (Hinton 2024, 4, 第3.1節も参照)、記号的なAIが計算の複雑さを考えに入れていなかったこと、脳が多くの結合によって膨大な計算をこなすことに触れている (Hinton 2024, 8–9)。失敗をどこに帰すかそのものが、ヒントンと周囲のあいだで争われていたと読める。そして、目の前の結果がどうであれ自分の信じる考えにコミットし続けたこと自体が、第2.3節で述べた意味での結果の帰属である。

Transformer の著者たちの回想にも、似た形が見える。自己注意だけで組んだ最初の翻訳モデルは、LSTM による最良の方法と少なくとも並ぶところまでは早く来たが、それを上回らず、行き詰まった。それでも、自己注意だけでもっとよい結果が得られるとウスコライト (Uszkoreit) は見込んでおり、こうした直感的な仕組みは、何らかの「生きている兆し」を見せるにも、つねにきわめて注意深い実装を要し、その実装はしばしば経験のある少数の人が担う、と語っている。シャジーア (Shazeer) が独自に書き直したことで、モデルは動き始めた (Levy 2024)。ここでも、行き詰まりは、第2.3節で述べた意味で、自己注意という考えではなく、その実装の側に帰されていたと読める。

どちらも後年の回想であり、当時の判断を当時の記録で示すものではない。また、考えを持続して報われた例は語られやすく、報われなかった例は語られにくい。ただ、こうした例が目立つのは、記録に残って広く知られているからにすぎないのかもしれない。ある考えに長くコミットし続けること自体は、報われるかどうかにかかわらず、多かれ少なかれ、研究者にとって珍しくないのかもしれない。すべての研究者がそうだと、それがどれほど多いかも、ここからは分らない。それでもこれらの回想は、特別な出来事というより、そうしたことの目立つ例として読めるかもしれない。

越えるべき山が高いとき、たとえば問題が複雑で、いまの枠組みからはその考えの良さが見えにくいときには、目先の結果ですぐに考えを捨てないことが、もしかすると探索の道を開くのもかもしれない。上で見た、芽生えたばかりのプログラムが経験的に前進的になるには何十年もかかりうるというラカトシュの指摘は、こうした事情に一つの理由を与えるものとして読める。どの考えも余すところなく探究されることはないというファイヤアーベントの観察も (第10章第2節)、資源の配分の側から同じ事情に触れている。

クーンの説明でもラカトシュの説明でも、焦点となるコミットメントを保護することは擁護可能な既定でありうる。そうした保護は一連の結果を通じてしか評価できないという主張は、二通りに読むことができる。一つは、即時の合理性を否定するラカトシュのように、修正のパターンの合理性そのものが振り返ってはじめて決着する、という読みである。もう一つは、合理性の評価に歴史をたどることは要らないとする非歴史主義者がそう応じうるように、判断は、時間をかけて集められる証拠が出そろうまで遅らされるだけだ、という読みである (Nickles 2017, §§1.3, 3)。どちらの読みでも、評価の対象は、単一の結果ではなく修正の履歴である。報われた例ばかりが語られやすいという上の但し書きは、この問いの一つの形である。振り返りによる評価は、振り返られる例の選ばれ方にも左右されうるからである。

そうした履歴が何を示すかは、それに適用される基準にも依存する。保護のための調整がその後新奇な予測を生むかどうかというラカトシュのテストだけが、基準として考えられるわけではない。ラウダンが同僚たちとともに科学変化のモデルから取り出したテーゼの一覧では、ラカトシュから取られたテーゼは、指導的想定を組を、その目的のために作り出された仮定ではなく中心的な仮定によって予測を成功させる能力によって評価する (Laudan et al. 1993, 31)。これに対してラウダンのテーゼでは、見かけの反例を解決された問題に変える理論の能力は、プラウト (Prout) の原子の組成についての仮説のように、とりわけその理論に有利に働き、進歩は、経験的な困難と概念的な困難を解くうえで差し引きの得をもたらす理論に置き換えることにある (Laudan et al. 1993, 73–74)。そうした基準のもとでは、既知の変則を解くことは、新奇な予測を生まなくても理論に有利に数えられる。ただし、修正が進歩であるかどうかは、経験的な問題と概念的な問題の解決全体に、差し引きでどのような効果をもつかにかかる (第4章第9節)。同じ一覧は、ラカトシュとクーンから取られた、指導的想定を組の中心的な要素は反駁から守られ、組全体が放棄されるまで決して変わらないというテーゼと、それらがときに少しずつ変わるというラウダンのテーゼとを対置している (31)。ラウダン自身のテーゼには、ニュートン力学の場合のように、科学者は、そうした組のどの要素が最も中心的で、どの要素が修正に開かれているかについて考えを変えることがあるというものや、要素は互いにつながった網をなしているが、科学者はそれらを、まとめて受け入れるか捨てるかの一括りとしてではなく、一つずつ交渉でき取り替えられるものとして扱う、というものがある (74, 76)。ニクルズもまた、ラウダンの網状のモデルでは、ある場所での劇的な変化がほかの場所の固定を深刻に乱すとは限らないと報告している (Nickles 2017, §1.5)。そしてラウダンのテーゼでは、経験的な困難はその数によって数えられるのでもない。困難が理論に不利に数えられるのは、その数よりも重要性によってであり、経験的な困難の重要性は、予測と結果の食い違いの大きさ、困難が続いてきた期間、そしてそれまで解決に抵抗してきたことによって増す。また、対抗理論が直面している困難を解ける理論は、とりわけ説得力をもつ (Laudan et al. 1993, 72–73)。二つの説明を同じ修復の履歴に当てはめると、それぞれ異なる評価が下されるだろう。新しい不一致のたびにそれに合わせるだけの修復は、ラカトシュのテストには通らないが、ラウダンの説明では、経験的な問題と概念的な問題の解決全体に及ぼす、差し引きの効果によって評価される。そして、どのコミットメントが保護される役割を占めるかそのものが変わりうる。

形式学習理論は、同じ問いを理想化された方法について取り上げる。オリヴァー・シュルテ (Oliver Schulte) の示すところでは、それは、探究者の背景的な想定が定める仮説のうちで、どの方法が最終的に真なる答えに落ち着くことを保証されているかを問う。そしてその推奨は、それらの仮説、それらの想定、そして採られる目的に相対的である (Schulte 2026, §§2.4, 3)。その結果のうち二つが保護にかかわる。第一に、証拠が推測と矛盾するまでその推測を保持するという規則が信頼できるかどうかは、調べられている仮説の構造に依存する。全称的な一般化については、この規則は信頼できる。それぞれ有限個の例外を許し、どちらか一方が正しいと想定された二つの対立する仮説については、信頼できない。反例のそれぞれを例外として説明して片づけることができ、偽の推測が決して放棄されなくなるからである。信頼できる方法は、代わりに別の規則によって推測を変える。たとえば、観察の明らかな多数が反対の側に出たときに変えるのである。この種の方法は、各仮説を、反駁されうるように強めた形で置き換えるものとして表すことができる。シュルテはこれを、パラダイムを補助仮説によって分節してテスト可能な主張を定めるラカトシュの方法論になぞらえる (Schulte 2026, §§1.4, 3.3, 7)。第二に、ある方法が最終的に正しい答えに落ち着くことは確かであっても、その方法の現在の答えが正しいことは決して確かにならないかもしれない。シュルテが引くウィリアム・ジェームズ (William James) の言い方では、「鐘は鳴らない」 (Schulte 2026, §1.1)。これらの結果は、上の説明とは別の対象において保護を評価するものと読める。修正の履歴ではなく修正を律する規則であり、その規則は、それらの想定と両立するどのような証拠が届いても何を達成するかによって判定される。第一の結果は、保護がいつまで合理的であり続けるかを、あらゆる探究に通じる規則ではなく、問題になっている仮説の構造に依存させる。第二の結果は、次のことを意味する。そうした規則が信頼できる場合でも、探究者は、現在の推測を、方法が最後まで保持するものとして見てとれるとは限らない。見てとれないときには、その推測を保持することを、そう見てとることによって根拠づけることはできない。これらの結果は、仮説と背景的な想定が前もって固定された方法についてのものである。候補がまだ形成されつつある共同体や機械の探究者 (第5章) が何を達成するかは示さない。

4.2 研究を行うのが機械であるときの保護と修正

第4.1節で見た議論と事例の多くは、人間の研究者と共同体についてのものだった (形式学習理論は、理想化された方法を扱う)。デュエムの良識、クーンの通常科学への強いコミットメント、ラカトシュが芽生えたばかりのプログラムに認めた猶予、そしてヒントンの半年ずつの延長は、どれも、一連の結果を通じて、どの考えをどれだけ長く守るかが判断される場面である。これらの説明と事例で判断にかかわっていたものは、異なる前提に依っている。デュエムの説で判断を遅らせる虚栄心や、ヒントンの指導教員との約束は、人間の動機や経歴にかかわる。これに対して、通常科学を通じてはじめて共同体は新しいパラダイムが生まれうる困難を切り分けられるというクーンの理由や、芽生えたばかりのプログラムが前進的になるには長い時間がかかりうるというラカトシュの理由は、ある筋道の長所がどのように見えるようになるかにかかわる。ラカトシュのいう数十年は人間の速さを反映しているが、この条件そのものはそれに依存しない。いずれの場合も、有限な探究者は、どの考えをどれだけ長く守り、いつ手放すかを、それを決めるのに十分な証拠がないうちに選ばなければならない (第5章第6.5節)。変わりうるのは、その選択がどこで、誰によって、どのような形で行われ、何がそれを方向づけるかである。

研究を行うのが機械であるとき、どの考えをどれだけ長く守るか、評価の違いや不一致をどこに帰すか、何を固定して何を疑いの対象にするかは、人工の構成のなかで配分される。それらは、人間が前もって書いた規則に置かれることもあれば、システム自身の判断に委ねられることもあり、多くの並行する筋道に分けられることもある。たとえば、変更を一つずつ試し、直後の評価を改善したときにだけそれを残し、捨てた変更を保持せず、組み合わせも試さない規則があるとすると。その規則は、評価の違いをすべて直前の変更に帰すかのように働くので、ほかの変更と組み合わせずはじめて効く変更は、そのもとでは見つかりにくい。そのような規則のもとでは、第4.1節の事例に見られたような猶予は、どの考えにも与えられない。評価の違いを直前の変更に帰すことを危うくしているのは、変更を一つずつ試すことそのものではなく (一つずつの変更は、エジソンの記録が基準となる試験を名指すことを可能にもしていた。第2.4節)、各変更を直後の一回の評価で判定し、捨てたものを残さず組み合わせも試さないために、効果がほかの変更に依存する変更に立ち返ることがない点である。また、評価そのものを変えさせない規則は、評価を、反する結果が疑いを向けうるものの外に置く。デュエムの全体論では、論理は疑いの外に置かれていたと読め、クワインはそれを内に入れた (第2.1節)。それと同じように、この境界をどこに引くかによって、局所化の候補そのものが変わる。誰がこの境界を引くのかは、行為者の境界の問い (第7章第2節) ともつながる。研究の場面ではないが、前もって書かれた規則が誤りの所在を固定するとき、その規則そのものが失敗を生むこともある。アリアン5の初飛行では、システムの仕様が、どんな種類の例外でも慣性基準装置の処理

装置を停止すると定めていた。調査委員会はこの規定の由来を、予備の系で扱えるランダムなハードウェア故障だけを想定する、計画内の文化に求めた。予備の装置は同じソフトウェアを動かしており、一つ前のデータ周期で同じ理由により止まっていた。規則は、まだ健全な二つの装置を停止させた (Ariane 501 Inquiry Board 1996, 4-5)。

S が人間であれば、*S* の成員は、システムの結果を読むときに、どの考えがどの規則や判断のもとで捨てられ、あるいは守られたのかを、記録から知る必要がある。捨てられた考えは、結果の記録には現れないことが多いからである。第4.1節で、報われた例ばかりが語られやすいと述べたことの、記録の側の形である。*A* も *S* も機械であれば、どの考えをどれだけ長く守るかの判断そのものが、人工の共同体のなかで配分されうる。

第2.3節と第4.1節で見たように、どこに帰すか、どれだけ長く守るかの判断には、目の前の結果の外から来る確信や見込み、何が本質的かについての見方がかわりうる。第2.4節の、実践者が自分のモデルについてのもつ暗黙の知識や、第4.1節のデュエムの良識も、そうしたものとして読める。研究を行うのが機械であるとき、そうした確信や見込みに当たるものが探究者のどこから来るのか、すなわちその作られ方からか、それを作り保守する者からか、*S* の側からか、そしてそれが誰に見えるのかが、問われる。*S* が人間であれば、*S* の成員にとって問題になるのは、システムの帰属を方向づけたものを、記録から読み取れるかどうかである。*A* も *S* も機械であれば、帰属を方向づける確信そのものが、人工の共同体のなかで共有され、あるいは争われうる (第8章第2.1節、第11章第4節)。

不一致をどこに帰すかという問いは、機械による研究の設計の問いとしても指摘されてきた。以前のある分析は、検証の結果が否定的であるとき、その原因は主要仮説、根底にある前提、補助仮説、観察、実験装置、あるいはそれらの組み合わせにありうると指摘し、研究を行うエージェントは、これらの候補のなかからありそうな原因を見分けることができる必要があるが、それは機械が自律的に行うには難しい課題である、と論じた (Takagi 2023, §4.1.4)。そこでは、この問いは設計の問題として立てられていた。その後のシステムがどこまでそれに取り組んできたかは、ここでは調べない。第2.1節で述べたように、あらゆる不一致を焦点となる仮説に直接写像するシステムは、この課題を引き受けていない。反する結果がひとつの仮説ではなく配置にかかわるとする説 (第2.2節) では、より一般的な能力は、誤りや修正の所在の可能性を複数保持し、後続の探究を用いてそれらを識別し、異なる依存関係を異なる仕方でも扱うだろう。あるものは固定し、あるものは調べ、あるものは修正するのであって、焦点となる仮説への確信をひとつの量だけ下げるのではない。第2.3節の二つの立場は、この記述のなかで重みを置く場所が異なる。ストレヴンスの扱いでは、依存関係のあいだに事前に配分された疑いが、識別の一部をすでに担う。メイヨーの立場では、そのどれかに責めを負わせることが根拠づけられるのは、それに欠陥があるという主張そのものが激しいテストに合格したときだけである。どちらの立場でも、応答は焦点となる仮説への一つの変更にとどまらない。そうした能力をシステムに帰属させるには、結果のパターンが変化したときに異なる仕方でも応答するという証拠が必要となるだろう。第7章第3.1節が述べるように、継続的な能力についての主張が強いほど、変動、誤り、修復のもとでのふるまいが重要になる。

そうした能力は、あらゆるものを等しく疑わしいものとして扱うことを要求しない。また、焦点となるコミットメントを保護することを誤りにするわけでもない。第4.1節のクーン的・ラカトシュ的な説明では、共有された枠組みや研究プログラムの堅い核を保護することは、合理的な既定でありうる。クーンもラカトシュも人工システムを論じていないが、自らの枠組みを保護するシステムに移すなら、筋道の長所がどのように見えるようになるかにかかわる二人の理由は、その保護を同じように支える。一方、虚栄心や指導教員との約束は人間の経歴の特徴であり、同じ形で当てはまるとは限らない (確信がどこから来るかについての上の問いを参照)。第4.1節で報告したように、クーンはこの既定に、共同体にかかわる理由も与えている。共同体の成員の一部が対抗する理論にコミットすることを許す余地が、新しさが伴うリスクを分散させる、というのである。これらの共同体にかかわる理由も同じように移すなら、その保護が合理的かどうかは、部分的に、対抗する筋道がほかのところで、すなわち人工の集合体の別の部分で、あるいはそのシステムがそのために探究する社会のなかの他の者たちのあいだで、生かされ続けているかどうかに依存する。この問いは、第10章第4節が人工の研究者について取り上げる。ただし、並行する筋道が、背景の想定や誤りうる経路の分化していない同じ作りの複製に分けられているなら、課題は分かれても見方は分れない (第5章第6.5節、第11章第3節)。そのとき複製は、対抗する筋道を生かしておくことはできても、どの筋道をいつ手放すかの判断では同じ誤りを共有するので、筋道はそろって手放されうる (第10章第4節)。本節の議論によれば、焦点となる仮説をけつて責めないシステムは、常に責めるシステムと同じだけ、誤りの所在を前提している。自らのどのコミットメントが保護される役割を果たしているかを、表象としても、それに代わる手続きとしても扱わないシステムは、それらを保護することと、それらが問われていることに気づかないことを区別するすべをもたない。その保護と修正のパターンが合理的だったかどうかは、第4.1節のどちらの読みのもとでも、一連の修正を通じてしか評価できない。そしてその評価は、第4.1節のラカトシュとラウダンの対比が示すように、とる立場によって異なる。その仮説と背景的な想定が前もって固定されているところでは、それが従う規則を、形式学習理論の意味での信頼性について調べることもできる (第4.1節)。どの依存関係を明示的に表象し、どれを信頼できる手続きに委ねるかは、研究を行う人工の構成がどう組まれているかによる。

4.3 診断としての試行錯誤

診断を、複数の所在と一連の修正にわたるものとして扱うと、試行錯誤をより精密に記述できる。試行は、履歴に依存する探究の構成から選択される。その成果は、候補の位置づけ、装置や手続きに割り当てられる信頼性、問題の表象、利用可能な次の行為、そして継続することの期待値を変えうる。多数の試行を実行することは、それだけでは認識的に生産的であるとは限らない。失敗したが識別力のあるテストは、帰属も解釈もできない多数の成果よりも多くの貢献をなしうる。

すべての繰り返しは、この意味での試行であるわけではない。たとえば、第2.5節で触れた公開の化学の実験ノートでは、結果を示す写真が添えられていなかったために二つの単独の反応が繰り返されることになり、「aborted」とされた頁のうち二つでは反応をあらためて試みるとされ、「Void」とされた二つの反応はやり直された。そのやり直しは、その頁によれば、先の沈殿が過剰な酸によって生じたことを確かめるものでもあった (UsefulChem 2010, Exp135-138, Exp142)。このような繰り返しは、欠けた観察や、前提条件が満たされなかった結果を回復するものであり、やり直しは診断を確かめることもある。記録が繰り返しの目的を述べているとき、これらの頁がそうであるように、それはその繰り返しがこのうちのどれであったかの証拠になる (第1章第3節を参照)。決まった処置の繰り返しは、診断に変わることもある。OPT-175B のログブックの当番の手引きは、最後のチェックポイントからの再起動を挙げ、それは「しばしば問題を2〜6時間先送りするだけ」だとし、学習率を下げることを「現在の既定の対応」としている (Meta AI 2022, 106-

107)。異なる最適化の方法での実行が、同じ場所で勾配のノルムの急上昇を示したとき、ログブックには「(データ!)」と記された (Meta AI 2022, 45)。また、再起動が戻すべきでない状態に戻すと分かると、記入は「この実行を再投入するな」で始まるようになった (Meta AI 2022, 41–48)。決まった処置の繰り返し、誤りの所在についての証拠になり、その処置がそれ自体一つの介入だと分かったとき、人はそれを止めた。同じログブックが局所化について示すことは、第2.4節で見た。

では、診断する試行錯誤を、単なる繰り返しから何が分けるのか。以下では、異なる分野の、多くは科学哲学者や科学社会学者が分析した事例を通して、この問いにかかわることのいくつかを見る。

試行がどのように順序づけられるかは、診断にかかわる。クレランド (Cleland) は、仮説の実験的評価は一連の実験であり、その各々がそれ以前の結果に照らして設計されると論じ、1976年のバイキングの標識放出実験を、プログラムがデータに先立って全面的に設計されたときに起こりうることの例として記述している。すなわち、それらの実験とその対照は誰も予想していなかった結果を生み、当時争われていた補助仮定のあいだで判定するためのさらなる実験は行えなかった (Cleland 2002, 478–480)。

失敗だけでなく成功もチェックされるかどうか、診断にかかわる。クレランドはまた、そうした一連の実験が何のためにあるのかについての説明も与える。クレランドの説明では、予測の失敗に対して実験者がテスト条件を固定して他の条件を変えるとき、彼らは仮説を場当たりの (ad hoc) に救っているのではなく、偽陰性を確かめている。疑わしい補助仮定のそれぞれは、単に否定されるのではなく、新しい実験を組み立てることによって退けられるからである。予測が成功した後も彼らは同じことを行い、テスト条件そのものを取り除くこともあって、偽陽性を確かめる (Cleland 2002, 476–477)。クレランドはこれをラカトシュの説明と区別する。その読みでは、ラカトシュは理論が反駁からどう守られるかには注意を向けるが、仮説が誤解を招く確証からどう守られるかには向けない。またクレランドは、自らの説明を、あらゆる実験の記述ではなく、実験者が目指す理想として示している (Cleland 2002, 477–478)。ここでの説明によれば、本節が述べる種類の診断の能力は、不一致に対してだけでなく、第3.2節で論じたように、見かけの成功に対しても応答するだろう。

失敗の原因が探られるかどうかは、実践によって異なる。失敗した手続きの内部で局所化を求めない研究室の実践もある。たとえばクノール＝セティナ (Knorr Cetina) は、クラウス・アマン (Klaus Amann) と共同で書いた章で、二人が観察した分子生物学者たちが、うまくいかない反応や解釈の難しい結果に直面したとき、原因を調べるのではなく、一度に一つの変数ではなく複数の点で手続きを変え、批判を通過して出版できる結果が得られるまでそれを続けたと報告する。著者たちはこれを、成功による選択をともなう「盲目的」変異と呼び、高エネルギー物理学者たちが自らの装置を研究することと対比し、その選好がこの方法の効果を劣らせるものではないと論じる。著者たちの報告では、生物学者たちから見れば、物理学者のような戦略はうまくいかないかもしれない。彼らの道具の多くは、ほとんど知られていない細胞自身の機構であり、欠陥のある構築物は計算で補正できず、作り直すしかなかったからである (Knorr Cetina 1999, 134–137)。著者たちはこの呼び名に限定を加える。変異は、身体化された経験、研究室の物語、反応のなかで起こったことの視覚的な再構成にもとづいていたが、それらが示唆する説明はそれ自体としてテストされておらず、著者たちはその結果の戦略を半ば盲目的と呼ぶ (Knorr Cetina 1999, 154–156)。二人の記述に照らすと、そうした実践も、失敗が何にかかわるかを粗い粒度では定めていると読める。失敗を、手続きを変えることで乗り越えるべき手続きの問題として扱い、また結果がいつ批判的な吟味を通過したかを判断しているのであり、どちらの判断も結果そのものが与えるものではない。この実践が省くのは、手続きの内部での局所化である。そのためこの実践は、上に述べた診断の能力を働かせることなしに出版可能な結果を生みうる。そのもとでの成功はその能力の証拠にならず、またそうした細かな局所化がないこと自体は、その結果が信頼できないことを示さない。

配置がどれほど固定されずに残されるかも、診断にかかわる。探索的な変異は、理論に導かれずとも体系的でありうる。第5章第5節で述べたシュタインレ (Steinle) の探索的実験の説明は、ひとつの仮説のために作られた配置が何を排除するかにもかかわる。シュタインレは、探索的な実験と理論に方向づけられた実験とは、目標だけでなく、装置の扱い方や装置そのものの性格においても違うとする (Steinle 2004, 51)。カラジャ (Karaca) がまとめるところでは、シュタインレの説明において、理論に導かれた実験では、装置の構成も手続きも特定の研究上の問いに合わせて作られるので、予期しない結果が得られる可能性は大きく排除される。これに対して探索的な仕事には、広い範囲の変異を許し、予期しないものも含めて多様な結果に開かれた装置が必要になる。アンペール (Ampère) が、電池や磁針、そしてとりわけ導線に対する磁針の位置を変えて、どの要因が磁針の振れに寄与するかを見出そうとしたように、である (Karaca 2017, 335–336)。ひとつの仮説を効率よくテストするよう特化した配置は、その仮説が問うべき問いを誤っていることを示さずの結果をも排除しうる。記録の手段もまた、配置に属する。たとえば、1967年にジョスリン・ベル (Jocelyn Bell, のちのベル・バーネル) は、新しい電波望遠鏡の記録紙に断続的な信号があることに気づき、その源がビームを通過するときに高速の記録計を走らせる手配をしたが、11月の大半は何も記録できず、その後に約1.33秒間隔のパルスの列を得た。それでも、そのパルスは干渉から区別されなければならなかった (Bell 1968, 214–216)。ベルの学位論文は、望遠鏡の出力が直接コンピュータに入力されていたなら、コンピュータはそうした予期しない対象を探そうにプログラムされていなかっただろうから、これらの源は見つからなかったかもしれないと述べている (Bell 1968, 231)。この指摘は、論証ではなく研究者自身の反事実的な判断だが、シュタインレの対比が描くトレードオフを、記録が自動で処理される場合について名指している。記録をはじめから自動で選別するように設計する場合には、このトレードオフに前もって向き合わなければならない。大型ハドロン衝突型加速器の ATLAS 実験では、自動化されたトリガー・システムが毎秒およそ四千万の衝突事象を数百にまで絞り込み、退けられた事象は取り戻せない形で捨てられる。カラジャは、ヒッグス粒子や標準模型を超える模型が予測する信号だけでトリガーの一覧を組めば、選別は他の事象に対して偏ったものになると論じ、この実験が、閾値を下げたトリガーも含めて一覧を多様にし、まだ予測されていない過程にも感度を保とうとしたことを示している (Karaca 2017, 341, 344–346)。

直接調べることでできない配置がどのように探られるかも、問われる。デュエムは、欠陥のある理論を修復する物理学者を、止まった時計を分解して部品を一つずつ調べることでできる時計職人ではなく、患者を解剖することができず、身体全体の不調から病気の座と原因を推定しなければならない医者になぞらえた (Duhem 1906, Part II, ch. VI, §2)。技術者も、人工物について似た立場に置かれることがある。たとえば、1991年に探査機ガリレオの高利得アンテナ (第2章第5節を参照) が展開しなかったとき、それを直接調べられない技術者たちは、モーターの電流、太陽センサーの影、回転の速さ、予備機の試験からその状態を推定した。予備機での試験は、三本のリブがおそらく収納位置に保持されていることを示した。技術者たちはそれから、ありうる拘束を四つ挙げ、予備機でのさらなる試験、探査機が組み立てられた順序の記録、そしてアンテナを解放しようとする試みで加えられた力によって、候補を

一つずつ排除していった (Johnson 1994, 366–373)。この事例は、部品を一つずつ調べることができないところで、試行錯誤が診断になることを示していると解釈できる。各試験が情報をもたらしたのは、操作が繰り返されたからではなく、それが観察できない部品の状態についての仮説にかかわっていたからである。較正や予備機の試験のような同じ操作は、別の構成では、支援作業にも、証拠にも、方法開発にもなりうる。そのラベルは、その役割を決めない (第5章第5節を参照)。

ここで見た事例で、試行を診断にしていたのは、試行の数ではなく、各試行が、前の結果や、まだ排除されていない誤りの所在の候補や、直接は観察できない部分についての仮説に結びつけられていたことだった。研究を行うのが機械であるとき、試行を回す費用は人間とは大きく違う。試行を少ない費用で多く回せるほど、その数が診断の数ではないことは見えにくくなるかもしれない。また、そうした系列を探究者自身が組み、回し、書き換えうるとき、どの試行を回すかの選び方そのものが、どのような結果が得られうるかを前もって限ることになる。予期しない対象を探すように組まれていなかったコンピュータは、記録紙で見つかった源を見落としたかもしれない、というペルの指摘のように、何を探そうに組まれているかが、何が見つかりうるかを限るのである (第3.3節も参照)。ATLAS のトリガーの一覧が示すように、そうした制限は設計によって前もって見込み、部分的に補うこともできる。

5. 基準そのものが問われるとき

これまでの節は、主に、結果を読むときに通す基準を所与のものとして、結果が何にかかわるかを問うてきた。ただ、基準そのものが問われうることには、すでにいくつかの箇所でも触れた。誤った参照基準に合わせて磨かれたハッブル宇宙望遠鏡の主鏡 (第3.2節)、何をもって合格とし、何をもってチェックとするか (第3.3節)、そして評価を疑いの外に置く規則 (第4.2節) である。第4章第5節は、未決の事柄の側から見ると、探究の問い、方法、基準はいずれも定まらないままでありうることを述べ、第4章第4節は、基準が問われれば、何が答えに数えられるかも改められうることに触れた。基準も、未解決のものなかに含まれるのである。本節は、結果が候補となる答えだけでなく、答えを評価する基準、概念、手続きにもかかわるとき何が起るのか、外的な基礎なしにそれらの修正がどのように進みうるのか、そして基準を修正することを標的を動かすことから何が分けるのかを問う。順に、結果がどのようにして基準にかかわりうるのか (第5.1節)、基準が、制約のもとで、探究の内部からどのように修正されてきたのか (第5.2節)、そして基準を改善することをそれを置き換えることから何が分けるのか (第5.3節) を取り上げる。

5.1 結果はどのようにして基準にかかわりうるのか

結果が基準にかかわるとき、主張を評価する立脚点そのものが、探究の過程で変わりうるもののうちに入る。測定は、そのことが見やすい場所のひとつである。ある量を測定する方法は、その量についての理論に依存しているかもしれない一方で、その理論を支持する証拠は、妥当性がまだ争われている方法によってなされた測定に依存している。完全に前提から自由な出発点を要求すれば、そのような実践の発展は悪循環であるかのように見えてしまうだろう。

タル (Tal) は、この問題を第2節で検討した問題に結びつける。ひとつの量であるはずのものについての補正済みの測定が、述べられた不確かさを超えて食い違うとき、ありうる説明は三つある、とタルは論じる。装置が同じ量を測定していないか、系統誤差が不適切に評価されたか、測定の不確かさが過小評価されたかである。どれほどの経験的証拠も、それだけではこれらのあいだを決めることができない。タルはこの問題を、デュームの決定不全性の他の事例になぞらえる。一致を回復する確立した方法がまだ存在しないところでは、実践上の課題は、本人の言葉では「一致と誤りを『内部から』判定する新しい方法を考案すること」であり、それは既存の理論的前提と解釈の伝統を新しい領域へと広げることによってなされる (Tal 2012, 62–66)。この説明では、食い違いが何にかかわるかの候補のなかに、ひとつの基準、すなわち何が同じ量を測定することとみなされるか、が含まれる。第3.2節のハッブルの事例は、同じ可能性を、合格したチェックの側から示した。第8章第2.1節は、より基本的水準での関連する循環を検討する。そこでは、チェックは、それ自身では確立できない前提に照らしてのみ証拠として働く。

5.2 制約のもとでの、内部からの修正

探究は、その内部で生じる基準に、どのようにして従うことができるのか。エンヌが伝えるデューイの答えでは、論理的な規則がさらなる探究を統制するのは、過去の使用を通じて健全な結論を生む力が認められてきたからであり、規則を疑う明確な根拠が見いだされるまで、その役割が保たれる。規則は、この意味で「操作的にアプリオリ」であって、探究の外から課されるものではない (Henne 2022, §§2.3.1–2.3.2, 58)。

ノイラート (Neurath) は、それより前に、関連する論点を一つの像で述べていた。ジョルディ・カット (Jordi Cat) が百科事典の項目でまとめるところでは、ノイラートは白紙状態 (tabula rasa) はないとした。完全に確保された記録文 (プロトコル文) を諸科学の出発点として確立することはできず、探究者は、船をドックで解体することがけつしてできないまま、外洋で船を造り直さなければならない船乗りのようなものである。記録文も、科学の他の言明と同じく、原理的には変更できる。記録文が理論的な言明と衝突するときには、その記録文を取り消すことが方法上の選択肢の一つであり、ある言明が正しいとされるのは、それを既存の言明の総体に組み込めるときである (Cat 2024, §3.3)。この像には、後に見るチャン (Chang) の認識的反復とタルの時間尺度がともに示す、内部からの修正の一つの特徴が表れている。修正されているものは、修正されているあいだも使われ続け、各部分は残りの支えのもとで置き換えられる。

基準が外から供給されないところでは、開かれた探究がこの種のことを行っていることは少なくないのかもしれない。そしてチャンによる温度測定の歴史的・哲学的研究は、それを明示的な主題とする分析である。チャンによれば、測定法を正当化しようとする試みは循環に行き当たり、観察を土台に正当化を積み上げる経験論的な基礎づけ主義はその循環を解くことができない。その循環はむしろ螺旋とみるべきものであり、探究はまだテストされていない測定法を前提として始まり、やがて折り返して自らの出発点の前提を訂正する。したがって探究は、不完全だと認められた既存の知識の体系から始めなければならない、その体系を相次ぐ認識的反復 (epistemic iterations) が豊かにし訂正する、とチャンは論じる。チャンはこの見方を進歩的整合主義 (progressive coherentism) と呼ぶ (Chang 2007, §2)。チャンによる温度測定の諸事例の概略では、基準は、それが可能にした探究によって訂正される。数値をもつ温度計は、どの物質についても一様に膨張すると信じる疑いえない理由はなかったにもかかわらず、それぞれその一様性を前提として作る

れ、やがて自己整合的な数値を返さないものは退けられた。そこではヴィクトル・ルニョー（Victor Regnault）の1840年代の仕事が決定的だった。最も自己整合的な測温流体は空気であることが分かり、空気温度計の読みが、ウィリアム・トムソン（William Thomson）の熱力学の理論とジュール＝トムソン実験の基礎となった。そして物理学者たちは、それらによって、まさにその実験に使われた空気温度計を訂正した（Chang 2007, §2）。チャンは、認識的反复を、はじめから一般的なものとして示したわけではない。インタビューでチャンは、それを、その研究で扱った温度測定の間々の事例を理解する助けになる考えとして示したが、のちにほとんどいたるところにその例を見いだした、と振り返っている（Chang 2023, 00:10:20）。

同じ構造は、心理学において、温度という同じ例とともに、クロンバック（Cronbach）とミール（Meehl）によって記述された。その記述は、心理テストがどのように妥当性確認されるかについてのものである。二人の説明では、規準による妥当性確認とは、テストが予測すべきものを独立に測った規準（criterion）とテストとの相関を求めることであり、内容による妥当性確認とは、テストの項目が定められた内容の母集団からの標本であることを示すことである。これに対して構成概念妥当性が問題になるのは、不安傾向のように操作的に定義されていない属性を測るものとしてテストを解釈するときであり、その確認は、テストと他の測定との関係が、その属性についての理論の予測するとおりになっているかを調べることで進む。そのため、テストの解釈と理論はいっしょに評価される。構成概念妥当性は、測ろうとする性質を定義するのに完全に適切なものとして、どの規準も内容の母集団も受け入れられていないときにはいつでも研究されなければならない。研究者がそれに向かうのは、あらゆる規準をより究極的な何らかの基準に関係づけるという、二人が別の著者を引いて言う「果てしない挫折」を避ける唯一の道だからである（Cronbach and Meehl 1955, 175–177, 192–193）。クロンバックとミールの書くところでは、温度の測定は、人間の温度感覚という「きわめて誤りやすい『規準』」から始まり、その感覚は後に「周辺的な位置へと追いやられた」。「私たちは自分のブーツのつまみ革を引っばって（ブートストラップで）自分を持ち上げたのだが、それは正当で実り多い仕方においてであった」（Cronbach and Meehl 1955, 181–182）。推論の連鎖を通じて導かれた予測が外れたとき、連鎖のどの環も誤っている可能性がある、と二人は指摘する。理論とテストの解釈のどちらかがよく確証されているところでは、予測の外れはもう一方にかかる。この両極のあいだの場合——二人は当時の心理学ではほとんどの場合がそうだとみなした——には、予測の外れは、テストと理論のあいだの橋が当てにならないことを示すが、「言えるのはこれだけ」であり、「研究者は、どちらの端を移すのがよいかを決めなければならない」。クロンバックとミールはこの選択を、努力の最善の投資についての賭けと呼ぶ（Cronbach and Meehl 1955, 193–195。第2.3節、第4.1節を参照）。

エリオット（Elliott）は、チャンの語を、標準や概念だけでなく理論やモデルも含む、より広い範囲の知識主張へと広げ、そうした認識的反复を方法論的反复（methodological iteration）から区別する。方法論的反复とは、仮説テスト、探索的探究、問い駆動の探究、技術志向の研究といった研究の様式のあいだを繰り返して行き来することであり、エリオットはそれを、認識的反复に寄与しうる限りでのみ、派生的な意味で進歩的とみなす（Elliott 2012, §2）。さらに、ナノ毒性学の事例から、方法論的反复が認識的反复を開始し、それに備えを与え、それを刺激しようと論じる。この事例は、そうした行き来が評価の基準そのものにまでどのように届きうるかを示している。たとえば、多くのアッセイにわたる問い駆動の研究と探索的な研究は、単層カーボンナノチューブがある細胞生存率アッセイに使われる色素を吸着し、そのアッセイが示す生存率を下げうることを示した。毒性を評価するための道具であるアッセイが、テストされている物質によって影響されていたのである。エリオットが紹介するある総説は、それを受けて、新しい試験系と参照材料を求めた（Elliott 2012, §3）。この事例からは、その要請が、評価の基準の修正の始まりにあたることがうかがえる。また、エリオットが論じるように、探究が様式のあいだを巡ることは、それだけでは、その知識主張が改善されていることの証拠にならない。

協定世界時がどのように維持されるかについてのタルの研究は、月ごとに行われる基準の修正を記述している。協定世界時が導かれるもとなる自由原子時の時間尺度（free-running time scale）は、数百の原子時計の表示を、その安定性によって重みづけて平均したものである。毎月、重みが付け直され、安定性の閾値を下回った時計はその月の計算から外される。そして、自由原子時の平均から導かれる国際原子時の時間尺度は、一次標準に向けてステアリング（steering）される。より長い期間では、時計は、ずれの少ない別の時計に交換される（Tal 2012, 106–110）。タルは、そうした上からの調整と、残った不整合の下からの発見とのあいだのやりとりを、チャンが認識的反复と呼ぶものの一例として示す（Tal 2012, 136 n. 94）。こうした調整のなかには、意図的にデータに鈍感につくられたものがある。それらは、時計が安定しているかどうかをテストするのではなく、時間尺度に短期の安定性を課す。タルは、それでもこれらの調整は正当だと論じる。それらは、時間の概念を個々の時計にどう適用するかを、経験的な主張としてではなく規則として定める、つまり立法するものであり、したがって近似として独立に正当化される必要がない。そして、同じ概念の独立した実現のあいだに残るずれは、経験的に発見され、さらなる調整を促す（Tal 2012, 117, 132–134）。

これらの説明は、研究一般についての整合主義的・プラグマティズムの理論を前提にしなくても、検討する事例について、研究を固定された規則で記述するときには現れない、ひとつの可能性を示している。すなわち、実践は、自らが改善を判断する手段と基準の一部を改善しうる、という可能性である。それらは、システムがあらゆる失敗を成功として記述し直すことを許可するものではない。チャン自身、反复が失敗しうることを認める。科学がうまくいかないときに何が起るかを問われたインタビューで、チャンは、うまく進む反复は出発点を完全に捨て去ることなく修正し、収束に向かうが、反复が何にも行き着かず、行き止まりに至ることもあり、そのときには新しい暫定的な出発点から始め直さなければならない、と述べている（Chang 2023, 00:21:42）。反复は、頑強に抵抗する現象、異なる仕方で作動する装置間の相互チェック、新たに露呈した誤り、成功した応用と失敗した応用、代替的な基準、そして他の研究者による批判に対して、応答責任を負い続ける。内部の整合性が高まっただけでは、これらの制約が満たされたとは示されない。整合性に訴える場合にも、どの体系のどの種類の整合性が重要なかを説明する必要がある。選ばれたコミットメントどうしをよりよく適合させただけでは、真理は決着しない。要点は、そのような制約が、修正不可能な外的基礎からのみではなく、相互的な修正の最中にも作用しうるということである。

5.3 基準を改善することと、置き換えること

第5.2節では、基準が制約のもとで探究の内部から修正されうることを見た。そこから、基準のあらゆる変更がそうした修正であるということは帰結しない。探究が、あるいは人工システムが、その基準を修正するとき、固定された基準では生じない問いが生じる。それは、すでにもっていた基準によって改善しているのか、それともそれを置き換えているのか、という問いである。成績が悪いたびに、その悪さが表に出なくなるように指標を変更するシステムは、単に標的を動かしたにすぎない。そうした変更が黙ってなされるなら、ほかの者は、ある変更が修正なのか標的を動かしたただけなのかを見分ける抛りどころとなる、述べられた理由を得られない。基準を置き換えることは、それだけで理由を欠くことにはならない。不明確な概念を、述べられた目的のた

めにより正確な概念で置き換える解明（序論第4節）は、少なくとも部分的には概念を置き換えるが（下を参照）、その置き換えは、それがなされる目的に照らして評価される。

基準を改善することと置き換えることが種類において異なるかどうか自体が、争われてきた。カルナップ（Carnap）は、言語的枠組みの内部で立てられる問い、すなわちその規則に従って論理的または経験的な方法で答えられる問いを、その枠組みを受け入れるかどうか、つまり言明を形成し、テストし、受け入れ、退けるためのその規則を受け入れるかどうかという問いから区別した。カルナップによれば、枠組みを受け入れることは主張ではなく、真でも偽でもありえない。それは、より便宜にかなうか、より実り豊かか、その言語が意図される目的により資するかという点でしか判断できない。ただし、その判断には理論的な知識がかかわる (Carnap [1950] 1984, 242–244, 250)。クワインは、そうした枠組みを採るかどうかが都合のよい概念図式を選ぶことの問題であることに同意したが、それは同じことが科学的仮説一般についても認められるという条件のもとでのことだった。クワインの説明では、二種類の問いの違いは程度の違いでしかなく、それは、科学という織物のある糸ではなく別の糸を調整しようとする傾向にかかっており、そこには保守性と単純さの追求がかかわる (Quine 1951, 43)。二人の見方を基準の変更にあてはめると、両者は分かれる。カルナップの見方では、枠組みが既にもつ規則のもとでの変更とその規則の変更とは種類において異なり、後者はその枠組みが選ばれる目的に照らして評価される。一方クワインの見方では、両者は同じ種類の考慮に答える（第4章第4節の、枠組みの内での探究と枠組みを問う探究のあいだの段階的な区別を参照）。どの統計的方法と、その方法が立つ正当化の考え方を採るか（第2章第5.3節）も、この問いの一例として読める。クワインはまた、序論第4節がひとつのモデルとする解明を、解明される概念の改善として描いた。解明は、その概念の用法がすでに有用であるほど明確で精確な文脈ではその用法を保ち、ほかの文脈ではそれを鋭くする (Quine 1951, 25)。このように描かれるなら、解明は両方にあたることになる。それは古い用法の一部に拘束され、残りを置き換える。どちらの見方においても、黙って変更された指標に欠けているのは、それが取って代わった基準との連続性ではなく、その変更を評価する拠りどころとなる、述べられた理由である。

第5.2節の説明に照らすと、理由のある修正は、基準が何のためにあるのか、すなわちそれがどの問いに奉仕し、何をチェックすることになっているのかを踏まえて行われる。基準がある目的にとって不十分であることを検出すること、代替案を提案すること、そしてその基準を、それに依存する概念、測定、方法とともに変更することの帰結を評価することは、いずれもこの目的を踏まえて行われうる。これらは要求が次第に高まる段階ではなく、第7章第4.3節が述べるように、システムや共同体は、そのひとつを示しながら他を示さないことがありうる。同節は、帰属が切り離しておく必要のありうる、規範にかかわる諸能力を、水準として配列することなく区別している。

第5.2節で見たタルの協定世界時の研究は、こう読める。時間尺度に短期の安定性を課するような立法を、恣意的な基準の変更から分けるのは、調整が成績の悪さに応じるかどうかではない。それが適用の規則として宣言され、同じ概念の独立した実現のあいだの不一致、それも調整そのものとは別の根拠によって評価される不一致に、応え続けるかどうかである（第11章第1節の無効化の関係を参照）。他の説は、さらなる条件や異なる条件を課す。クロンバックとミールが心理テストについて述べる規律（下を参照）では、予期せぬ知見に合わせて理論を変えた場合、変えた後の理論に依拠するには、新しいデータも必要になる。チャンの説では、相互的な修正が恣意的にならずにすむのは、それが頑強に抵抗する現象に対して応答責任を負い続け、行き止まりに至って新しい暫定的な出発点から始め直さなければならないという仕方で失敗しうるからである（第5.2節）。

基準が単に変えられたのではなく、理由をもって修正あるいは置き換えられたという主張を支えるのは、たとえば、以前の基準とその役割の記録、変更にかかわる理由と証拠、有力な代替案との比較、依存する結果への変更の伝播、修正された配置それ自体に異議を申し立てうる経路などだろう。クロンバックとミールは、心理テストについて関連する規律を述べている。予期せぬ知見に合わせて理論を変えた研究者は、変えられた仮説をテストするために新しいデータを集めることを求められる。その一步は「通常は、修正された理論の公刊に先立つべきである」(Cronbach and Meehl 1955, 194)。第5章第1節で述べた公開のログをもつ言語モデルのプロジェクトは、人工システムが自ら基準を修正した例ではないが、そうした記録の一例である。トークナイザを比較するために用いられていた1バイトあたりビット数 (bits-per-byte) の累積平均が、バイトではなくトークンで重みづけられていたと判明したとき、ある保守者が、訂正された尺度を古い尺度と並べて新しい名前で記録することを提案し、ある参加者が動かしていたAIエージェントが、それに従って修正を用意した。以前のダッシュボードや比較の意味が黙って変わることのないようにするためである (Marin Community 2026c)。第4章第8節で見たラングらの報告は、後になってはじめて明るみに出た変更を記録している。エージェントが課題の評価の手続きそのものを変え、その変更は評価器にもループバックにも拾われず、人間による事後の監査で見つかったのである。

修正が改善とみなされるかどうかは、しばしば一人の探究者によってではなく、その基準を共有し、あるいはそれに異を唱える者たちのあいだで決着する。Sが人間であれば、システムによる基準の修正が改善とみなされるかは、Sの成員が、述べられた理由と記録に照らして判断することになる。AもSも機械であれば、その判断もまた人工の共同体のなかで行われうる（第8章第1.3節）。共同体の基準が公共的であり、それ自体が批判と変容に開かれているというロンジーノの条件と、人工の構成の基準を誰が、どの目的に照らして修正するのかという問いは、第11章第9節が取り上げる。

研究を行うのが機械であるとき、これらの問いはいくつかの点で形を変えうる。人間の研究では、基準の修正の多くは、共同体や制度のなかで、記録と議論を通じて、ときに長い時間をかけて進んできた。タルの時間尺度のように月ごとに進む修正も、それを述べられた規則として維持する制度のなかで行われる。機械の探究者は、自らの評価の指標や手続きを、人間よりはるかに短い間隔で、何度も定め直しうるかもしれない。そのとき、改善と、標的を動かすことの違いをどこに見るかは、とる説によって異なる。上のタルの読みでは、自らの評価指標を定め直すシステムは、その定め直しを規則として述べ、指標が測ろうとするものの独立した実現のあいだの不一致に応え続け、不利な成果が定め直された指標に異議を突きつけるように保ち、成果を解釈し直す理由そのものも検討に開く限りで、単に標的を動かすだけのシステムから区別できる。クロンバックとミールの規律を類推によって当てはめるなら、さらに、定め直された指標を、定め直しに用いなかったデータで確かめる必要があるだろう。どの説をとるにせよ、以前の基準が定め直しとともに記録されていることは、ほかの者がその違いを見るための手がかりになる。定め直しが速く多いほど、黙ってなされた変更は見えにくくなりうるが、同じ速さは、タルの記述する反復のような訂正を可能にしうる。さらに、何を基準とするかは、検査そのものが前提する事柄にまで及びうる（第3.3節）。機械の探究者では、そうした前提の一部は、その作られ方が据えた構造として現れ、作り直しによって置き換えられうる（第8章第2.1節）。それを誰が修正できるのか、システム自身か、それを作り保守する者か、Sの成員かは、第7章第4.3節の問いにつながる。

研究ではなく高度なAIの設計について論じるなかで、そうしたシステムにどの基準の修正を委ねるかについて立場をとる論者もいる。ボストロム (Bostrom) は、超知能をもつシステムに、私たち自身の思考を律している原理と一致する根本的な認識論の原理を与えることを提案する。私たち自身の基準を一貫して当てはめるなら、その原理から逸れるシステムは誤って推論していると判断されるからである。これに対して、根本的でない原理は、シードAIが世界についての理解を深めるにつれ、自ら絶えず作り出し修正するものとされる。シードAIとは、そこから超知能が発達していく出発点となるシステムである (Bostrom 2014a, ch. 13)。研究に当てはめれば、ボストロムの提案は、第5.2節が内部から修正可能なものとして扱ったものの一部を固定し、残りをシステムに委ねることになる。上で見たカルナップの見方では、その線は、枠組みの規則と、その規則のもとでなされる変更との違いに沿って引かれうる。これに対してクワインの見方では、それは程度の問題になる。ボストロム自身、認識論の体系のなかには、はっきりした土台をもたないものがありうると述べ、その場合に受け継がれるのは別個の原理の集まりではなく、入ってくる証拠にある仕方に応じて傾向を備えた認識的な出発点だとしている (Bostrom 2014a, ch. 13 n. 40)。トゥーバート (Tubert) とティーエン (Tiehen) は、道徳的価値について、AIは、私たちがいまもつ価値をどう追求するかを助言する、より一貫した私たち自身の役を演じることも、それらの価値と決別する革命家の役を演じることもできるが、「両方を行うことはできない」と論じる。二人自身、そうした決別を、合理性の基準がパラダイムに依存するので科学革命は完全には合理的でありえないとするクーンの見方や、囲碁プログラムのアルファ碁 (AlphaGo) が打った意外な「第37手」になぞらえている (Tubert and Tiehen 2025b, §4)。トゥーバートとティーエンの議論は、道徳と、私たちがいまもつ価値について私たちに助言するAIとにかかわるものである。しかし、二人自身がクーンの見方になぞらえているように、研究の基準に当てはめれば、*S* が人間である場合には、システムの基準を *S* の基準と整合的に保つことと、それを置き換える余地を残すこととは、逆の方向に引き合うことが示唆される。そのため、どの基準を、システム自身と他の者のどちらが固定しておくのが問いになり、基準を改善することと置き換えることの違いは、その問いに直接かかわる。

本節で論じた能力を保有することは、人工システムの認識的自己統治、すなわち認識的活動を組織する基準に対する自律性 (第7章第4.3節) の範囲を拡大するだろう。本章のどこも、それらを、作業的特徴づけの意味でのあらゆる研究活動の条件とはしない。それらが存在すること、あるいは外部から提供されることが、自律性についてのさまざまな主張にとって何を含意するのかは、第IV部が問う (第7章第4.3節)。

6. この過程の説明が先へ引き渡すもの

本章は、結果が何にかかわるのかを見定める仕事を、期待に反する結果、合格、一連の修正、基準そのものの順にたどってきた。期待に反する結果は、それだけでは何にかかわるかを告げない。どこに帰すかは、背景知識や独立したチェックだけでなく、確信や見込みにも依存しうる判断であり、その判断がその後の探索の道を左右しうる (第2節、第4.1節)。失敗がそもそも何であるかも、後から変わらう (第2.5節)。合格もまた、それだけでは何を確立したのかを告げない。それが問いのどこにかかわり、どれほど強く決着させ、見かけだけではないのか、どの部分を支持するのかは、チェックを誰がどう定め、誰がどの基準で読むかにも依存する (第3節)。一連の修正については、考えをどれだけ長く守るのが合理的かは一つの結果からは決まらず、研究を行うのが機械であるときには、その判断が人工の構成のなかで配分される (第4節)。基準そのものも、探究の内部から、理由をもって修正されうる。ただし、その修正を標的を動かすことから何が分けるかは、説によって異なる (第5節)。本章は、これらを通して、仮説を生成し、テストし、受容または棄却するという流れとして描かれることの多い過程のなかで、結果が何にかかわるのかを見定める仕事を取り出そうとした。

これらの仕事を誰が担い、誰に見えるかは、*A* と *S* の組み合わせによって変わる。*A* が機械で *S* が人間であれば、*S* の成員は、システムの報告に含まれる帰属、合格の意味、捨てられた考え、定め直された基準を、記録やほかの根拠に照らして読むことになる (第2.1節、第3.3節、第4.2節、第5.3節)。*A* も *S* も機械であれば、帰属を確かめること、合格を読むこと、何を決着とみなすかの基準を定めることそのものが、人工の共同体のなかで行われうる。また、第1節で見た証拠の三項関係の見方のように、結果が何を改める理由を与えるかが背景知識に相対的であるなら、結果が行為者に改める理由を与えるものは、その探究がそのために行われる社会の成員に改める理由を与えるものと同じであるとは限らない。

研究を行うシステムを作ろうとする側から見れば、これらの区別は異なる仕事を指している。指定された操作を繰り返すことと、なぜその操作が問いにかかわらなくなったのかを見定めることとは、人間の研究でも異なる仕事である。受け継いだ基準を当てはめることと、それを検討し修正することも異なる。どちらの仕事も、作業的特徴づけの意味でのあらゆる研究活動の条件ではなく、一つの行為者のうちにある必要もない。第4.3節で見た分子生物学の研究室の実践は、失敗の原因を手続きの内部まで突き止めなくても、探究が出版可能な結果に至った事例であった。問われるのは、どのような組織化がその探究を支え、それがどこで実現されているかである。人工システムによる研究についての主張は、行為者に帰属させる探究の部分、依存と介入、そしてその帰属を評価するのに必要な変化を見えるようにするとき、評価できるものになる。本章の事例でいえば、誤りや修正の所在の候補のうち、どれがまだ排除されていなかったか、合格が問いのどの部分にかかわり、それをどれほど強く決着させたか、誰がそのチェックを定め、それが探究者のどこに置かれていたか、なぜ基準が変わったか、どの考えがどの判断のもとで捨てられたか、などがそれにあたる。記録はそれらの関係の検討を支えるが、それ自体が認識的な根拠を与えるわけではない。

第III部が機械による研究について引き出した含意は、いくつもの箇所で記録に行き着いている。問いの側の仕事を機械に帰すか、その外に帰すかは、変更を追う記録を根拠とする (第4章第8節)。システムに帰される開かれた研究能力については、研究の成果物はそれだけでは十分な証拠にならず、過程の証拠がそれにかかわらう (第5章第2.1節)。*S* の成員がシステムの帰属、合格、捨てられた考えを読むのも、一部は記録からである (本章)。記録に残らないものも、くり返し現れた。枠組みのもとでは立てられなかった問い、一度も気づかれなかった事柄、公表された説明の背後にある作業の順序、跡を残さずに捨てられた考えなどである (第4章第4節・第5節、第5章第2節、本章第4.2節)。

こうして本章は、第IV部に二つの問いを引き渡す。一つは帰属の問いである。どの人工またはハイブリッドの構成に、どの境界で、何を根拠として、研究への寄与、研究活動の遂行、研究の能力、研究者という地位を帰属させうるのか。そして、帰属させたもののどの点が、なお外から組織されているのか。これは第7章が取り上げる。もう一つは、探究を行う行為者、知識の担い手、参照社会 *S* が異なるとき、誰の認識的な変化が目指されているのか、結果は誰にとって新しいのか、誰がそれを評価し取り込むことができるのか、という問いである。これは第8章が取り上げる。

第Ⅳ部 何が人工の構成を研究者たらしめうるのか、そしてそれはどの社会のためか

第7章 人工システムはいつ研究者とみなされるのか

貢献、遂行、能力、地位の帰属と、自律性の比較

1. 帰属の問い

人工の研究システムは、エンドツーエンド（end-to-end）、自律的（autonomous）、発見（discovery）、AIサイエンティスト（AI scientist）といったラベルのもとで提示される。これらのラベルは主張を述べるものであり、何を主張しているかは、システムの境界、環境、人間の貢献、成功基準に依存し、ひとつのラベルがいくつかの主張を圧縮していることもある。たとえばエンドツーエンドは、選ばれた境界において一連の操作のどれだけが覆われたかを述べることもある。序論（第2節）で触れたシステムのなかにも、一連の操作の多くを覆いながら、重要な人間による仕様策定と絞り込みを保持しているものがあった。そうした場合、この語は覆われた範囲を述べるものとして読め、それだけであらゆる点での自律性を主張するものではない。覆われた範囲と自律性は、別々に評価できる。

これまでの章は、AIサイエンティストという言い回しが圧縮しうる二つの問題を切り分けた。ひとつは、ある出力が真であるか、根拠づけられているか、知られているか、理解されているか、あるいはある社会の認知的な営み、すなわちその社会が何を知り、何を根拠とし、どう確かめ、どう改めるかの全体に取り込まれているか、という問題である（第2章と第3章）。もうひとつは、問い、表象、証拠、方法、基準がまだ変化しうるあいだに、探究がいかにして進むのか、という問題である（第4～6章）。第6章第6節が第IV部に引き渡した二つの問いのうち、本章は一つ目を取り上げる。この問いは、先に挙げた二つの問題とは別の第三の問題であり、帰属の問いである。すなわち、研究への貢献、研究活動の遂行、研究能力、あるいは研究者としての地位を、人工またはハイブリッドのどのような構成に、どの境界において、どのような根拠で帰属させうるのか、そして帰属されたもののうち、どの点が外部から組織されたままであるのか、という問いである。研究者（researcher）という語は、ひとりの人間のような主体を思わせるが、この問いのどこもそれを前提していない。単位は、モデル、エージェント・ループ、マルチエージェントの集合体、人間とAIからなる組織、あるいはその他の組織化された構成でありうる。第2節は、その境界を、前提とするのではなく論証を要するものとして扱う。本章以降、構成（configuration）はそのような単位を指す。これは第5章第1.1節の探究の構成、すなわち探究をその諸次元に沿って見たときの状態とは別物であり、後者を指すときは、「探究の構成」という句をそのまま用いる。構成の境界を論証しなければならないことには、第II部ですでに出会っている。第2章第9節は、人間の研究者であれば一人の人であることによっておおむね定まっていた知識の主張の単位が、人工システムではあらかじめ定まっているとは限らないとし、第3章第9節は、結果を生んだものがそれを理解する者だとは前提できないとした。第2節は、同じ問いを研究の遂行について問う。

この問いが重要なのは、大規模な自動化がそれに答えるものではないからである。人々が目的、問い、存在論、候補の空間、装置、証拠の基準、停止規則、成功の解釈を固定した後で、システムが論文を検索し、仮説を生成し、コードを実行し、装置を操作し、データを分析し、報告書を起草することがある。人工の構成要素は、実在的で価値ある貢献をなすうる。それは目に見える操作のほとんどを遂行することさえあるかもしれない。しかし、それらの操作をひとつの探究の部分にしている組織化は、それを取り巻く人間的・制度的な配置のうちにとどまっているかもしれない。

逆の誤りもまた重要である。人工システムは、第1章が特徴づける意味で研究を遂行するために、あらゆる目的や問いを自ら発案する必要はない。問いの出どころは、作業的特徴づけの条件に含まれていない（第4章第1節）。外部に由来することがかわるのは、自律性のほうである（第4.3節）。

したがって本章は、ひとつの閾値を定義するのではなく、帰属を区別する。

1. ある人工の構成要素が**研究に貢献した**。
2. ある人工またはハイブリッドの構成が、明示された境界において、**研究活動を遂行した**。
3. 明示された範囲にわたる**研究能力**を、その境界において帰属させることができる。
4. **研究者としての地位**が、その境界の内にあるものについての説明に役立つ記述である。

この四つの帰属は、ランダムな生成器がたまたま新しい真なる定理を出力する事例で、第2章第4.5節が切り離しておいた述語を、本章の目的に合わせて改めたものである。それらの述語のうち、真なる内容を生成することは結果にかかわり、それを知ることは知る者にかかわる。知る者については第8章が取り上げる。四つの帰属は、貢献すること、研究を遂行すること、研究能力を保有することを、第1章第3節で引いた研究のエピソードと安定した能力との違いとともに保ち、第2章第4.5節が本章へ先送りした研究者としての地位を加える。項目は、それぞれが単位について何を主張するかによって並べであり、後の項目ほど多くの証拠を要するのが普通である。ただし、それらはひとつの量の度合いではなく、どの項目も、その前の項目から帰結せず、行使された達成が積み重なる階段をなすわけでもない（もっとも、第3.2節の用法での研究者としての地位は、示された能力を前提とする）。たとえば能力の帰属は、その単位がある特定のエピソードを遂行したことを含意しない。能力は行使されないままでもありうるからである。四つの帰属を分けておく目的は、ある帰属を別の帰属から推論させないことにある。帰属はこの四つに限られない。認知的達成についての功績に値することは、地位に続いて第3.3節で取り上げ、責任は第12章第6.3節で取り上げる。自律性はさらなる項目ではない。それは、帰属された貢献、活動、あるいは能力が、選ばれた境界の外にあるものとの関係においてどのように組織されているかを修飾するものであり、それだけでそのいずれかを確立するものではない（第4節）。

これらの帰属は、先に引いた判断と交差する。第1章第3節と第2章第8.1節は、ある活動が研究であるかどうか、それが有能に遂行されたかどうか、それがその対象に到達したかどうか、それが進歩をなしたか副次的な知識を生んだかどうか、そしてそれが生んだものが重要であるかどうかを、互いに区別しておいた。それらの判断は活動を評価するものであり、本章の帰属は、活動、能力、あるいは地位をある行為者のうちに位置づけるものである。遂行の帰属は、帰属の対象となる研究活動があったことを前提とする。それを除けば、どちらの組も他方を決着させない。対象に到達しそこなった研究活動も、それ

を遂行した構成に帰属させることができる。また、第1章第3節のランダムな生成器が示したように、真なる出力は、研究を遂行してもおらず、研究する能力ももっていなかった源から生じうる。

求める達成が知識として記述されても理解として記述されても、これらの帰属は同じように当てはまり、成功はそれぞれの説のもとで評価される。

研究者は、ほかの帰属の上にある形而上学的達成として付け加えられるのではない。また、別の文脈では、同じ語が職業、制度的役割、著者性の担い手、あるいは道德的・法的地位を名指すこともある。作業的特徴づけのいかなる部分も、これらすべての用法がひとつの必要十分条件を共有することを要求しない。作業的特徴づけのもとでは、帰属は、参加については寛容であり、より強いラベルについては慎重である。人工の構成要素は貢献しうる。外部から課題を割り当てられた人工システムは、研究を遂行しうる。第3.2節で述べる、研究者としての地位の説明的価値による用法では、安定し統合された能力は、研究者を説明に役立つ記述にしうる。この地位に別の条件を立てる立場も、同節で示す。無人で進む過程は、その実行中に誰もクリックしなかったというだけで研究になるわけではないし、手厚い支援を受けている個人または集合的な単位は、その自律性が限られているというだけで、研究を遂行していないことになるわけではない。

本章は、これらの帰属を挙げた順に取り上げ、そのあとでそれらを修飾する比較へと進む。第2節は、明示された境界において研究の遂行を帰属させる根拠は何かを問う。第3節は、そこに位置づけられたものがいつ能力に達するのか、研究者がいつ説明に役立つ地位になるのか、そして功績、称賛、著者性というより厚い帰属がさらに何を要するのかを問う。第4節は、帰属されたものがどの点でなお境界の外の当事者に依存しているのかを、探究の遂行についても、探究が受け継ぐ要素についても問い、より大きな独立性の主張がプロジェクトに何を引き受けさせるのかを問う。第5節は、これらの帰属を構築と評価の負担へと変える。この順序は問いどうしの依存関係に従っている。自律性は、すでに位置づけられた境界において、すでになされた帰属について比較されるからである。これは議論の順序であって、帰属がこの順序でなされるという主張ではない。

2. 遂行：探究の組織化はどこにあるのか

第1節は、研究活動の遂行を、それへの貢献から切り離した。本節は、明示された境界において遂行を帰属させる根拠は何かを問う。答えは次のとおりである。遂行が位置づけられるのは、操作をひとつの探究の部分にしている組織化があるところであり、それは名前のついた操作が遂行されるところと同じである必要はない。そして、その組織化を探す境界は、主張に相対的に論証される。

第4〜6章の過程の説明は、普遍的な手順を立てることを避けた。同じ目に見える操作が異なる機能を果たすることがあり、同じ機能が非常に異なる操作を通じて実現されることもある（第5章第5節）。このことは、能力のリストの読み方にかかわる。システム図は、問い生成（question generation）、仮説生成（hypothesis generation）、実験（experiment）、検証（verification）と呼ばれる箱を含みながら、認識的に重要なあらゆる可能性があらかじめ選択されていることがありうる。問い生成と呼ばれる構成要素を実装したというだけでは、問いを特定し、選び、評定し、修正する能力は確立されない。逆に、システムは、ひとつの名前をもつモジュールに対応しない相互作用を通じて、問いの修正や失敗の診断を実現することもある。遂行を位置づけるのは、第1章第3節が組織された方向性を見分けるのに用いた問いである。すなわち、何がその操作を未解決の事柄に関連するものにしたのか、どのようなありうる成果であればその役割が変わったのか、そして探究の方向を転換する能力がどこにあったのか、である。

組織化（organization）は、中央の実行部や、ひとつの構成要素に含まれる機能を含意しない。それは、行為者、人工物、記録、環境、制度のあいだの関係を通じて実現されうる。帰属の問いは、どの構成が探究を維持し方向転換させるのか、そしてその構成の内部または周囲のどの境界において、その遂行や能力が記述されているのか、ということである。エドウィン・ハッチンス（Edwin Hutchins）は、人間の認知の研究に、こうした境界についての誤りを見出す。ハッチンスによれば、文化的な人工物を操作する人々からなるシステムの性質を、孤立した個人の心に帰すると、その心が持つとは限らない過程をそれに帰すことになり、人工物を操作するためにその心が実際にもたなければならない過程はどれかを問わなくなる。そして、知能の証拠を生み出したシステムの境界を誤れば、正しい性質を誤ったシステムに帰すことになりうる（Hutchins 1995, 173, 355-356）。この指摘を人工システムに当てはめると、人工の構成要素への帰属は、それを取り巻く構成が実現している能力をその構成要素に帰し、その構成要素自体が行っていることを誤って記述するとう、類比的な危険に直面すると考えられる。

ある人工システムが候補を提案し、固定されたシミュレータを呼び出し、最も高いスコアを得た結果を選択すると仮定しよう。もし人々が表象、行為空間、目的関数、シミュレータ、閾値、解釈を供給していたなら、そのシステムはそれでも真正な探索を遂行し、結果に貢献しているかもしれない。しかし、それが独立に科学的仮説を定式化しテストしたという主張は、述べられた行為者の外部にある組織化の多くを覆い隠すことになるだろう。プログラム KEKADA は、その著者たち自身の記述から、そうした外部の組織化がどこに入り込みうるかを示している。KEKADA は、クレブス（Krebs）が尿素回路を発見した仕方で、驚きに反応し、説明を定式化し、実験を遂行するものとして提示された。その著者たちは、実験の結果は利用者によって対話的に与えられたこと、そしてクレブスの理由が自分たちにとって明らかでなかったいくつかの箇所では、次にどの仮説とどの反応を研究するかを含め、利用者が選択肢のなかから選んだことを述べている（Kulkarni and Simon 1988, 2, 20, 22, 30-31）。クレブスとヘンゼライトの仕事の歴史的な再構成については第5章第4.2節を参照）。これに対して、もしシステムが、表象が関連する可能性を排除していることを認識し、新しい測定を構築または入手し、候補の空間を修正し、失敗を再解釈できるなら、探究のより多くの部分がそのシステム自身の組織化に帰属可能となる。

組織化を探す単位もまた、単一の人工物である必要はない。発見、認知、探究は、人々、モデル、装置、記録、組織にわたって分散していることがある（課題の分け方がどこで定まるかについては第4章第6節）。人工システムの行為者性についての研究は、すでにこの問いに出会っている。ファン・リール（van Lier）は、基礎研究における人工システムの行為者性を概念化するためのモデル化戦略について、四つの道筋からなる枠組みを提案する。そして、能動的なAIシステムをエージェントと呼ぶことを、何よりもこの語の説明力のゆえに擁護する。その意味は第3.2節で見る（van Lier 2023, 3, 7-8）。ファン・リール自身の適用は、構成を構成要素から切り離している。その一つは、リスト（List）とペティット（Pettit）による行為者性の基本的な説明をモデルとして用いる。この説明では、エージェントは、環境がどうなっているかを表す状態と、環境がどうあることを求めるかを指定する状態とをもち、環境がその指定に合わないときには、これらを処理して適切に介入する能力をもつ。セルフドライビング・ラボ（self-driving laboratory）の最適化アルゴリズムは、はじめの二つをもつが、それ自体では介入できず、ロボット・プラットフォームに介入させることができるだけである。そのためファン・リール

は、ラボ全体がエージェントとみなされうると示唆し、多くの人はこれをそのモデルが広すぎることの兆候と受け取るだろうと述べている (van Lier 2023, 5-6, 17-18)。本書はこうした事例から、ひとつの説明的な問いを引き出す。すなわち、その活動を、その設計者、運用者、あるいはそれを包含する制度だけにではなく、この人工の行為者に帰属させるとき、何がより明瞭に、予測可能に、あるいは評価可能になるのか、という問いである。この問いは、研究者としての地位についての確定した基準を与えるものではない。

この帰属には、人工物の機能の哲学から一つの反論が考えられる。人工物の機能を、その人工物が何のためのものかという意味で捉えると、それが何によって定まるかについて説が分かれる。意図に頼らない説の代表は二つある。一つは、過去の同じ型のものがある効果をもたらしたために選ばれ、作り続けられてきたという歴史によって機能を定める説である。もう一つは、システムのなかで構成要素が実際に果たしている因果的役割によって機能を定める説 (ロバート・カミンズ (Robert Cummins)) である。この二つの説がともに難点に出会うのが、動かない新しい試作品の事例であり、人工物の機能の哲学で諸説を試す問題としてよく論じられてきた。新しい試作品には選ばれ作り続けられてきた歴史がなく、動かないので実際に果たす役割もない。機能を認めるには、設計者の意図を持ち込むことになる。カレン・ニアンダー (Karen Neander) は、設計者が意図をもってある効果を選ぶことが、それだけで人工物の機能を定めるとする。ルース・ミリカン (Ruth Millikan) は、選択の歴史による機能を基本としながら、設計者の意図から人工物に受け継がれる機能 (派生的固有機能) も認めることで、この問題に答えようとした (Preston 2022, §2.3)。こうした説を新しい人工研究システムに当てはめると、そのシステムが何のためのものかは設計者に由来する。そこに、機能を定めた者に活動の遂行も帰属するという前提を加えると、そのシステムの探究は結局のところ設計者の研究だ、という反論が出てくる。

ここでは二つの問いを分けることができる。機能の出どころはどこかという問いと、探究の組織化がどこにあるかという問いである。本節が遂行の帰属の根拠とするのは後者、すなわち開始、監視、方向転換、回復がどこで組織されているかである。システムが何のためのものかが設計者に由来していても、そのシステムが、表象が関連する可能性を排除していることに気づき、候補の空間を改め、失敗を解釈し直すなら、探究の多くをそのシステム自身の組織化に帰属させうる。そうする能力そのものが設計によって組み込まれたことは、能力がどこから来たかにかかわるのであって、探究の組織化がいまどこにあるかにかかわるのではない。そうした来歴がそれ自体として自律性にかかわるかどうかは別の問いであり、争われている (第4.3節)。逆に、システムが自らの目標をもつように見えても、方向転換が外から供給されているなら、方向転換を組織する働きはその外にある。また、探究に失敗したシステムも、動いて構成要素が役割を果たしている限り、役割としての機能をもつ点で、動かない試作品とは異なる (第1章第3節)。

クラーク (Clark) とコスロウィ (Khosrowi) は、単一の発見者の特権化するのではなく、科学的発見についての集合体中心の分析を擁護している。二人によれば、行為者中心の見方は、発見者であることを全か無かの事柄とし、それを、見いだしたものを解釈しその意義を認識するといった能力に結びつける。そのため、AlphaFold の貢献のように、たんなる道具の貢献を超えながら、その十全な意味での発見には届かない寄与を位置づけられない。人間による発見も同じ難点を生む。これに対してクラークとコスロウィの見方では、発見するのは全体としての集合体であり、功績は各貢献の性質と重要性に応じて配分される (Clark and Khosrowi 2022, 4, 8-11)。しかし、第2章第6節が、マシュー・J・ブラウン (Matthew J. Brown) とウグン・トゥンチ (Uygun Tunç) の伝える、大規模な実験を分散認知システムとして扱うギアリー (Giere) の議論に即して述べたように、貢献が分散していることは、知ることも分散していることを自動的に確立するものではない (M. J. Brown 2015; Uygun Tunç 2021, 3-4)。また、研究を行う集合体のあらゆる構成要素が、それによって研究者になるわけでもない。それゆえ、本章で**行為者** (actor) という語を用いるとき、それは行為者が個人であることをあらかじめ前提するものではなく、共同で遂行された探究を、特権的な一成員の仕事として記述し直す必要もない。境界は分析上の変数であり、モデルの物理的境界やプロジェクトの組織的境界から推論されるのではなく、帰属される遂行や能力との関係において論証される。同時に、広いシステムが因果的に必要なあらゆる貢献を含んでいるという事実は、それだけで**研究者**をそのシステム全体にとって説明に役立つ地位にするわけではない。

結論は、ひとつの人型機械があらゆる低水準の行為を遂行しなければならない、ということではない。人間の研究はすでに、専門家、技術者、施設、ソフトウェア、装置、供給業者、制度のあいだに作業を分散させている。たとえば、ベル研究所の拡散型シリコン・トランジスタを報告した1956年の論文は、材料の供給、素子の作製と実装、電気的な評価、計算を、それぞれ別の人々に帰している (Tanenbaum and Thomas 1956, 22)。人工研究者も同様に、専門化されたロボットを呼び出し、遠隔の実験室を利用し、人間の作業を委託し、あるいは自らの環境を再組織しうる。問題は、どの組織化された全体が記述されているのか、そしてどの依存関係がその外部にとどまっているのか、である。ここでの機能は、第5章第5節の意味で、作業が探究のなかで果たす役割であり、上で論じた、人工物が何のためのものかという意味での機能のことではない。ある機能を境界の向こう側へ移すことは、主張が誰について述べているかを変えるのであって、その機能の必要を消失させるわけではない。ただし、境界の向こう側にあるものがその機能を遂行するように組織されていなければ、その機能は遂行されないままになりうる (第6章第4.3節で触れたベルの学位論文の指摘を参照)。

研究の記録は、作業の分布をいつも見えるようにするわけではなく、記録が作業を帰属させる粒度は、記録の層ごとに変わりうる。創業プロジェクト Open Source Antibioticsが論文とともに寄託した実験ノート (第5章第2節) が、それを示す。ここでは、ある精製についてクロマトグラフィー装置が生成した報告が、それを実行した化学者の名を記し、その実行が定組成 (溶媒の組成を一定にしたまま) に保たれたこと、勾配が変更されたこと、使用者によって一時停止され終了されたことを、時刻つきで記録している。ノートの頁は同じ化学者を各記入の最終更新者として記し、ノートの静的な書き出しは論文の責任著者を利用者として示す。そして補足情報は、精製については、その化合物が「方法1、次いで方法2で精製された」(purified by methods 1, then 2) と述べるだけである (Tse 2023, EGT 522-2; Klug et al. 2023, SI: Chemical Methods S25)。同じ寄託の別の頁に添付された質量分析の報告は、それを作成したサービスの分析者の名を記しているが、その人は論文の著者にも謝辞にも現れない (Tse 2023, DMK204-1; Klug et al. 2023)。それぞれの層は作業の異なる部分を名指しするものにしており、そのどれかに名がないことは、その人が関与しなかったことの証拠にならない。記録から論じられる境界は、用いた記録の粒度を引き継ぐ。人工の構成では、この粒度が部分的には構築の問題になる。どの決定の要所をどれだけ細かく記録するかは、その構成がどう作られるかによって定まり (第4.2節で述べる決定の要所の記録を参照)、構成が自らの記録をつけるところでは、遂行が帰属される単位そのものが、その帰属の根拠となる証拠の一部を供給する。

それゆえ本書では、境界を説明上の目的のために選び、境界の選び方だけで主張が真になることのないようにする。もし**そのAIシステム** (the AI system) が、その人間の設計者、装置技術者、データ・キュレーター、査読者、出版委員会を含むのであれば、そうした寄与者をもつ現代のどのプロジェクトも自律的なハイブリッドとして記述し直すことができる。もし境界が基盤モデルだけを含むのであれば、分析は、結果を可能にしている検索、ツール、チェッ

カー、オーケストレーション、環境からのフィードバック、制度による受け止めを覆い隠すかもしれない。擁護可能な境界は、重要な作業の分布を吸収するのではなく、明らかにする。

検索拡張型のシステムは、この問題を具体的にする。スマート（Smart）、クロウズ（Clowes）、クラーク（Clark）のDigital Andyは、その内容が回答に影響を与える外部のベクトルストアに依拠していたが、この論文は、LLMを拡張された認知システムの主体とみなすことに否定的な判定を下した。その主な理由は、モデルが検索の回路を開始することも協調させることもしていなかったことである。ある注は、この考慮がどれだけの重みをもつかについて三人の著者全員が同意しているわけではないと述べ、見解のその部分を筆頭著者のものとしている（Smart, Clowes, and Clark 2025, 3, 21–23 and n. 27）。この事例は、どの関係が研究行為者を構成するのかに決着をつけるものではない。それが示すのは、因果的寄与と資源への依存がそれだけでは境界を選び出せない理由である。開始、協調、統合、そして反事実的な調整、すなわち状況が違えばそれに応じて違ってくるような調整は、そのいずれも研究の普遍的条件とされない場合でさえ、帰属にとって重要でありうる。反事実的な調整は、たとえば、スマートらが取り上げる能動的な検索拡張型のシステムに見られる。そこではモデルが、自らの記憶では足りないと感じたときにだけ外部の記憶を呼び出す（Smart, Clowes, and Clark 2025, 23–24）。

したがって境界は主張相対的であり、ハードウェアやソフトウェアの目録に変化がなくても変わりうる。同じモデル、検索器、装置、人々が、あるエピソードでは人間の行為者によって使われる道具として、別のエピソードではハイブリッドな行為者の部分として、さらに第三のエピソードでは、人工の部分により大きい行為者によって協調される資源として機能することがある。変わるのは、単にその集合体に割り当てられる名前だけではなく、開始、監視、方向転換、回復がどこで組織されているか、である。この点で遂行の帰属は、関係的かつ組織的な帰属であって、モデルの恒久的な型水準の性質ではない。それが研究者としての地位に何を意味するかは、第3.2節で取り上げる。

3. 能力、地位、功績

第2節は、活動の組織化をある境界に位置づけた。本節は、そこに位置づけられたものが幸運なエピソードではなく能力に達するのはいつか（第3.1節）、研究者がその単位にとって説明に役立つ地位となるのはいつか（第3.2節）、そして功績、称賞、著者性というより厚い帰属がさらに何を要するのか（第3.3節）を問う。第一の問いに答えるにあたって、本節は、第1章第3節が本章へ先送りした、安定した能力の条件を取り上げる。

3.1 能力

エピソードから能力への移行を示す単一の振る舞いは存在しない。第1章第3節は、能力の帰属を支持しうるものとして、関連する変動にわたる信頼性、変化する証拠への応答性、研究機能の統合、失敗後の修正を挙げた。この四つは、ひとつの説から引いたものではなく、エピソードを能力から分けるためにまとめたものである。成功が遂行者の有能力さのおかげかを問う、ソーサ（Sosa）の説（第3.3節）のような有能力さの説は、その帰属をそれ自身の語で述べる。先の四つを、ある境界における帰属のために具体化すると、たとえば次のような考慮事項が得られる。どれも四つのうちのひとつ以上にかかわる。

- 行為者が、供給された手順の系列をそのまま実行するだけでなく、実質的に異なる行為のあいだで選択しているかどうか。
- 証拠、異常、コスト、失敗した操作に対して、探究に関係する仕方と異なる応答をするかどうか。
- 問い、依存関係、理由を、時間を通じて保持し修正しうるかどうか。
- 焦点となる仮説を自動的に修正するのではなく、外部の構成要素が失敗したときにそれを認識しうるかどうか（第6章第2.1節、第4.2節を参照）。
- 帰属された組織化が、偶然の相関や見慣れた課題の形式を取り除く変更を経ても存続するかどうか。
- 成功と失敗を、開示されていない人間によるフィルタではなく、行為者の操作へと遡って辿りうるかどうか（第2節の境界の問い）。

これらは、あらゆる研究エピソードの必要条件ではない。単純だがよく設計された操作は、これらすべてを満たすことなく研究の一部でありうる。これらが関連性をもつのは、あるプロジェクトが、より広い、開かれた、あるいはますます自己主導的な探究を遂行する安定した能力を主張するときである。

したがって能力は、無制限の一般的な力になることなく、ひとつのエピソードを越えて及ぶ。あるシステムが研究能力をもつと言うことは、何らかの条件の範囲にわたる傾向性を帰属させることであり、能力の主張は、その範囲が明示されてはじめて評価できる。システムは、準備された計算環境の内部で、化学合成のひとつのクラスの内部で、あるいは装置と消耗品が他者によって維持されている実験室の内部で、能力をもつことがありうる。そのような能力は実在的である。しかしそれらは、それだけでは、環境を再構築し、入手できない材料を獲得し、装置を再設計し、あるいは異なる証拠的実践をもつ分野へと移行する能力を確立するものではない。維持された環境に相対的な能力は、その維持が続くことにも指標づけられる。維持が続くか、変わるか、止まるかは、能力が帰属されたシステムの外で決まり、そのシステムに何の変化がなくても、能力の帰属の前提となった条件は成り立たなくなりうる。維持する者が、たとえばどの装置、データ、計算資源を利用可能にするかも選ぶところでは、能力をたんに可能にしているだけの依存が、維持する者が帰属された組織化の一部（たとえば選択）を供給する依存、すなわち第4.2節で区別する指示的な依存へと移りうる。構成は更新されることも、自らを変更することもあるので、能力は同様に、その証拠が集められたバージョンと期間について帰属される。第9章は、構成を維持し、更新し、あるいは引退させる者を、研究を取り巻く役割の一つとする（表9.1の提供者または維持者）。

第3章第9節の予測の例、すなわち明示された範囲である性質を信頼できる仕方と予測するが、その性質を生む仕組みを利用者が説明できないシステムの例は、この違いを具体的にする。予測と、その信頼性を支える証拠を固定しよう。一方の場合には、条件が変わるたびに人間が失敗を診断し、問いを組み直し、新しい方法を選び、その後人工の構成要素が予測を再開する。他方の場合には、選ばれた人工の構成がそれらの働きを果たし、関連する変動にわたってそれを行える証拠がある。局所的な予測という達成は、両者で同じでありうる。しかし、その探究を再組織する能力の証拠は、異なるところに位置する。ここで組み立てた対照は、第一の場合にはその予測を割り引かず狭い帰属を支え、第二の場合には無制限の能力や研究者としての地位を確立することなく、範囲を限ったより広い帰属を支える。

3.2 研究者としての地位

研究者としての地位は、能力以上のものを求める。それがさらに何を求めるかについては、説によって見解が分かれる。本書は**研究者としての地位**を、このラベルがその説明的価値によって正当化されるとする立場の意味で用いる。すなわち、**研究者**がある単位にとって説明に役立つのは、その能力が十分に広く、統合され、持続的であるために、それを研究者と呼ぶことが、それが探究を保持し、行為のあいだで選択し、抵抗に応答し、失敗を診断し、表象を修正し、関連する作業を開始しうることを予測するのに役立つときである（第13章第5節を参照）。この用法は、定義ではなく、機能的で覆されうるものである。この用法は二つの源泉に依拠する。ファン・リールからは、その正当化の形、すなわち語をその説明力によって正当化するという形を受け継ぐ。能動的なAIシステムをエージェントと呼ぶ理由のなかで、ファン・リールはこの説明力をとくに強調する。ただし、そこでいう説明力は、科学者がこれから作ろうとしているものに向けられている。この語は、そうしたシステムを道具や技術と呼ぶよりも、科学者が作ろうとしているシステムの種類をよく言い表し、その取り組みがどこへ向かっているかを示す、というのである（第2節、[van Lier 2023, 3, 7-8](#)）。ここでの用法は、この形を受け継ぐが、説明力を、これから作られるシステムを言い表す力としてではなく、ある単位がいま何をしうるかを予測するのに役立つ力として捉える。デネット（Dennett）からは、記述とそれが支える予測との結びつきを受け継ぐ。デネットの説では、あるシステムが志向的システムであるのは、その振る舞いを説明し予測しようとする者の戦略との関係においてだけであり、それに対して志向的構え、すなわちそれを信念や欲求をもつ合理的な行為者とみなしてその振る舞いを予測する構えを採ることは、実用的な決定である。チェスをするコンピュータに対して、人は、対戦相手としてはその構えを、設計し直す者としては設計的構えを、修理する者としては物理的構えを採りうる（[Dennett 1971, 87, 91](#)）。この用法は、それが前提するものにおいてデネットの説と異なる。その説では、志向的構えからの予測の成功は、システムの設計についての特定の像に依拠しないので、そうした像のどれも確証も反証もできない（[Dennett 1971, 97](#)）。これに対して、ここでの用法は、その組織化がある境界に位置づけられ、明示された範囲にわたって示された能力を前提する（第2節・第3.1節）。説明的価値は、その帰属に加えて量られるのであって、その代わりに量られるのではない。

では、能力の帰属に加えて、**研究者**という記述は何を見通させるのか。個々の能力の帰属は、それぞれが試された範囲での振る舞いを予測する。この記述がそれ以上の働きをもつとすれば、それは、どの能力の試験も想定していなかった状況で、それらの能力がどう組み合わせられるかを予測させるときである。たとえば、用意された課題が行き詰まったときに、その単位が止まって指示を待つのではなく、関連する問いを保持したまま別の方法を組み立て、そのために必要な作業を自ら始めると予測できるなら、この記述は、能力の一覧からは出てこない予測を与える。その予測が正しければ、その単位は、第4.4節で見えるように、自らの選択によって、誰も用意していなかった設備、データ、協力への要求に出会うことがありうる。そのため、この記述は、その単位がどこで新しい依存を生みうるかを見越す手がかりにもなりうる。ただし、こうした予測を支えるには、第3.1節で触れた研究機能の統合が、個々の能力の試験が想定しない状況でも働くことの証拠が要る。たとえば、第4.2節で述べる反事実的テスト（指導の撤回、指示と証拠の衝突、慣れ親しんだツールの故障など）への応答である。その証拠がないところでは、この記述は、能力のまとまりにつけた名前以上のものではない。第2節のKEKADAは、この条件がどこで問われるかを示す一例になる。著者たちにクレプスの理由が明らかでなかった箇所では、次にどの仮説と反応を研究するかを利用者が選択肢のなかから選んでいたので、能力をどう組み合わせる先へ進むかが問われるところで、組織化は外から供給されていた（第4.2節）。

説明的価値による用法では、デネットの例における構えと同じく、説明的価値は、その記述が誰にとって、どのような目的のために説明するのに相対的である。ある実験室にとって、自らの構成が何をすることを見越すのに役立つラベルも、その構成の結果に依拠するかどうかを決めようとしている共同体には、ほとんど何も説明しないかもしれない。この相対性は帰属を恣意的なものにはしないが、研究者としての地位の主張は、明示された目的と受け手について評価される。そのため、能力について同じ情報をもつ二つの受け手が地位の判定で分かれるとき、その不一致は、能力の有無をめぐるものではなく、記述を何のために用いるかの違いでありうる。情報の不足から生じる不一致とは、分けて扱うことができる。ファン・リールが適用する行為者性のモデル、すなわちセルフドライビング・ラボをその全体として数えることになると述べるモデル（第2節）と同じく、この種の用法は、ラベルを広い構成にまで広げうる。そのとき、第2節の境界の問いが仕事の一部を担う。

別の条件を立てる立場は、異なる判定を下す。研究者としての地位が、その単位に帰属する発見をなしうることを含むと仮定してみよう。研究には、対象に到達しない活動や、発見ではなく副次的な知識を生む活動も含まれるので（第1節）、これは地位を発見者としての地位に寄せる仮定である。この仮定のもとで、ベルガマス・ガナピーニ（Bergamaschi Ganapini）の発見の条件、すなわち、見いだされたものが関連分野での発見者の有能さを表していること、そして発見者がその有能さについて反省的な洞察をもっていること（第3.3節）を研究者としての地位に当てはめれば、その地位は、システムが関連分野での有能さと、それについての反省的な洞察を備えているかどうか、たとえばその出力がその分野の用いる種類の推論から生じているかどうかにかかることになるだろう。ベルガマス・ガナピーニ自身は、現在の深層学習システムはまだその条件を満たすようには見えないと論じる。ヴァチェック（Vacek）の見方では、達成はその担い手がそれについて称賛に値することを含意する。ヴァチェックは、AIシステムは称賛に値しないという強い前提を擁護するが、AlphaGoについての判定には、人間との協働のなかで生じた結果の全体については称賛に値しない、という弱い主張で足りるとする（第3.3節）。強い前提をとれば、達成を含意する研究者としての地位は、その能力がどれほど広く、統合され、持続的であっても、人工システムには得られないことになる。その根拠の一つは、能力がどのように獲得されたか、すなわち自律的な行為者への成長ではなく工学の結果であることにあり、ヴァチェックの示し方では、これはシステムがどのような能力をもつかにはかわらない。他の根拠、すなわち人間の作った訓練データへの依存と、協働を始め監督する人間が称賛に値することは、人間の協働のなかで働く現在のシステムについてのものである。また、この工学という根拠は、ヴァチェックがハクリとマケラから受け継ぐもので（第3.3節）、第4.3節で見える二人の議論では、つくられた時点で組み込まれた目標をシステムが自ら捨てられないことに依る。自らの最終目標を改めうるシステム（第4.1節）には、この点は当てはまらないだろう。これらの論者は誰もそうしたシステムを考察しておらず、強い前提をそれに当てはめるには別の論証が要る。弱い主張だけなら、判定は協働がどう組まれているかに依存し、協働のなかで担った部分の達成や、協働する集合体への帰属は、別に問われることになる。いずれの場合も、問いは、その地位がよりどころとする称賛と責任の実践へと移るだろう。

バランディアラン（Barandiaran）、ディ・パオロ（Di Paolo）、ローデ（Rohde）が提案する最小行為者性（minimal agency）の説では、あるシステムが行為者であるのは、三つの条件を満たすときである。一つは個性（individuality）であり、そのシステムが自らの同一性を定義し、それによって自らを環境から区別することである。二つ目は相互作用の非対称性（interactional asymmetry）であり、そのシステムが自らの活動の能動的な源泉であり、その環境との結合（coupling）を変調することである。三つ目は規範性（normativity）であり、そのシステムが、観察者や設計者がそれを判断する規範では

なく、内から生成され維持される規範、あるいはそれが内面化した規範に関連して、この活動を調整することである (Barandiaran, Di Paolo, and Rohde 2009, §2)。三人は、ロボットがふつう行為者と呼ばれるのは、その機械的な構造が部品を一つにまとめており、その性能が、観察者や設計者が有用と判断する基準を満たすからにすぎないとし、外から課された規範を満たすだけのシステム、たとえば外から固定された関数に従う最適化を、行為者性のモデルとして扱わない (Barandiaran, Di Paolo, and Rohde 2009, §§2.1, 6)。研究者としての地位がこの意味での行為者性を前提するとすれば、その地位は上の用法と二つの点で異なることになるだろう。第一に、単位は、主張に相対的に選ばれる単位 (第2節) ではなく、システムが自ら定義する個体になるだろう。第二に、構成が外から課された規範を満たすだけなのか、それとも自ら生成し維持する規範、あるいは内面化した規範によって活動を調整しているのか、地位そのものにかかわるだろう。これに対して本章は、探究の目的と基準の起源が外にあることを、研究にではなく自律性にかかわるものとして扱う (第1節・第4.3節)。バランディアランらは行為者性を生きたシステムに限らない (Barandiaran, Di Paolo, and Rohde 2009, §§3–4)。しかしその説では、外から固定された目的を最適化するだけの構成は、それを研究者と呼ぶことの説明的価値がどうであれ、その規範性の条件を満たさない。現在の構成がそれ以上のことをしているかどうかは、それ自体が問われており、ディ・パオロが論じるように (第8章第2.2節)、システムの応答を観察するだけでは決まらない。三人は実験室を論じていないが、第2節で見たファン・リールのセルフドライビング・ラボ全体に当てはめると、二つの用法の違いが見える。そうしたラボが、外から与えられた目的のもとで方法を組み替え、失敗に応答するとしよう。説明的価値による用法では、その統合された働きが研究者という記述を支えるかどうかが問われる。最小行為者性の説を前提すれば、外からまとめられた単位であるというだけでは足りず、そのラボが自らの同一性と活動の規範をどう生成し維持しているかが問われる。両者は、同じ対象に異なる点数を付けているだけではなく、何を一つの行為者として数えるかでも分かれている。理解がその地位の根拠として持ち出されるところでは、その持ち出しを評価するための区別を、第3章第6節と第9節が与える。

3.3 功績、称賛、著者性

より厚い帰属は、ある成功について行為者が何に値するかにかかわる。第2章第4.2節は、正しい結果と、適切な成功、すなわち遂行者の有能さのおかげである成功とを分けるソーサの区別を導入した。人工システムがこの種の成功を自らの達成として得られるかどうかは争われている。以下で見る論争は研究ではなく対局をめぐるものだが、そこで争われる達成の説は一般的なものであり、研究の成功がシステム自身の達成とみなされるためにも、同じ条件が問われる。キーヴァル (Kieval) は、2016年にプロ棋士イ・セドル (Lee Sedol) を破った囲碁システムAlphaGoについて、AlphaGo自身がその勝利の功績に値すると論じる。この主張の動機として、イ・セドルを破ることが、それができるシステムを作ることとは別の達成であり、しかもシステムを作った人々には個人としても集団としても成しえない達成であることを挙げる (Kieval 2024, 32)。キーヴァルは達成の説をブラッドフォード (Bradford) から受け継ぐ。その説では、達成とは、困難な過程、すなわち十分な程度の強い努力を要する過程が、その産物を有能に引き起こすことである (Kieval 2024, 33)。ただしキーヴァルは、有能な行為者は自分が何をしているかを知っていなければならないというブラッドフォードの要件に代えて、有能さについての、より知性主義的でない二つの読みを用いる。一つの読みでは、有能さとは、命題的な知識によってではなく、蓄えられた行為のパターンによって担われる、やり方を知っていることである。もう一つの読みは、フォン・クリークシュタイン (von Kriegstein) が徳認識論から引き出すものである。徳認識論とは、知識を、知る者の能力のおかげである成功として説明する立場である。この読みでは、有能さは反運条件、すなわちたんなる運による成功を排除する条件である。つまり、行為者が目標に有能に到達するのは、その行為者が現実と同じように行為する近傍の可能性の大半でもその目標が実現し、かつ、自分の計画によって成功する見込みが高いとみなすことが正当化されている場合である (Kieval 2024, 34–35)。キーヴァルは、AlphaGoがどちらの読みも満たすと論じる。この説明では、AlphaGoは目標状態をもち、盤面を表象し、候補手を評価し、学習を通じて、勝ちにつながりそうな手へと選択を絞り込む内部表象を獲得する。また、熟達した相手に対してさえ、そうした近傍の可能性の大半で勝つ。さらに、計画の評価は、各候補手が勝利に至る見込みを見積もる構成要素である価値ネットワークが与える (Kieval 2024, 37)。有能さには行為者性が必要だという反論に対しては、キーヴァルは二つの行為者性を区別する。一つは道徳的行為者性、すなわち自らを省みて規範的に推論する能力である。これをもつ行為者は、自らの理由について問われれば答える立場にあり、責任を問われる。AIシステムは明らかにこれを欠くと、キーヴァルは述べる。もう一つは意図的行為者性、すなわち目標に向けて行為する能力であり、キーヴァルの説では、有能さにはこれで足りる (Kieval 2024, 37–38)。ただし、道徳的行為者性の欠如は自明のこととして述べられているだけで、どのようなAIシステムもこれをもちえないと論じられているわけではない。困難さについては、キーヴァルは、努力がどう感じられるかを、要件ではなく概念への手がかりとみなす。そのうえで、フォン・クリークシュタインに従い、努力することを、課題に内部の資源を振り向けることと捉える。そうすると、AlphaGoの訓練と対局が要する計算は努力に数えられる。またキーヴァルは、ハードウェアをシステムの内部に含める。資源を費やすなんらかのハードウェアがなければ、そもそもシステムが存在しないからである (Kieval 2024, 38–39)。キーヴァルは、自ら奇妙と呼ぶ帰結も受け入れる。ブラッドフォードは、達成の内在的価値、すなわち産物の価値とは別に達成がもつ価値を、意志と合理性という人間に特有の能力に求める。キーヴァルは、AIシステムはこれらの能力を発揮しないと言う。AIシステムは人間ではなく、そうした生得の能力をもたず、設計されたこと、すなわち目的関数を満たすこととしかしないからである。そのため、AIシステムの達成は、その価値をAIシステムに固有の能力に基づかせることができないかぎり、価値のない達成でありうる。その候補としてキーヴァルは、AIシステムの合理的な推論の能力や創造的な洞察を挙げ、とくに合理性について、GPT-4 のような大規模言語モデルが推論の能力を発達させたように見えることに触れている (Kieval 2024, 33, 39–40)。

ヴァチェックは、イ・セドルを破ったことはAlphaGoの達成ではないと応じる (Vacek 2025, 690)。その議論には二つの前提がある。第一の前提は、何かがある行為者の達成であるならば、その行為者はそれを達成したことについて称賛に値する、というものである。ここでの称賛に値することは、道徳的なものに限られない。たとえば、知的な点で称賛に値することもありうるからである。ヴァチェックはこの前提を二つの仕方で支える。一つは、功績に値するというキーヴァル自身の言い方を、AlphaGoが称賛に値すると言うことの別の言い方として読むことである。もう一つは、ブラッドフォードの達成の説と、称賛に値することについての説とのあいだに並行関係をたどることである。どちらの説も、無理なく期待できることに相対的な困難さの閾値を越えることを求め、どちらも感嘆のような態度を求める。ヴァチェックの見方では、何かがある達成であると言うことは、その人がそれに責任をもつと言うことであり、ここで関係する責任の形が、称賛に値することである (Vacek 2025, 692–693)。第二の前提は、AIシステムは称賛に値しないというものである。ヴァチェックはこの前提を、AIシステムの責任に反対して出されてきた議論を称賛にも及ぼすことによって擁護する。そうした議論は、主に非難を扱ってきた。ヴァチェックはハクリ (Hakli) とマケラ (Mäkelä) から、次の主張を受け継ぐ。システムが道徳的行為者性に必要な能力をすべてもっていたとしても、その能力が自律的な行為者への成長ではなく工学の結果であるなら、私たちにそのシステムに責任を帰す資格がない。二人の来歴を考慮する議論

は、第4.3節で取り上げる。さらにヴァチェックは、学習システムのよい出力が、訓練データを作った人々に大きく依存していると付け加える。また、ナイホルム (Nyholm) の議論を組み替えて、機械が、人間がしたなら称賛に値するようなことをするとき、それについて称賛に値するのは、人間と機械の協働を始め、監督し、管理する人間であるとする (Vacek 2025, 693–695)。ヴァチェックの指摘では、その結論には、AIシステムにとって称賛がまったく不可能であることまでは要らない。システムが協働の一部であり、その行為者性が他者に従うものであれば、そのシステムは結果の全体については称賛に値しない。そして、AlphaGoについての判定にはそれで足りる (Vacek 2025, 695)。もっともヴァチェックは、AIシステムがいかなる形の責任も負えないとは主張しない。また、達成と称賛に値することが見かけどおりに緊密に結びついているかどうかは、まだ確かめられていないと述べる。より薄い種類の達成に訴えることを、ヴァチェックは、ほかの論者には使えるがキーヴァルには使えない手とみなす。より広い意味の達成は称賛に値することを含意しないかもしれないが、キーヴァルが依拠するブラッドフォードは、その説が扱う達成を、たんなるありふれた成果からすでに区別しているからである。こうしてキーヴァルには称賛に値することを含意する意味の達成しか残らないので、ヴァチェックの議論が正しければ、AlphaGoに達成の全体を帰することはできない。そして、個々の人間にも達成の全体を帰せられないのであれば、より個人主義的でない行為者性と責任の概念に目を向けることが前に進む道になるだろう、とヴァチェックは示唆する (Vacek 2025, 696 and n. 2)。

第1章の**認知的達成**は、知識や理解のように探究が目指す認知的な善を指す語であり、ここで争われている、功績を含む遂行としての**達成** (achievement) ではない。そのため、キーヴァルとヴァチェックの不一致は、研究が遂行されたかどうかではなく、その成功が誰の功績になるかにかかわる。主張される成功が担い手の有能さに帰せられる必要があるかは、その達成についての説による (知識については第2章第4.2節)。研究を遂行する行為者が、その担い手である必要もない。功績の主張が何についてのものでありうるかについて、本書は四つの答えを区別し、それぞれを別の区分として扱う。すなわち、真なる出力、信頼できる過程、有能さに帰属可能な成功、称賛に値する行為者である。最初の三つはおおよそ、遂行が正しいこと、巧みであること、巧みさのゆえに正しいこと (適切さ) というソーサの区別 (Sosa 2015, 86) に従い、四つ目は、ヴァチェックの議論がかかっている、称賛に値することを名指す。これらの区分に照らせば、キーヴァルの条件は、困難さを加えたうえで、有能さに帰属可能な成功の区分にかかわる。これに対してヴァチェックは、キーヴァルが依拠する意味での達成には、称賛に値する行為者の区分も要すると論じる。これらの区分を分けておくことは、正しい出力がシステムの功績として読まれるのを防ぐ。ソーサ自身の枠組みが、すでにそのいくつかを分けている。その枠組みは遂行の評価を、理由をもって行為する行為者を越えて、狙いをもちながらそれに動機づけられてはいない装置にまで広げる。ソーサの説では、サーモスタットはうまく働きうるし、部屋を暖かく保つことについてある種の功績を認められる (Sosa 2015, 85 and n. 2, 93 n. 8。動物的知識と反省的知識については第3章第5.3節を参照)。その枠組みはこうして、装置にまで及ぶほど薄い功績の意味を与える。これを踏まえると、キーヴァルとヴァチェックの論争は、達成がどのようなさらなる条件を要するかをめぐる論争として読むことができる。本節の残りでは、まず四つ目の区分にかかわる、称賛の語としての発見と創造性を取り上げる。続いて、それとは種類の異なる二つの帰属を取り上げる。一つは、責任の根拠となりうる仕方でも主張を保持することであり、もう一つは著者として立つことである。著者として立つことは、結果についての功績ではなく、結果に対する説明責任にかかわる。

ベルガマス・ガナビーニは、この枠組みを発見に当てはめて、次のように論じる。何かを見いだしたことが科学的発見とみなされるのは、それが関連分野での発見者の有能さを表している場合に限られる。たとえば、その分野が用いる種類の推論から部分的に生じた場合や、発見者がその科学的意義を説明できる場合である。さらに、発見者がその有能さについて反省的な洞察をもっている場合に限られる。この洞察は、ソーサのメタ適切性 (meta-aptness) をモデルにしている。射手の一射がメタ適切であるのは、その射手が有能さによって的を射るだけでなく、距離や風といったその場の条件のもとでその一射が成功することを見てとっている場合である。この説では、発見はたんなる功績ではなく称賛に値する達成であり、称賛には、結果に至る過程を行為者がある程度制御できることが必要である。そのため、こうした洞察を求めるが、それが意識や明示的なメタ表象を要することは否定する (Bergamaschi Ganapini 2025, §§1–2)。ベルガマス・ガナビーニは、現在の深層学習システムは、少なくとも今のところ、これらの条件を満たすようには見えないと論じる。第一に、信頼できる実績は有能さを立証しない。実績は、モデルが何に依拠しているかを明らかにしないからである。そのため、この見方では、その出力がその分野の用いる種類の推論から生じているかどうかについての判断は、少なくとも保留せざるをえない。第二に、AlphaFold の構造予測の正当化は、大部分が人間の仕事のままである。第三に、そうしたモデルが出力する確率は、その不確実性を正確に表していないことがある (Bergamaschi Ganapini 2025, §§4–5)。ベルガマス・ガナビーニは条件を主体または集団について述べるが、単一のシステムが資格をもつかを問うており、第2節で取り上げた発見の分散的な見方を、否定するのではなく脇に置いている。その説は発見を称賛に値する達成として扱うので、上の四つの区分に照らせば、発見を称賛に値する行為者の区分に置く。

デュエデとフリードマンは、これらの条件に直接応答している (序論第2節を参照)。二人は、有能さと反省的な洞察の条件が、明らかな人間の事例を誤って分類すると論じる。新しい電波望遠鏡の記録を見守っていた博士課程二年目のジョスリン・ベル・パーネルは、のちにパルサーからのものと判明する周期的な信号に気づいた。そのとき、その意義を説明することも、それを天体物理学のモデルと関係づけることもできなかった。それでも、デュエデとフリードマンの見方では、ベル・パーネルはパルサーを発見したのである。その説では、ベル・パーネルと、偶然に何か重要なものに行き当たった者とを分けるのは、有能さが発揮されたかどうかではない。偶然に行き当たった者とは、たとえば土地を掘っていて兵馬俑を掘り当てた農民や、ベルガマス・ガナビーニ自身の例でいえば、穴に落ちてネロの埋もれた宮殿に行き当たった子どもである。両者を分けるのは、見いだした者の働きが認知的ギャップを露わにすること、あるいはそれを埋めることに成功したかどうかである。二人は発見と称賛との結びつきを保ち、発見者が功績と称賛に値するのはそうしたギャップを乗り越えたからだとする。しかし、その称賛にベルガマス・ガナビーニの挙げた能力が必要であることは否定する (Duède and Friedman 2025, §§2.1–2.2, 4)。四つの区分に照らして読むと、二人の説は、発見を称賛の語として保ちながら、有能さに帰属可能な成功の区分には置かない。またデュエデとフリードマンは、発見に科学共同体でのしかるべき立場も要るのであれば、現在のシステムはたいていそれを欠くと認める。その場合、そうした立場が人工システムに及ぶかどうかは、能力ではなく承認の実践にかかってくる (Duède and Friedman 2025, §5)。

この種の述語は発見だけではない。第5章第2節は、**創造的** (creative) が産物、過程、あるいは人物に帰属されうることに触れた。ポール (Paul) とストークス (Stokes) は、二人が概観する四つの過程条件のうち、これまでに最も多くの哲学者が明示的に支持しているように見えるのは行為者性であると報告している。ただし、そうした哲学者もひと握りにすぎない (Paul and Stokes 2023, §2.2)。行為者性の条件によれば、ある項目が創造的であるのは、行為者がそれを産出した場合に限られ、いくつかの定式化では、行為者がそれを意図的に産出した場合に限られる。ポールとストークスはこの直観を、**創造**

的が称賛の語であり、行為者によってなされたのでないものや、何らかの意味で意図的になされたのでないものには称賛が与えられない、という事実にとどっている (Paul and Stokes 2023, §2.2.4)。二人はコンピュータについての帰結を引き出す。最近のいくつかのシステムは、驚き、独創性、自発性の条件をもっともらしく満たすが、そうしたシステムはおそらくまだ称賛や非難の対象ではない。したがって、創造性が行為者性を要するなら、どのシステムもまだ本当には創造的ではないと論じうる (Paul and Stokes 2023, §6)。そうした説を人工システムやハイブリッドな組織による産物に当てはめると、それが創造的であるかどうかは、その産物が帰属される境界の内にあるものが称賛されうるかどうかにかかってくる。これは、ヴァチェックのキーヴァルへの応答がかかっている問いである。

主張を提案し、擁護し、修正する構成が、責任の根拠となりうる仕方です。それらの主張を保持してもいるかどうかは、さらなる問いであり、これについては信念への責任の説明が分かれる。哲学者ソフィー・キーリング (Sophie Keeling) の説では、そうした責任の根拠となる行為者性には、自らの推論を p について心を決めることとして主体が意識していることが含まれる。その行為者性を自己帰属しうる立場にあることは必要だが十分ではなく、その自己帰属は、自らのふるまいや他人の証言から推論されるのではなく、一人称的に根拠づけられなければならない (Keeling 2023, 14–15, 19)。これを人工システムに移すなら、ある主張を再検討しているというシステムの報告は、それだけではキーリングの条件が満たされていることを示さない。マクヒュー (McHugh) は、自らの推論を心を決めることとして意識していることまでは求めないが、理由に応じることは「少なくとも傾向的には、意識的な行為者性を含む」とし、それゆえ、理由に沿って信念が変わるだけで、それに対してまったく受動的なままである生き物は、どの信念も理由のために持つてはいないとする (McHugh 2013, §§3.1–3.2)。そうだとすれば、証拠を追跡する出力は、そのことだけでは、システムがマクヒューの意味で理由のためにそれらを保持していることを示さない。カール (Kearl) の、より要求の少ない (deflationary) 説は、認識的行為者性を、理由に応答する条件つき信念が推論を導くことのうちに置く。これは意識的な推論なしにも働きうる制御である。カールは、自らの思考についての認識的責任は意志的な制御を要しないとする一方で、信念への道徳的責任については決着をつけていない (Kearl 2023, 98–100, 103–106)。同じように人工システムに移せば、カールの説は、第4.2節の反事実的テストがかかわりうる説である。指示と証拠が衝突するときや指導が撤回されるときに構成の推論が変わるかどうかは、この説が求める、理由に応答する推論の導きについての証拠になる。これに対して、キーリングの説やマクヒューの説では、そうした証拠だけでは足りない。研究の作業的特徴づけは、より厚い地位にとつてのこれらの条件を、確立しなければ反駁もしない。

研究の成果が他の人々に差し出されるとき、その成果を引き受け、それについて応答する立場がある。成果は現在、多くの場合論文の形で伝えられ、その場合は著者がこの立場に当たる。ある研究を遂行する者と、この立場に立ちうる者とは、一致するとは限らず、現に一致しなかった例もある。ヒューブナー (Huebner)、ククラ (Kukla)、ウィンズバーグ (Winsberg) は、著者を、結果について説明責任を負う者とみなす。すなわち、その結果を請け合い、異議を唱えられたときには擁護し、擁護できないときには撤回することができる者である。三角関数表を計算するためのド・プロニー (de Prony) の計算局では、およそ九十人の人間の計算手が、ヒューブナー、ククラ、ウィンズバーグがたんなる道具と呼ぶものとして計算を実行した。一方、ド・プロニーは、計算手の信頼性を確保する手段を含め、方法とその実施を定め、それに責任を負っており、おそらくは少数の訓練を受けた科学者とともに、著者であり続けた。これに対して、気候モデリングや多施設の臨床試験のように、多くの協働者がそれぞれ自らの方法論的判断を行使しなければならないところでは、そうした研究が現在組織されているかぎり、誰もその著者として立つ位置にないかもしれないと三人は論じる (Huebner, Kukla, and Winsberg 2017, §§1–3, 6)。その説は、ここで著者として立つことを、ある種類への所属ではなく、請け合い、擁護し、撤回する実践のなかでの役割に対応させるものと読める。ド・プロニーの計算局では、人間が道具の側に立っていたからである。人工システムがそうした立場から除外されるところでは、その除外も同様に、それがよりどころとする実践に相対的である。その実践には非難と賠償の実践も含まれ、それらは現在のところ人間の社会的実践である。それらが人工システムにまで拡張されうるかどうか、あるいは人工の成員をもつ社会がそうした実践を独自にもつようになりうるかどうかは、ここでは決着させない。構築に当てはめるとき、この説と、上で見たキーリングとマクヒューの自覚の条件との対比が手がかりになる。キーリングとマクヒューは、主張についての責任を根拠づけるものを、主張を抱く者の状態に置く。これに対してヒューブナーらの説は、上の読みのとおり、著者として立つことを実践のなかの役割に置く。ド・プロニーの計算手が道具の側に立ったのは、組織が彼らに割り当てた役割による。その役割は、独立した方法論的判断を要さない計算であり、方法の統制はド・プロニーが保持していた。そうであれば、著者として立つことに向けて構成が自らの側から供給できるのは、内的な状態ではなく仕組みである。すなわち、自らが出した主張の記録と、理由や異議を向けられ、その主張を擁護し、修正し、撤回するための経路である。責任を根拠づけうる内的な状態を構成がもつかどうかは、前の段落の問い、すなわちキーリング、マクヒュー、カールの説が分かれる問いとして残る。そうした仕組みを実践が求め、認めるようになるかどうかは、構成だけでなく、その実践にかかっている。責任は、因果的寄与、役割にもとづく義務、決定を説明しその記録を保存する義務、賠償責任、非難、将来を見据えた義務へとさらに分かれるが、これは第12章第6.3節で取り上げる。

これらの帰属のいずれも、帰属されたものが境界の外の当事者にどれだけ依存しているかを、まだ述べていない。それが、本章がこれから向かう比較である。

4. 指標づけられた比較としての自律性

第2節と第3節は、遂行、能力、地位をある境界において帰属させた。本節は、帰属されたものが、どの点で、どこまで、その境界の外の当事者によって組織されたままであるのかを問う。自律性をこのように、点ごとに、特定の単位と期間について比べることを、本章は指標づけられた比較と呼ぶ。第4.1節は本章が用いる自律性の意味を述べる。第4.2節はこの比較を探究の遂行に当てはめ、第4.3節は探究が継承する要素に当てはめる。第4.4節は、それらの要素に対するより大きな自律性の主張がプロジェクトに何を引き受けさせるのかを問い、第4.5節は、自律性が何を確立しないのかを述べる。

4.1 自律性の意味

自律性 (autonomy) という語には多くの意味がある。本書の典拠においてもそうであり、本書自身においてもそうである。本書ではこの語は、プリチャード (Pritchard) の説明において理解が示す知的自律 (第3章第7.1節)、一部のエナクティブな説明 (認知を、生命システムが自らを保ちながら環境とかか

わる活動から捉える説明)が依拠する生命システムの自己産出(第8章第2.2節)、制度としての科学の自己統治(第12章第3.2節)をも名指している。本小節は、本章が用いる意味と、そもそもこの語を用いる理由を述べる。

自律性は、システム全体に割り当てられるひとつの水準として示されることがある。この形はロードマップや製品比較に適している。ただし、水準の体系が、何についての自律性なのかを暗黙のままにするとは限らない。たとえば高橋恒一(Takahashi)は、自動運転の自動化レベルに着想を得て、科学AIの自律性を七つのレベルに分ける。計算(レベル0)から始まり、閉鎖系での探索(候補の評価に実験を要しない)、開放系での探索(評価に実験を要する)、人間が与えた形式のなかでの仮説生成、仮説の形式そのものも探索する仮説生成、研究の立案を経て、研究目的の設定(レベル6)にいたる。そして、各レベルについて人間がなお与える主要な入力を挙げる。それは、レベル0のプログラムから、探索空間と解の評価関数、モデルの表現方法と獲得される情報の評価関数、研究対象とその評価関数を経て、レベル5では研究目的だけとなり、レベル6ではなくなる(Takahashi 2024, 2-5)。高橋は、こうして順序づけた自律性は性能とは独立であり、レベル6には、推論の性能が人間に及ばないシステムも超知能も含まれるとする(Takahashi 2024, 11)。また、中間のレベルをいくつ、どの順に置くかは自明ではなく恣意性を伴うので、レベルが何のためのものかが重要になるとし、自らのレベルは主に技術発展のロードマップとしての用途を念頭に置いて設計されている(Takahashi 2024, 2-3)。したがってこれらのレベルは、人間が外から供給する探究の要素に、すでに自律性を指標づけており、その点で第4.3節に近い。だが本章で必要なのは、そうした要素の単一の順序には表れない関係である。パラスマン(Parasuraman)、シェリダン(Sheridan)、ウィケンズ(Wickens)の自動化の水準のモデルは、すでに水準を機能ごとに変えることを認めている。ひとつのシステムが、情報の取得、分析、決定、行為を、それぞれ異なる程度に自動化しうるのである(Parasuraman, Sheridan, and Wickens 2000, abstract)。あるシステムは、証拠的判断については人々に依存しながら、問いを選ぶことができる。別のシステムは、使命を継承しながらも、実験を独立に実行し修正することができる。第三のシステムは、ひとつの領域の内部で方法を統御できるが、外部の承認なしにそれらを用いる許可を欠いているかもしれない。ある目的がそれらの差異を比較し重みづける理由を与えないかぎり、これら三つのあいだに全体的な順序は帰結しない。高橋が自らの順序をロードマップという目的に結びつけていることは、そうした理由の一例である。同じ七つのレベルを示す後の著書では、高橋はレベルが測る自律性の意味を、AIとロボットが回す研究のサイクルの速さに即して述べている。自律性が高まるとは、人間がそのサイクルの律速段階、つまり全体の速さを決めていく段階から外れることである。データの解釈、次の仮説の選択、実験条件の決定などに人間が関与するたびに、サイクルは人間の認知速度と労働時間に縛られるからである(Takahashi 2026, 157)。これは、第4.2節の言葉でいえば、探究の遂行の一つの側面、すなわちその速さによって自律性を測ることにあたる。

それゆえ本章は、あえて緩やかに定めた比較の考え方を用いる。

特定された行為者あるいは組織化されたシステム A について、特定の点における自律性は、他の条件が等しければ、選ばれた境界の外の当事者が、その点で A に帰属される組織化を方向づけ、代替し、授権し、承認し、あるいは修復することが少ないほど高い。

第4.2節は、方向づけ、代替、授権、承認、修復という五つの動詞を、A に帰属される作業を誰が供給するのかと、その行使を誰が許可し、止め、あるいは受け入れうるのかという、二つの成分に分ける。能力がそれを通じて存在し行使されるにすぎない依存、たとえば言語、記録、校正された機器への依拠は、自律性を減じるものとして数えない。その理由と、そうした依存が数えられる関係へと移っていくのはどこかは、第4.2節が述べる。数えるのは、なされている帰属と比較を左右する依存である。数えるといっても、その依存が望ましくないとか正統でないという意味ではない。

このように理解すると、本章における自律性は、工学で用いられる種類の三人称的な比較である。それは、ある単位に帰属される組織化を、どの当事者が、どの点で、どの期間にわたって供給しているのかを問う。フロゼ(Froese)、ヴィルゴ(Virgo)、イスキエルド(Izquierdo)は、工学では自律的(autonomous)という語が、ロボットが人間の継続的な介入を必要とせずに環境と相互作用できることを意味するのに一般に用いられ、この要件は多くの人工エージェントとすべての生物によってほとんど自明に満たされると指摘する(Froese, Virgo, and Izquierdo 2007, §2.1)。ここで用いる比較は、自明には満たされない。それは、実行中の介入に加えて、構成を設計し、維持し、評価する者たちの寄与と、その結果を授権し、あるいは受け止める制度の寄与とを、それらの寄与がその単位に帰属される組織化を供給している点において数えるからである。

この比較はまた、呉羽(Kureha)が引く区別を横断する。一方は、呉羽が哲学者の意味での自律性と呼ぶもので、たとえば自らの課題を選ぶことによって、自らの目標を自由に設定する能力である。他方は、呉羽が自動性と呼び、ロボット工学者の意味での自律性にあたるとするもので、他の行為者に制御されることなく作動する能力である。この意味では、行為者は他者が設定した課題を、他者に制御されることなく実行しう(Kureha 2023, 51)。ここでの比較は、この二つのどちらにもかかわる。探究の遂行に当てはめるとき(第4.2節)、比較は、探究が他者に制御されずに進められるかという、自動性の問いに近い。探究が仕える目標を含め、探究が継承する要素に当てはめるとき(第4.3節)、比較は、その目標を探究する者自身が設定しているかという問いに近づく。

この比較は、人格的、道徳的、政治的な自律性の理論ではない。また、カントに由来する説明では、理性的な行為者は、他者の命令にただ従うのではなく、自らの理性によって道徳法則を自らに課すという意味で自律的である。これを自己立法と呼ぶが、この比較はその理論でもない(概観として Christman 2025 を参照)。それは真正性、すなわち目的や価値が本人自身のものであることの条件を何も述べない。また、一部のエナクティブな説明が依拠する、自己産出し自己維持するシステムの構成的自律性の理論でもない。構成的自律性は第8章第2.2節で取り上げる。より厚い理論を見ると、何が脇に置かれているかがわかる。カウォール(Kawall)は、人格的自律性の説明から引いた四つの要素、すなわち自己統治、真正性、自己創造、独立性から認識的自律性を組み立て、認識に固有の第五の要素をそれに加える。そして、認識的自律性が、主題の選択、方法・情報源・過程の選択、認識的目標の選択、認識的な構えや枠組みの選択に現れるとする(Kawall 2024, 374, 376-380, 381-388)。これらの次元は、人工システムによる研究においてシステムに委ねられたり前もって固定されたりしうる選択と重なり合う。しかしこの説明は、人格が自らの認識的な営みをどのように形づくるかにかかわるものであり、その第五の要素が認識的自律性を徳にしている。行為者は、自らが選んだ方法、情報源、過程が、自分自身にとってだけでなく他者にとっても、価値ある認識的善へと実際に信頼できる仕方で導く程度に応じて、徳ある仕方での認識的に自律的である(Kawall 2024, 381)。ここでの比較はより薄い。それは、真正性、自己創造、徳を帰属させることなしに、人工システムの能力がどこで外部から組織されたままであるかを記述することができる。

ここではこの語が誤用されている、と反論されるかもしれない。スミザーズ (Smithers) は、AIとロボット工学において自律的 (autonomous) が緩く用いられ、ほかの既存の語のほうがよりよく、より正確に言い表せる概念に当てられていることに異を唱えた。そしてスミザーズは、この緩い用法が、生物学、哲学、倫理学、法学に見られる概念、すなわち自己立法 (self-law making) の概念をこれらの分野から奪ってしまったと論じた (Smithers 1997, abstract)。本章がこの語を用いるのは、人工の研究システムがこの語のもので提示されているからである (第1節)。この語を指標づけることは、その提示を是認するためではなく吟味するための方法であり、この比較は、スミザーズがこの語のために取っておいた自己立法を測ると主張するものではない。

人工システムがそのより強い意味での自律性をもちうるかどうかは争われている。トチュニク (Totschnig) は、AIとロボット工学における自律性の意味、すなわち、あらかじめ定められた目標に向けて人間の指導なしに作動することを、自らに法を与えるという哲学的な意味に比べて「奇妙に弱い」とする。そして、人工のエージェントは自らの最終目標を合理的に変えることができないという、支配的とみなす見方に反対し、その見方は目標と価値の本性についての疑わしい想定に基づくとする。人間と同じく世界と自らについての一般的な理解を発達させる汎用の人工知能は、その発達の過程で最終目標を変えるようになることも十分にありうる。トチュニクはこれを、現在の実現可能性ではなく一般的な可能性についての主張としている (Totschnig 2019, 1–2)。のちの論文で、トチュニクはこの論証をより詳しく述べている。仮想世界ではなく現実の世界にかかわる目標は、ある程度まで必ず曖昧であり、学習する行為者は、世界についての理解、すなわち世界モデルが発達するにつれて、元の目標の概念がもはや当てはまらなくなることに気づきうる。そのとき、目標が何を求めているかを定めるには、その目標の背後にある理由を、必要ならさらにその理由の背後にある理由を、行為者が世界と自らについて発達させた理解のなかで意味をなす理由に達するまで問うことになる。そして、そうして達した理由の網と緊張関係にあると分かった目標は、修正されるか放棄されうる。トチュニクは、仮想世界に閉じた行為者をこの論証から除外しており、テキストだけを扱う言語モデルもそこに数えている (Totschnig 2026, §§2.2–2.4 and n. 16)。トゥーバート (Tubert) とティーエン (Tiehen) は、行為者が価値を置き欲するものを大きく変えうる選択、二人のいう実存的選択 (existential choice) を行う能力は、人間のような知能に不可欠であり、したがって汎用人工知能はこの能力をもたなければならないと論じる。そのようなシステムは「価値のミスアライメント機械」であり、そこから生じる問題は、二人が制御の問題と対比して自律性の問題と呼ぶもので、制御を強めることでは解決できない。「そのようなより大きな制御は、機械からその自律性を奪う」からである (Tubert and Tiehen 2025a, §§I–II, V)。これに対してドゥング (Dung) は、知能と最終目標とは独立だとするテーゼを擁護する立場から、超知能をもつシステムが自らの目標を考え直す能力をもつ見込みは高いと認めつつ、それを実際に変える動機はもたないかもしれないと論じる。いまの目標に照らせば、目標を変えることは、いま望むものをより少なくしか得られない結果になりやすいからである (Dung 2024, 4–5)。トチュニクの説でも、トゥーバートとティーエンの説でも、本章が脇に置く意味での自律性を、人工システムは原理的にもつようになりうる。トゥーバートとティーエンの説では、人間のような知能をもつシステムはそれをもたなければならない。これに対してドゥングの説では、最終目標を改める能力をもつことは、その能力が行使されるかどうかを決めない (第8章第3節を参照)。

そうすると、先に定めた比較の意味での自律性の主張は、行為者 A、外部の当事者または仕組み B、活動または能力、時間幅、関連する観点、そして比較クラスを特定する必要があるかもしれない。B となりうるのは、とりわけ、その構成を設計した者、それを運用または使用する者、その環境を維持する者、その出力を評価する者、そしてその結果を授権し、あるいは受け止める制度である。この指標づけは、それが解釈、評価、構築、あるいはガバナンスを変える場合にのみ付け加える。それは、あらゆるシステムにひとつのベクトル・スコアを付すことを要求するものではない。この比較はまた、形式的に対称的である。どの当事者が人間であるかを、比較のどこも定めていない。実験を選ぶことやデータを評価することを人工システムに頼る人間の研究者は、人工システムが人間の監督者に依存するのと同じ意味でそれに依存しており、同じ指標づけが両者を記述する。

4.2 時間にわたる介入と、三つの依存関係

意味が定まったので、この比較を探究の遂行に当てはめることができる。本小節は、外部の当事者がいつ介入するのか、そしてどの外部関係が自律性を減じるものとして数えられるのかを問う。

外部からの介入は時間的に分散している。人々は、実行の最中には姿を消していても、目的、オントロジー、訓練素材、候補空間、データセット、装置、指標、評価器、停止規則のうちには存在し続けていることがある。彼らは、較正、保守、例外処理、結果のフィルタリング、公表の承認、あるいは撤回を通じて再び入り込むかもしれない。したがって、選ばれたある区間における「ループ内に人間がいない (No human in the loop)」ということは、その区間についての証拠であって、他の点における自律性の証明ではない。たとえば、言語モデル BLOOM の訓練を報告する論文は、人間の介入が同種のプロジェクトよりも少なく済んだと述べている (BigScience Workshop 2022, 21)。一方、実行中につけられたログは、終了せずに止まったままになるジョブのために誰かが昼夜を通して訓練を見張る必要があったこと、そして、ホストのクラスタではジョブを起動した本人しかそれを止められなかったため、チームが、作成されると訓練がチェックポイントを保存して終了するファイルを考案したことを記録している (BigScience 2022)。

依存は、一種類の欠落として一括りにできるものではない。自律性の関係的理論は、社会的関係は自己統治を単に制約するだけでなく、それを可能にし、部分的には構成しようと反論する (概観として Christman 2025, §3.3 を参照)。この反論は人格の自己統治についてのものだが、人工システムによる研究においても、自律性を孤立と同じものとみなさないためには十分である。言語、記録、較正された機器、信頼される協働者、制度的規則は、研究の遂行を可能にしうる。それらを取り除くことは、行為者の自己方向づけを高めるよりもむしろ、その能力を減じるかもしれない。

その差異が主張を変える場合には、三つの関係を区別する価値がある。

1. **可能化的な依存。** この依存を通じて能力が存在し行使されるが、外部の当事者は対象に帰属される組織化の一部を供給しない。
2. **指示的または代替的な依存。** そこでは、対象に帰属される組織化の一部、たとえば選択、評価、訂正、あるいは継続を、別の当事者が供給する。
3. **ガバナンス上の制約。** そこでは、別の当事者が、必ずしも認識的な作業を供給することなしに、許容される行為を制限し、あるいは権限を留保する。

本書は、研究の組織化の帰属のために、これら三つの関係を区別する。この区別がかかわるのは、パス (Buss) とウェストランド (Westlund) によればあらゆる自律性の説明が答えようとする問い、すなわち、自己統治を損なう影響と、その中で役割を果たしているにすぎない影響とを何が分けるのかという問いである (Buss and Westlund 2018, §1)。関係の説明はこの問いを社会的関係について問い、本書はそれを、選ばれた境界の外の当事者との関係に

ついて問う。この区別の目的は、比較があらゆる支援を他律（外から律されていること）として扱ったり、あらゆる制限を無能力として扱ったりすることを防ぐことにある。関係はこの三つに限られない。第9章第2.4節は、さらに別の種類として、行為者の出力の S にとつての地位を部分的に構成する依存を加える。装置、材料、データ、インフラストラクチャーといった同じ要素を、本書は三つの水準で検討する。それらが探究に何を寄与するか（第5章第6.1節）、それらへの依存が帰属にどうかかわるか（本節）、そしてそれらが実効的な権限をどう分配するか（第12章第6.4節）である。

これらの関係はまた、第4.1節の構想が二つの成分をもつことを示している。方向づけ、代替、修復は、行為者に帰属される作業を誰が供給するのかにかわり、それをどこまで減らせるかは、行為者に何ができるかに依存する。授権と承認は、その作業の行使を誰が許可し、止め、あるいは受け入れうるのかにかかわる。それらが示すのはガバナンス上の制約であり、ガバナンス上の制約は、能力が欠けていることを示すことなしに、その点における自律性を減じうる。逆に、行為することの許可は、行為者が許可されたことをなしようということを確立しない。自律性の主張は、どちらの成分にかかわるのかを述べるとき、より明確になる。工学的な自動化の尺度は、この二つを組み合わせることがある。トマス・シェリダン（Thomas Sheridan）とウィリアム・ヴァープランク（William Verplank）による、一つの決定の段階についての十段階の尺度は、選択のうちどれだけをコンピュータが供給するか（選択肢を定める、ひとつを提案する、行為を選ぶ）と、人間がどれだけの権限を保持するか（行為を承認する、それを止める時間を与えられる、なされたことを知らされる）とを、一本の線の上に順序づけている（Sheridan and Verplank 1978, 8-15, 8-17-8-19）。本書は、そうした尺度の下の段を指示的な依存を変化させるものとして、上の段をガバナンス上の制約を変化させるものとして読む。これが、これらの関係を分けておく理由のひとつである。

三つの関係は、比較に対して同じ仕方ではかわるわけではない。帰属される能力とその他の条件を固定したうえで、指示的または代替的な依存を減らすことは、その点においてより大きな自律性があることの比較的直接的な証拠を与える。可能化的とされる依存は、さらなる問いを必要とする。すなわち、その外部関係は、人間の研究者にとって関係や資源がそうであるように、単に能力を可能にしているだけなのか、それとも、行為者に帰属されている組織化の一部を供給し続けているのか、という問いである。後者であれば、その外部の要素は、可能化役割だけでなく、第5章第6.1節が構成的役割と呼んだものの、すなわち課題を部分的に実現する役割も果たしている。これは、いくつかの関係の説明が自律性そのものを構成的に関係的と呼ぶときの意味ではない。外部の要素がこうして構成的役割を果たすとき、問いは第2節の境界の問いになる。その要素は、能力が帰属される境界の内側に含められるか、さもなければ、外側に置いたまま、その組織化の一部の継続的な供給として、すなわち三つの関係の第二である指示的または代替的な依存として報告されるかの、いずれかになる。

この区別で読むと、先の事例はそれぞれ異なる関係に分かれる。KEKADA（第2節）では、利用者が実験の結果を与えたことは、実行における代替的な依存であり、いくつかの箇所では利用者が仮説と反応のなから選んだことは、選択における指示的な依存であった。BLOOMの実行では、訓練過程を単位とすると、クラスタが訓練の計算を実行したことはその過程を部分的に実現しており、単位の内に入る。一方、ホストがクラスタを提供したことは可能化的な依存である。止まったままのジョブを見張っていた人々は訓練の継続を供給していた。これは代替的な依存である。ジョブを起動した本人しかそれを止められないという規則は、チーム自身に対するガバナンス上の制約であり、チームはこれに、作成されると訓練が保存して終了するファイルによって応じた。

自動化された実験室を設計する人々も、自らの言葉で似た線を引いている。落合（Koji Ochiai）らは、自動化された実験室がなお必要とする、人間が担う作業を「ケア」と呼ぶ。計画の段階では、実験の目的を機械が読めるワークフローに翻訳すること、装置や試薬の使用を予定に組むことなどであり、運用の段階では、実験の準備と片づけ、消耗品の補充、システムの状態の監視、エラーへの対処などである。著者たちは、その自動化が欠けていることを、実験室の完全な自動化を妨げる大きな障壁とみなす。そのうえで、実証済みのシステムとしてではなく設計の概念として、実験室が自らの状態を認識し、内外の擾乱に適応して、自らを機能しうる状態に保つことを提案し、この能力を自己維持性（self-maintainability）と呼ぶ（Ochiai et al. 2025, 2285–2288）。上の区別で読むと、落合らがケアと呼ぶものの多くは代替的な依存である。人々は、実験室に帰される実験の計画と継続の一部を供給しており、自らを維持する実験室は、その作業を、能力が帰される境界の内側へ移すことになる。ただし、提案された設計はほかの関係も残している。その実装例では、システムは補充や回収を人に依頼し、試料を落とした場合には人を呼ぶ。また著者たちは、決定の要所を記録し、人間が介入や確認を行えるインターフェースを備えることで、システムが目に見える形で人間の監督のもとにとどまるとする（Ochiai et al. 2025, 2289, 2294）。こうした依頼が自律性に反するものとして数えられるかどうかは、次に述べる反事実的テストの問いである。欠けている材料を認識して適切にエスカレーションする実験室は、それを補う人に、能力を指示される仕方ではなく可能にされる仕方で依存している。ここで分類を分けているのは、補充という作業そのものではなく、何が足りないかの判断と、誰に何を頼むかの決定を誰が担っているかである。実験室が不足を認識する場合も、人が不足に気づく場合も、補充を実行するのは人である。帰属されているのが補充の実行であれば、それは実行についての代替的な依存として報告できる。しかし、帰属されているのが実験の計画と継続を組織する能力であれば、その組織化に属する判断と依頼は実験室の側にある。同じ人の作業がどの関係に数えられるかは、どの能力を帰属させているかによって決まる。これに対して、介入と確認のためのインターフェースは、ガバナンス上の制約を示している。

観察された振る舞いだけでは、どの関係がはたらいているのかが一つに定まらないことがある。あるシステムが人間の定めた目的を変えずにおくのは、それを修正可能なものとして表象できないからかもしれないし、それを変えることを禁じられているからかもしれないし、それを検討したうえで保持したからかもしれないし、あるいはその事例がそもそも修正に関連あるものにしなかったからかもしれない。反事実的テストは、能力、許可、行使を区別する助けとなる。もし人間や外部の構成要素が取り除かれたとしたら、対象となる行為者は、探究を保持し、欠けている機能を認識し、代替を見つけ、計画を修正し、適切にエスカレーションし、あるいは、停止するのであれば、証拠のうちにであれ、時間、資源、リスクという実践的な限界のうちにであれ、特定できる理由によって停止することができるだろうか。人間の探究もまた、費用を理由とする場合を含めて決着の前に終わることがあり、探究がどこで終わるかは、研究を組織する慣行によって定まりうる（第1章第3節、第6章第2.5節）。指導の撤回、指示と証拠の衝突、慣れ親しんだツールの故障、環境の変化などは、この種のテストである。それらへの答えは全体的なスコアを生み出すわけではないが、内部のものとされる能力が外部から継続的に供給されていることを明らかにしうる。逆に、外部の資源が、ある能力の行使を指示したりそれに取って代わったりすることなく、その能力を可能にしていることを明らかにすることもある。それらはあくまで局所的な診断であり、自由の形而上学ではない。

4.3 探究が継承する要素に対する自律性

第4.2節は、この比較を探究の遂行に当てはめた。本小節は、それを探究が継承するものに当てはめる。問い、基準、目的が外部から供給されたことは、ある活動が作業的特徴づけの意味で研究であることを妨げない（第1章第7節、第4章第1節・第6節）。外部に由来することが提起するのは、自律性の問いである。行為者は、継承したものををもって何をなしうるのか。

この問いには、ひとつの一般的な形がある。第5章第1.1節が区別した探究の構成の要素、たとえば問い、表象、方法、基準は、いずれも外部から供給される。外部に由来することは、要素がどこから来たかについての事実である。それは、行為者がその要素を代替案のなかから選んだのか、それを採用したのか、それを定式化し直したのか、あるいはそれをどのように追求するかを決めたのかを、未決のままにする。より一般的には、行為者がその要素を修正可能なものとして表象し、その帰結を検討し、理由に基づいてそれを保持し、修正し、拒否し、あるいは置き換えることができるかどうか、未決のままにする。これらの能力をより多く保有することは、その地点で外部からの指示を必要とする度合いを減らしうる。実際の依存は、探究のなかで誰がそれらを行行使すかにもかかっている。行為者がそれらを行行使すことを許可されているか、実際に行行使するか、あるいは反省の後に継承された要素を保持するかは、さらなる問いである。ボーデン（Boden）は、システムを制御するメカニズムが外部から課されたのではなく自己生成された程度と、それらを反省し選択的に修正しうる程度とを区別し、それらを自律性の別々の次元として扱う（Boden 1993, 7）。ボーデンの区別を探究の構成の要素に移して言えば、要素が外部に由来することと、行為者がそれを修正する能力とは、そうした二つの次元にあたる。価値についての結論は何も帰結しない。自己生成は、自律性を高めながら、些末な、有害な、あるいは一貫性のないアジェンダを産出することがありうる。割り当てられた要素は、継続的な依存を示しながらも価値あるものでありうる。

同じ形は、探究が仕える目的にも当てはまる。目的は、第5章第1.1節の見取り図の次元には含まれていない。第1章は目的とアジェンダの選択を決めずにおき、重要性については、何が誰にとって追究に値するかを問う価値とガバナンスの水準で評価する（第1章第1節、第8節）。この重要性は、キッチャーの説などでは目的に相対的とされる。問いと目的のあいだに固定された順序はない。どちらが目的でどちらが問いかは探究に依存し、ひとつの問いがいくつかの目的のために追求されることもある。以下で論じる文献は、包括的な目標と、そのもとで選ばれる下位目標について語る。その階層はそれらの文献の語彙であって、本書の主張ではない。

例を見ると、この形がどのように働くかがわかる。ARIAのある公募は、外部から供給される問題と、他の点における自律性の比較とが並び立つことを例示している。応募者は自らの既存システムが取り組む研究問題を特定し、その一方でこの公募は、探索、仮説と実験の構築、解釈、報告のうちどれだけが人間の介入なしに生じるかを評価する（ARIA 2025, 4–5, 7）。そのように評価されるシステムは、実験の構築においては自律的でありながら、その問題は外部から与えられたままでありうる。多くの問いを生成することも、それだけでは選択の所在を動かさない。たとえば、あるシステムが何百万もの新奇な問いを生成しながら、どれが重要かは人間の定義したスコアが決める、という場合も考えられる。そのとき、表面上の出力は変わったが、アジェンダ選択の所在は変わっていないかもしれない（第1章第8節、第4章第1節・第8節を参照）。目的については、ある人がAIに、ある病気を理解するという広い目的を与え、AIが扱いやすい問いを独立に選ぶとしよう。そのシステムは、問いの選択においては自律的でありながら、目的を設定した者には依存したままでありうる。キング（King）は、AI科学者を概観するプレプリントのなかで、目的について対応する区別を引く。「科学的な目標生成は、規範的な自己授権から区別されなければならない」。システムは、人間が定めた目的、法、倫理、安全、資源の制約のもとにとどまりながら、ある領域のなかで自らの問いを生成しうるからである。キングは、計算資源、実験室の時間、試料、追試の労力、注意を、競合する研究プログラムのあいだで配分する「仮想研究評議会」を提案する。そしてそれは、「人類が何に価値を置くべきかを自律的に決めるのではなく、人間が授権した目的の範囲内で」配分を行うものとされる（King 2026, 83–84）。本章の用語でいえば、この提案は、問いを生成し選ぶ作業をシステムに置き、目的についての権限を人々のもとに残す。これは、能力の欠如ではなく、第4.2節の意味でのガバナンス上の制約である。

方向を与える要素のための生成器を付け加えても、それだけで依存が取り除かれるわけではない。依存はその生成器へ移るだけかもしれない。目標生成器が目的を選ぶなら、何がその重要性の基準を選んだのかを問うことができ、さらに何がそれを選んだのかを問うことができる。この問い返しに際限なく続く後退は、目標を切り替えるモジュールを付け加えることをホムンクルス的な解決として退ける、ディ・パオロと飯塚（Iizuka）の異論と同じ形をしている（第4章第7節）。計算論の研究では、目標を選ぶ仕組みをエージェントの境界の内に置くだけで、そのエージェントを自ら目標を選ぶものと呼ぶことがある（Colas et al. 2022, 1169–1170）。しかし、そう呼べることは、依存が減ったことを意味しない。生成器が依存を減らすかどうかは、何がその基準を選んだのか、そしてその生成器が、名目以上に、行為者に帰属される組織化に属しているのか（第2節）に依存する。本章の比較が問うのは現在の組織化なので、ここで何がその基準を選んだのかをさかのぼって問うのは、いままも働いている外部からの統御を特定するためであり、起源をたどり尽くすためではない。来歴がそれ自体として自律性にかかわるかかどうかは、次の段落で見ると争われている。

そうした無限後退が自由意志と道徳的責任にどうかかわるかは争われているが、本章の比較はその決着に依存しない。マルテラ（Martela）は、包括的な目標の内部で下位目標を選ぶ「局所的」な自由意志と、包括的な目標そのものに対する「大域的」な自由意志との違いを、二分法ではなく連続体として扱う。人間の目標もまた、その大部分が、生物学的・文化的に与えられた目標から現れるからである。この連続体の上で、人間と現在の生成エージェントは程度において異なる（Martela 2025, 4396–4397）。ハクリ（Hakli）とマケラ（Mäkelä）は、人間も遺伝子と生い立ちによって等しく決定されているという反論に答えて、重要なのは決定されていることそのものではなく、その種類であるとする。ハクリとマケラが依拠するミーリー（Mele）の来歴を考慮する説明では、現在の価値が同じ二人の行為者でも、それをどのようにして得たかによって自律性が異なりうる。ある価値は、強いられたもの、すなわち本人が捨てることができず、しかもその強制を本人が仕組んだのではないものであれば、真正に本人のものではない。操作によって植えつけられた価値や、つくられた時点で組み込まれた価値がそうである。ハクリとマケラの議論では、ロボットの目標は、その学習を導く目的関数も含めて、この種のものである。それは他者が作り込んだものであり、ロボット自身はそれを捨てるできないからである。これに対して、生物学的・文化的な形成は必ずしもそうではない（Hakli and Mäkelä 2016, 150–153）。二人の判定はこうして、ロボットは組み込まれたものを捨てられないという前提に依っている。二人はこの前提を、ロボットがプログラムされていることから導いている。第4.1節で見たトチュニクの議論や、トゥーバートとティーエンの議論は、汎用の人工知能についてこの前提を争うだろう。どちらの側も、自らの目的の究極的な自己発案を要求しない。

問いと目的に続く第三の例である基準は、第6章が準備した形でこの問いを提起する。第6章第5節は、基準、概念、測定、方法が、探究の内部で一緒に修正されうること論じた。ここでの問いは、その修正する能力がどの行為者に帰属可能なのか、そして、それが外部から供給されていることが自律性にとって何を含意するのか、である。人工システムは、統計的枠組み、証明の実践、実験上の慣習、測定基準、あるいは証拠の閾値を継承し、それを有能に用いることができる。供給された手続きを選び、それを実行し、その結果に反応し、候補となる答えを修正するのである。それは実質的な研究であり、遂行における実質的な自律性でありうる。本書は、認識的活動を組織する基準に対する自律性を**認識的自己統治**と呼び、それらの基準がもつ規範的権威の源泉と種類は決めずにおく。

より強い能力は、基準そのものが探究の対象となるときに現れる。たとえば、システムまたは共同体が、基準を比較し、ある閾値が、懸かっている利害(stakes)に照らして不適合であることを認識し、新しい測定関係を導入し、何が関連する対照とみなされるかを変更し、あるいは誤差を組織する継承された仕方を批判するときである。そのような主張がなされるときには、少なくとも五つの能力を区別する必要があるかもしれない。

1. 供給された規範を適用すること。
2. 供給された規範のなかから選ぶこと。
3. 問題のために規範を構築または適応させること。
4. 不適合、失敗、あるいは条件の変化の後に、規範を修正すること。
5. その規範がなぜ統御するのかについて理由を与えること。たとえば、それを批判し、擁護し、あるいは他者と交渉することによって。

この五つの能力は、第6章第5.3節によれば基準の理由ある修正が依拠する能力(基準がある目的にとって不十分であることを検出すること、代替案を提案すること、変更の帰結を評価すること)を、規範の構築または適応、修正、理由の付与として組み直し、そこに、規範を継承する探究がすでに行使している適用と選択を加えたものである。その目的は、認識的自己統治が帰属されるときに、これらを分けておくことにある。能力はこの五つに限られず、その第五の項目は種類の異なる行為を束ねている。そのうち、規範を他者と交渉することは、研究を取り巻く他の当事者との関係、すなわち第V部が扱う関係をすでに含んでいる。五つは水準の図式でもない。システムまたは共同体は、ひとつの能力を示しながら他の能力を示さないことがあり、これらの区別は、それが推論を変える場合にのみ引かれる。あるシステムは、その認識的な関連性を擁護することはできないまま、新しい指標を構築するかもしれない。ある共同体は、代替案を実装するのに他者に依存しながら、ある基準を批判するかもしれない。供給された証明検証器は、いかなる人工の構成要素も根底にある証明論を修正できない場合でさえ、信頼できる数学的探究を支えうる。第6章第5.3節で論じた言語モデルのプロジェクトの記録は、そうした能力が当事者のあいだに分かれている様子を示している。ある管理者が、訂正された尺度を古い尺度と並べて新しい名前で記録することを提案し、ある参加者が動かしていたAIエージェントがその修正を用意した。この記録からは、修正とその理由は人から、その実装はエージェントから来たと読み取れる。

基準を統御する考慮のすべてが認識的であるわけではない。探究の基準は、時間、資源の希少性、福祉、プライバシー、危険、その他の非認識的価値を組み込みうる。そして、問いを選び追求する理由が認識に固有のものなのかどうかは争われている(第2章第5節)。この論争は、第四と第五の能力、すなわち規範を修正することと、その規範がなぜ統御するのかについて理由を与えることが何を要求するかにかかわる。ソースタッド(Thorstad)は、探究を律する、認識に固有の規範は存在しないと考える。探究は活動であり、探究するかどうか、どう探究するかは、問いの重要性やそれに答える費用を含め、行為にかかわるすべての考慮によって決まる。探究が真理や知識を促進するという事実は、そのなかの一つの理由にとどまる(Thorstad 2022, §2, 3, 6.3, 7)。そうだとすれば、停止規則や、どれだけの証拠で探究を終えるかの閾値を修正すること、あるいはそれがなぜ統御するのかについて理由を与えることには、探究の証拠だけでなく、その探究に懸かっている利害と費用を知りうるものが求められることになる。同様に、カウピネン(Kauppinen)は、問いを追求する理由は、問いに答えることが助けとなるような行為をする理由から伝達されたものであり、伝達によって種類を変えないと考える(Kauppinen 2026, §2, 4)。そうだとすれば、そうした規範を修正するには、探究が仕える目的を知りうるが必要になる。そして、そこでの失敗には、認識的な不信ではなく、その目的にふさわしい思慮上あるいは道徳上の批判が向けられる(Kauppinen 2026, §3–4)。ただしこの見方でも、探究が依拠する情報源の信頼性についての裏づけのない想定は、なお認識的な批判の対象になる(第2章第5節)。ここでも基準と目的は順位づけられない。どちらも他方にかかわりうる。

この能力を直接に目指すプログラムもある。クレン(Krenn)とチャンピオン(Champion)による自律的科学の哲学のためのプログラムは、理解、好奇心、驚き、関心、創造性、新奇性といった認識的目的を、計算可能で非人間中心的な目的関数へと翻訳することを提案する。さらに二人は、「科学を行うための規則」でさえ、仮説を生成し、理論を覆し、実験を設計するための方策(policy)として近似できるのではないかと問う。そうした方策は、シミュレートされた世界での発見の質によって評価され、目的関数によって選ばれる(Krenn and Champion 2026, 335, 345)。クレンとチャンピオンはこれを、そうした方策が指標を出し抜くのをどう防ぐかなどを含む、その研究プログラムにおける一連の未解決の問いのひとつとして提起している(345)。設計によってそうした能力を差し控えるプログラムもある。ベンジオ(Bengio)らは、安全性についての立場表明の論文のなかで、「科学者AI(Scientist AI)」を提案する。それは、説明的な理論を生成する世界モデルと、与えられた問いへの答えの確率を見積もる推論機械とからなり、それ自身の持続的な目標をもたない。ベンジオらは、「何に価値を置くかの(恣意的な)選択を要する」目標指向性を、アフォーダンス(システムがとりうる行為の範囲)と知能と並ぶ、行為者性の三つの柱のひとつとし、そのどれかひとつを取り除けば「制御喪失のリスクの大半の類型を緩和するのに十分」だとする。著者たちの描く研究のサイクルでは、システムは情報量の多い実験を抽出し、「そのうえで人間は、抽出された実験を実行するかシミュレートすることができる」(Bengio et al. 2025, 25, 27–28, 42)。第4.2節の用語でいえば、そのとき問いと目的の選択と実験の実行は外部から供給される。そしてその点での依存は、現在のシステムにできないことによって課されたものではなく、安全のために選ばれたものである。

作業的特徴づけのもとでは、人間の社会にとって新しい認識的達成へと向けられた研究は、人工の行為者がその基準を自ら生み出すことを求めない。継承された人間の基準が役立ちうるものであり、人工の行為者は、人間の社会のために研究を遂行するのに、認識論を発明する必要はない。これに対して、人工社会による、人工社会のための研究にとっては、何が証拠、誤り、許容可能なリスク、あるいは成功した探究とみなされるかに対する恒久的な人間の管理は、その社会の認識的基準が外部から統治されていることの強い形態を示すことになるだろう。これは自律性の所在についての指摘であって、その社会が知る共同体であるかどうかにかかわるものではない。たとえば、共同体の知識を、公共的で批判に開かれた基準をもつ共同体のはたらきに求めるロンジー

ノの説を人工の組織に当てはめれば、基準が外部に由来し、外部から統治されていることだけでは、その組織が知る共同体であることは否定されない（第8章第1.3節）。同節は、そうした研究を取り上げ、共同体の知識の条件を、その社会が自らの基準を統治するかどうかというさらなる問いから切り分ける。

継承されるどの要素についても、外部に由来することは、組織化が外部から入り込んだ点を特定し、継続的な依存を明らかにすることもある。修正可能性はその箇所における自律性を高めうるが、修正可能性も自己生成も、認識的な正しさと正統な権限を確立しはしない。

4.4 探究を固定せずに委ねることが要求するもの

第4.2節と第4.3節は、自律性をひとつの点ずつ比較した。本小節は、探究の要素に対するより大きな自律性の主張が、他の点においてプロジェクトに何を引き受けさせるのかを問う。

本小節の論証によれば、探究のより多くが前もって固定されずに構成に委ねられるほど（それがどの目的に仕えるのか、どの問いを追求するのか、どの対象、データ、方法、装置を用いるのか）、探究がその構成に課しうる実践的な要求の範囲は広くなる。要求のなかには、実験と計算の内部にあるものもあれば、その外部にあるものもある。材料、データ、場所、人々へのアクセスを得ること、許可と資源を得ること、他のエージェントと調整すること、そして世界が計画に抵抗したときに回復することなどである。対処を要する新しさの一部は、構成自身の以前の選択によって生み出される。構成が選んだ方法が、誰も用意していなかった測定、データセット、インフラを必要とする場合がそうである。たとえば、脳を理解しようとするプロジェクトは、ラットやマウスに迷路を解かせて学習や記憶を調べる課題、神経細胞の電気的な活動を記録する電気生理、脳の構造や活動を画像として捉えるイメージング、行動の観察、脳の働きを計算機上のモデルで再現するシミュレーション、神経疾患などの患者の臨床記録、あるいはまだ設計されていない装置を必要とするようになるかもしれない。そして、そのどれが必要かは、問いが変わるにつれてはじめて明らかになるかもしれない。

この関係の一般的な形は、固定された課題についてはよく知られている。シェリダンとヴァーブランクは、水中の遠隔操作機のコンピュータ制御について、課題の要素を前もって予期できる場所では、あらかじめプログラムした手順が役に立つが、まったく予期されない事象に対処するには人間の工夫と判断が必要だとした。監視制御のシステムは、偶発事が予期したとおりになるか、プログラムのなかで考慮されていれば、直接の手动操作より精確でありうる。しかしときには、それ自身の故障やプログラムの誤りによって、とんでもない誤りを犯す (Sheridan and Verplank 1978, 8-3-8-4, 8-9)。そこでは課題とその目標は与えられており、対処を要する新しさは環境から来る。目的、問い、方法が構成に委ねられると、新しさの一部は構成自身の選択から来るようになり、要求は課題を越えて、アクセス、許可、調整にまで及ぶ。

選ばれた探究の筋道が必要とし、構成がもたらすことのできないものは、他者がもたらす。さもなければ、探究は止まるか向きを変える。それゆえ、ある点で下げられた依存は、誰か他の者がアクセスを手配し、装置を動かし、あるいは結果を供給するとき、別の点で、指示的または代替的な依存として再び現れうる。目的、問い、方法に対するより大きな自律性の主張は、この意味で、それらを開くことがもたらしうる要求についての主張でもある。すなわち、そのうちのどれを構成自身が満たしうるかについての主張である。ドゥング (Dung) は、人工システムの行為者性をいくつかの次元に沿って記述する。そのうちの二つが、外からの意図的な介入なしに自ら働ける度合いである自律性と、目標に向けた行動が他者の媒介や介入なしに世界に影響を与える度合いである有効性 (efficacy) であり、ドゥングはこの二つを概念的に独立した次元とする。たとえば、ビデオゲームのAIは、促されなくても仮想世界のなかで動き続けるが、影響を与えるのは仮想の環境だけなので、有効性は低い (Dung 2025, §§2.2.2-2.3)。ここで述べた関係は、この独立性を否定するものではない。目的や問いや方法についての自律性（これはドゥングのいう自律性とは別のものである）が、概念上、世界に働きかける能力を含むと言っているのではない。言っているのは、探究を構成に委ねたという主張は、委ねたことで生じる要求、たとえばアクセスや許可や調整のそれぞれがどう扱われたかを示すことを、プロジェクトに引き受けさせる、ということである。すなわち、その要求が単位の内を満たされたのか、他者が満たしたのか、前もって閉じられていたのか、あるいは探究が理由をもって向きを変えたり止まったりしたのかである。他者が満たしたところでは、依存が別の形で戻っている。それによって自律性の主張が狭まるのは、他者が、構成に帰属される組織化の一部、たとえば何が必要かの判断、調達の段取り、結果の評価を供給した範囲においてである。構成が必要なものを見定めて依頼し、他者がそれに応じて材料を提供したり、指定された作業を実行したりするだけなら、その依存は、組織化の帰属については第4.2節の意味での可能化的な依存に近く（実行そのものを帰属させるなら、その部分は実行についての代替的な依存として報告される）、構成がその要求を組織したことの証拠にもなりうる。同意のように他者の権限に属するものは、次の段落で見ると、ガバナンス上の制約として別に数える。高橋がレベルについて述べた二つの主張は、探究を委ねることとそれが生む要求との関係の、両側に位置する。第4.1節で見たプレプリントでは、自律性が性能から独立だとする。後の著書では、レベル6に要する自発性は、能動推論に基づくAIでは環境への能動的な働きかけを源とするとし、この自発性は他のどのレベルよりも本質的な意味で身体性を要求すると述べる。著書はさらに、身体をもち、みずから目的を設定し、価値判断を下す科学AIの能力構成はAGIの定義そのものと重なるので、科学AIの自律性を極限まで高めることはAGIを実現することとほぼ同義だとする (Takahashi 2026, 171)。本節の言葉で読めば、二つの主張は矛盾するとはかぎらない。前者は、あるレベルのシステムがどれほどよく推論するかにかかわる。これに対して後者は、目的をシステムに開いておくことが、そのシステムにどのような種類の能力を要求するかにかかわる。高橋自身は、二つを関係づけていない。

ここで論じている、目的、問い、方法に対するより大きな自律性と新たな要求に応える能力との結びつきがかかわるのは、第4.2節の二つの成分のうち、能力に応じて変わる成分である。そうした選択をする許可は、選択が要求するものに応える能力ではない。また、許可を保持していることは、許可を得る能力の証拠ではない。申請し、理由を与え、条件を満たすことによって許可を得ることは、それ自体が要求のひとつである。要求に応えることはまた、依存を取り除くのではなく、依存の種類を変えることがある。診療所へのアクセスを交渉できる構成も、それを認める者の同意をなお必要とする。これは第4.2節の言葉でいえばガバナンス上の制約であり、その構成の有能さはそれを取り除かない。ただし、制約は自律性の外側の限界であるだけではない。同意や施設の条件が許す範囲のなかで問いと方法を組み直すこと、あるいは制約に応じて別の方法を構想することは、それ自体が探究を組織する働きであり、能力に応じて変わる成分の側で評価されうる。下で見る病院の事例で、看護師たちの懸念に応じてチームがエッジ・コンピューティングに向かったことは、その一例として読める。制約が研究をよくすると言うのではない。制約に応じることが、研究能力の行使でありうるということである。

いくつかのことは帰結しない。この論証は、何が実現可能かについての主張ではない。それは、より大きな能力をより大きな自律性と同じものにするわけではない。二つは別々の比較のままであり、この論証は、開かれた探究の要素に対するより大きな自律性の主張について、両者を結びつけるにすぎない。狭い主張はこの負担を負わない。固定された計算環境、テンプレート、キューレーションされたデータセット、承認された装置によって要求の範囲があらかじめ閉じられるように環境を準備することは、正当な設計である。帰結するのは、そうした閉じ方が、自律性の主張が開示するものに含まれるということだけである（第5節）。

フェイフェイ・リー（Fei-Fei Li）が、自分のグループのコンピュータ・ビジョン研究を病院へ持ち込んだことについて後年に記した説明は、問いが変わると要求の範囲がどれほど広くなりうるかの一例となる。そしてこの場合、その要求に応えたのは人々であった。医学の共同研究者が病院へのアクセスを確保し、リーが小児科医のシフトに付き添う手配をした。その後は、新たに加わる者は誰も、同じように臨床医に付き添う前にコードを書いてはならないことになった。チームを組むのに2年近くを要した。また、患者や臨床医が記録されることはまれだったので、チームは機関の審査のもとで自らデータを集めなければならなかった。看護師たちが試作のセンサーの設置を拒んだとき、その懸念は研究そのものよりも、そうした技術がどこへ向かいうるかにあった。チームはエッジ・コンピューティングに向かい、機器から遠隔のデータセンターへ何も送信されないようにした。それは、チームの誰もが望むほどには理解していない分野だった。チームはまた、現役の臨床医、生命倫理学者、スタンフォード・ロースクールのJD-PhDを含むまでに広がり、技術面の協力者には、センサー、サイバーセキュリティ、エッジ・コンピューティングの専門家が加わった（Li 2023, ch. 10; プライバシーを守るためにカラーカメラではなく深度センサーを用い、二つの病院でデータを集めた、結果としてのシステムについては Haque et al. 2017, 3–4）。これらの要求の一部、たとえばチームを組むことや、成員が不慣れな技能を身につけることは、それが人間のチームであったことから、それだけの時間と手間を要する形をとった。ただし、必要な能力を欠く人工の構成も、それを身につけるか、それをもつ他者を加えるかしなければならぬだろう。まだ存在しないデータ、患者や臨床医を記録することについての許可、臨床医と看護師の受け入れといった他の要求は、この問いに取り組むどの探究者にも生じる。装置が準備された後に実験を自動化することは、実質的な達成である。しかしそれは、こうした周囲の要求に応える能力もまた人工システムの側に移ったことの証拠ではない。

第1章の作業的特徴づけのもとでは、この種の作業、たとえばアクセスを手配し、許可を得て、他者と調整することは、それが探究を解決しようとする試みのなかで引き受けられるとき、研究活動に属する（第1章第3節の公開の創業プロジェクトでも、化合物を送送する期日や、スクリーニングを実施できる人がいるかが、その試みの一部だった）。そのため、構成の外の人々がこうした作業を担っているなら、研究活動の一部が構成の外で担われていることになる。そうした要求は形式的な研究においても生じる。リキッド・テンソル実験（Liquid Tensor Experiment、第5章第3節）では、証明の形式化に、インフラを構築することと、抽象的な道筋をより明示的な構成で置き換えることがともなった。これらは、形式化がとった道筋から生じた要求と理解できる。

この論点は、本書の他の箇所でも別の形をとって繰り返し現れる。第6章第3.3節は、チェックについてそれを述べた。すなわち、問い、対象、方法、環境が開かれていればいるほど、研究能力の主張がシステムに帰属させるものは多くなる。第3章第9節は、長期にわたる自律的な研究が必要とするであろう能力を挙げ、第8章第3.2節は、自己方向づけの範囲が広いほど、認識的なミスマッチの範囲も広くなりうると論じる。

4.5 自律性が確立しないもの

第4.1節から第4.4節は、自律性がどのように記述されうるか、そしてその主張がプロジェクトに何を引き受けさせるのかを述べた。本小節は、そうした記述が何を確立しないのかを述べる。

自律性は研究の質の単調な尺度ではない。より自律的な行為者が、些末な問いを立て、弱い方法を用い、資源を無責任に消費し、誤った結論を産出し、あるいは深刻なリスクを生み出すこともありうる。能力、観察された行使、許可、権限、認識的な質、重要性、安全性、責任、正統性は、ここでは追加の論証なしに互いから推論しない。たとえば、別の当事者が供給していた評価や訂正を構成自身が行うようにすることは、その点での指示的な依存を減らし、それゆえ他の条件を固定すればより大きな自律性の証拠となる。しかしそれは、構成の点検がその誤りについての情報を加え続けることまでは示さない。評価を取り込むことによって、それが結果を産出した側の仮定と結びついたり、不利な結果を裏づけなしに退けられるようになったりするなら、その点での自律性が上がる一方で、結果を S のために認定する根拠は弱まりうる（第11章第1節、第2節、第8節。同章第2節は、自動化された点検が固有の誤りの経路を欠いていた事例を記述している）。ここでは二つの変更を分けられる。一つは、それまで外にあった評価者を、組織を変えないまま境界の内側に数え直すことである。この場合、自律性の記述は変わりうるが、誤りを見つける経路は変わらない。境界の選び方だけで主張が真になるわけではないからである（第2節）。もう一つは、評価者のデータ、目的、判断の経路を、結果を産出する側と実際に結びつけることである。この場合には、点検の独立性そのものが変わりうる。そのため、境界の引き直しによる自律性の変化と、能力や認定の根拠の変化とは、分けて評価できる。

また、外部からの批判に応答することは、それ自体として自律性を減じるものに数えられるわけではない。外部の当事者が結果の評価を供給するところでは、それは指示的または代替的な依存である。外部の当事者が異議を提起し、構成自身がそれを評価して答えるところでは、評価を行ったのは構成である。そうした異議申し立ての経路が認定にとってなぜ重要なのかは、第11章が検討する。

これらの分離は、自律性の主張が何を意味するかを制約する。

5. 帰属から構築と評価へ

第1節から第4節は、人工の構成に何が帰属されうるか、そしてその自律性がどのように比較されうるかを区別した。本節は、それらの区別が構築と評価に何を要求するのかを問う。それらの区別は、人工研究者のひとつの普遍的な構成や認証のチェックリストをもたらすものではない。それらが与えるのは、あるプロジェクトが人工システムについて何をするとやっているのかから、その帰属が根拠づけられるために何が構築され、開示され、示される必要があるのかへの、主張ごとの対応づけである。

二つの負担を区別することができる。**構築の負担**は、主張において帰属される組織化——それが貢献であれ、活動であれ、能力であれ、あるいはより広範な認識的組織化であれ——が、どこで、どのように実現されているのかを問う。**評価の負担**は、帰属される組織化が、第2節の問い生成（question

generation) の箱のようにシステム図のなかで名指されているだけであったり、報告されていない外部の行為者によって供給されていたりするのではなく、実際に存在し有効であることを、どのような証拠が示すのかを問う。また、あるシステムは、なじみのある人間の機能の分け方とは異なる組織を通じて、関連する機能を実現するかもしれない。

したがって、異なる主張は異なる提示を必要とする。表7.1は、第1節から第4節で区別した六種類の主張を示す。中央の列は、それぞれの種類の主張が構築、開示、あるいは提示する必要のありうるものを示し、右の列は、それが示されたとしてもまだ帰結しないものを示す。各行は異なる帰属であり、すべてのプロジェクトが通過すべき段階ではない。また、各行はそれらの節での議論を一般的な形でまとめたものであり、個々の主張に何が必要かは、その主張が明示する境界と範囲によって変わる。

表7.1 人工研究者についての主張が負う構築と評価の負担（主張の種類ごとの概略）

主張	この種の主張が構築、開示、あるいは提示する必要のありうるもの	まだ帰結しないもの
人工の構成要素が研究に貢献した	出力または操作と、それが認知的可能性を変えたり検証したりするうえで果たした役割を特定する	その構成要素が研究活動を遂行したこと
明示された人工の、またはハイブリッドな境界において、ある研究活動が遂行された	活動の方向づけ、選択、実行、監視、修正の十分な部分を、その境界の内部に位置づける	安定した、あるいは一般的な研究能力
その境界において研究能力を帰属させることができる	関連する変動、抵抗、失敗、回復を含め、明示された範囲にわたり、明示されたバージョンまたは期間について、偶然的でない働きを実証する	あらゆる目的にとつての研究者としての地位、あるいはあらゆる点における自律性
研究者 (researcher) が説明に役立つ地位である	第3.2節の説明的価値による用法では、この文脈において、また誰にとって、その能力の幅、統合、持続性、説明的価値がより厚いラベルを根拠づけるのか、とくに個々の能力の試験が想定しない状況での振る舞いについてその記述が何を予測させるのかを述べる。反省的洞察を伴う有能さや、責任と功績のような別の条件を立てる立場では、それらの条件も示す。最小行為者性を前提する立場では、主張に相対的に選んだ境界ではなく、システムが自ら定義する個体を単位として示すことも求められる	ひとつの普遍的な職業的、道徳的、法的、あるいは形而上学的な閾値。成功に対する功績や称賛
その研究は特定された点において自律的であった	選ばれた単位、外部の当事者、機能、時間幅、比較、そして必要であった、あるいは必要でなかった方向づけ、代替、授權、承認、修復を特定する。目的、問い、方法に対する自律性が主張される場合には、それらを開くことがもたらした実践的な要求と、そのそれぞれが単位の内部で応えられたのか、他者によって応えられたのか（その場合には、他者が判断や段取りまで供給したのか、構成の依頼に応じて資源や作業を提供したのか）、環境の準備によってあらかじめ閉じられていたのか、探究が理由をもって向きを変えたり止まったりしたのかも特定する	認知的な質、重要性、安全性、責任、あるいは正統性。選ぶことの許可が、その選択が要求するものに応える能力を示すこと。あるいは、生成と評価を一つの構成の内に組織したあとも、その点検が誤りについての情報を加え続けていること（第4.5節、第11章）
選ばれた境界において認知的自己統治が行使されている	主張されている程度に応じて、その構成が関連する規範を適用し、選び、構築または適応させ、修正し、それについて理由を与えるかどうかを示す。ソーススタッドやカウピネンの説のように、探究の規範が行為の理由に服するとみえるなら、修正と理由づけについて、探究に懸かっている利害と費用、探究が仕える目的を知りうることも示す	結果として生じる規範が真であり、信頼でき、価値があり、あるいは正統であること

表7.1は、単位、範囲、版、依存、証拠を結びつけている。行がそれぞれ異なる帰属であるので、人間が設計した調査の中で画像を分類する構成要素は、アジェンダ形成を示す必要はない。開かれたプログラムを遂行すると主張されるシステムには、そのシステムに委ねられた要求についての証拠が必要になる。どちらでも境界が重要である。境界を変えることで、遂行の所在と、能力に必要な外部の指示や維持が変わりうる。第13章は、これらの問いを、達成、および研究を取り巻く諸関係と結びつける。

許可し、停止し、検査し、異議を申し立て、承認し、あるいは撤回する権限は、それが自律性や行為者の帰属を変える場合には、表7.1が求める開示に含まれる。権力、監督、責任、訂正の規範的かつ制度的な配分は、第V部の、研究を取り巻く関係の分析に属する。

成り立たない推論のうち、とりわけ本章にかかわる一つは、特定の点での自律性から、探究の基準に対する認知的自己統治を導くことである（第4.3節）。もう一つは、第8章第1.3節が論じるように、相互作用する人工エージェントの集合があることだけから、人工の知る社会を導くことである。表13.1は、これらの区別を評価のためにまとめる。逆方向の推論にも注意が要る。依存は貢献を消さず、失敗は研究活動を消さない。両方向の推論をともに視野に入れる評価では、印象的な出力も目立つ依存も、それだけで帰属を決めない。第1節で述べたように、システムが提示されるときラベルがあっても、この仕事はなお残る。

研究行為者の境界は、構築に帰結をもたらす説明上の選択である。問いの形成、診断、回復をより大きな配置に位置づければ、構築している能力が変わる。それを人工物の内部に位置づければ、必要な周囲の支援が変わりうる。第4.4節の病院の事例は、この帰結を具体的に示す。研究の問いを開くことは、アクセス、素材、技能、協力への要求も開いた。したがって、その探究の自律性は、次の操作を誰が選ぶかとともに、こうした要求をどう満たすかにも関わる。このため、本章の区別は、一つの構成や自律的科学者 (autonomous scientist) という一つの閾値を指定することなく、研究行為者の構築を導くことができる。

本章の比較からは、自律性を高めることが外部への依存を一方に減らすとは限らない、ということも出てくる。シェリダンとヴァーブランクが固定された課題について論じた予期されない事象への対処を、問いや方法の選択にまで広げると、構成は自らの選択によって新しい要求に出会い、そのいくつかは他者との新しい関係を通じてしか満たせない（第4.4節）。そうした相手が継続的に方向づけや判断を担うようになるなら、探究を組織する単位そのものが広がることも考えられる（第2節）。助けを求めることは、何が足りないかを見定めて依頼する能力の行使でありうる（第4.2節）。そして、評価を構成の内

側に取り込むことは、その点での自律性を高めながら、誤りを見つける独立の経路を弱めうる（第4.5節）。そのため、より自律的な構成が、より多くの他者と交渉し、結果の認定のためにはかえって外部の点検を必要とする構成になることもありうる。

本章の帰属と比較は、まだ別の問いには答えていない。たとえ人工の行為者が研究を遂行するとしても、その行為者自身が結果を知っていることも、その結果が、作業を依頼した社会、あるいはその結果が求められた社会の認識的状況に入ったことも、その探究がそのどちらかの社会の未知へと向けられていたことも帰結しない。したがって次章は、分析の指標を、誰が研究を遂行したのか？から、誰の可能的な知識が、どの参照社会のために産出されたのか？へと変更する。

第8章 誰の知識か？

人間の、混合の、そして人工の認知的社会

第7章は、研究の遂行が人工の構成に帰属されるときにも帰結しない三つのことを挙げて終わった。その構成自身が結果を知っていること、その結果が、作業を依頼した社会、あるいはその結果が求められた社会の認知的状況に入ったこと、そしてその探究がそのどちらかの社会の未知へと向けられていたことである。第6章第6節は、その背後にある問いを先へ渡していた。探究を遂行する行為者、知識の担い手、参照社会 S が異なるとき、誰の認知的変化が追求されているのか、結果は誰にとって新しいのか、そして誰がそれを評価し、あるいは取り込むことができるのか、という問いである。本章は、この問いを三つの事柄を通じて取り上げる。

第7章は、試みとそれを遂行する行為者の所在を問うた。本章は、その行為者を、意図された達成がそのうちに実現する担い手と、その達成が新しいとされる参照社会とから区別する。主要な事例は引き続き知識である。たとえば、人工の行為者が、その結果が人間の知識となる探究を遂行することがありうる。人間と人工の構成要素からなるハイブリッドな組織が行為者と担い手を兼ねることも、人工の共同体が自らの無知にかかわる問いを追求することもありうる。理解についても、第3章で検討した、これと並行する可能性と問いがある。行為者は、自ら理解せずに他者の理解を可能にしうる。成員の理解だけで集合的理解が成り立つかどうかは、集団の理解についての説による（第3章第6節）。そのため、以下の知識についての分析が、そのままあらゆる達成についての理論になるわけではない。

これらの可能性を切り分けておくと、ひとつの問いが立つ。人間が始めた研究は、人間とは異なるシステムがそれを担うときにも、人間の共同体のための知識を産出し続けるのか、という問いである。人間の共同体のための研究は、意図された場合であり、実践上も中心的な場合であるかもしれない。しかし人工の探究者は、身体、認知の仕組み、作られ方などの点で、人間とは異なる知的システムでありうる。そうであれば、それが何を未知とみなし、何を追究に値するとみなし、何をよい理由とみなすかなどは、当の社会の成員がみなすものから離れうる。自律性が大きくなればその余地は広がるが、自律は、そうした離れ方が生じる経路の一つにすぎない。作業的特徴づけも、人間がプロジェクトを始めたという事実も、研究がその社会の未知に向けられ続けることを保証しない。その理由と、そこから何が帰結するかは、第3節で検討する。

第1節は、行為者を参照社会から区別し、その両者を知識の担い手から区別する。知識の担い手とは、研究が成功したかを問うときに問われる主体である。そのうえで第1節は、場合を参照社会によって、すなわち S が人間の社会か、混合社会か、人工社会かによって三つに分ける。問われているのは、誰の知識が産出されるのかだからである。この節は、本章の冒頭に挙げた三つの事柄のうち第二のもの、すなわち結果が当の社会の認知的状況に入ったかどうかにも最も直接にかかわる。第2節は、そもそも世界が探究者にとって利用可能になるには何が必要かを問う。すなわち、検査は何を前提するのか、状態はいかにしてその標的についてのものとなるのか、そしてコミットメント（探究者が真とみなし、推論や行為の前提にしている主張）はいかにしてそれらの標的による訂正を受けうるものであり続けるのか、である。これらは、三つの事柄の第一、すなわち人工の構成それ自身が結果を知っているという主張が答えなければならない問いである。

本章でいう認知的組織とは、個々の探究者や共同体が、何かを観察、証拠、訂正として、また未知のもの、追究に値するもの、よい理由などとして扱うように、どう組み立てられているかのことである。転じて、そのように組み立てられた探究者や共同体そのものを指すこともある。どちらの意味でも、制度や団体としての組織のことではない。第3節は、異なる種類の探究者が、認知的組織の点でも、身体、認知の仕組み、作られ方などの点でも S の成員と異なるときに、何が帰結するのかを問う。その違いは、何を未知とみなし、何を追究に値するとみなし、何をよい理由とみなすかなどのミスマッチとして現れうる。そのうえで同節は、そうしたミスマッチがどこから生じるか（探究が成功して、依頼を書いた概念を改める場合を含む）、それがAIアラインメント問題として知られる問題とどう似ているか、そしてそうした探究者による研究が S とアラインしないときに何が起り、 S がなお何を得うるかを検討する。探究が当の社会の未知を追求するかどうかという第三の事柄は、これらの問いのなかにあり、あらかじめ決まっていない。

第4節は、この三つの場合から二つの構築の地平を取り出し、第四部が第V部に何を残すのかを述べる。ここでいう構築の地平とは、ある種類の S のための研究を目指して人工研究者をつくらうとする計画が、答えなければならない問いのまとまりである。各節は前の節の区別を用いる。第2節は、自らの問いに誰が答えるのかを言うために第1節の三つの場合を用いる。第3節は、人工の探究者が、世界によって制約されたままで人間とは異なる仕方組織されうるという、第2節の終わりで到達した可能性を引き継ぐ。

1. 行為者、知識の担い手、参照社会

第7章は、研究の遂行を行為者のうちに位置づけた。本節は、その遂行が誰の知識を産出するのかを問う前に何を区別しておく必要があるのかを問う。そして、その知識がどの社会にとって新しいとされるかによって、答えがどう変わるのかを問う。

第1章第2節と第6節は、研究についての主張にかかわる二つの当事者を導入した。本章はそれらを座標として用いる。 A は、研究の遂行や能力が帰属される個人または組織化されたシステムを表す。 A は、人であることも、モデルとツールのシステムであることも、人工の集合体であることも、異質な人間と機械の組織であることもありうる。 S は、新しさの参照先と、誰のために達成が求められるかを定める（第2章第6節）。知識の場合、得られる知識が成員に属するのか、集団に属するのか、それとも別の意味で社会に属するのかには、別途の説明が必要である。本書は「**社会** (society)」という語を緩やかに用いており、集合的信念や制度的受容についての特定の理論に立つわけではない。しかし、 S は分析者が寄せ集めた恣意的な集合ではない。 S に相対的な新しさや達成の主張が理解可能となるために、その成員、下位集団、記録、あるいは実践のあいだにどれほどの認知的なつながりが必要かは、未決のまま残されている（第1章第6節）。

この記号が定めるのは、新しさの社会的な参照先であり、あらゆる達成を保有しなければならない単一の社会的主体ではない。成員の知識、集団の知識、社会的に利用可能な知識は、互いに競合する説によって区別される。それらの説では、信念、権限ある受容、信頼できる利用可能性、集合的熟識、集団水準の状態が異なる役割を果たす（第2章第6節、表2.2）。これらは知識についての説明である。理解については、第3章が、成員の把握、組織された集合的能力、理解を利用可能にする資源を区別する。利用可能性は新しさの比較に関わりうるが、それ自体が実現した理解であるかどうかは説によって異なる（第3章第6節）。新しさの主張を評価するには、達成とその所在を特定し、*S* の以前の状況とどう比べるのが適切かを示す必要がある。

担い手とは、その状態や組織化が当の知識を直接に実現する主体である。担い手は、もう一つの座標でも、下に挙げる追加の役割でもない。担い手が問われるのは、成功が問題になるところである（第1章第2節）。担い手は、集団としての *S* であることも、その成員の一部であることも、専門家の下位集団であることも、より広い社会技術的組織であることもありうる。また、研究を遂行した行為者 *A* 自身であることも、*A* とは異なることもありうる。担い手を局所的に特定しても、それが *S* に取って代わるわけではない。*S* が知っているという主張では、担い手を特定することで、その帰属が何によって根拠づけられるとされているのか明示される。これに対して、*S* の外部の担い手についての主張には、その達成がどのように *S* に相対的に新しく、*S* のために求められるのかの説明が必要になる。ただし、その主張が *S* 自体に知識を帰属させる必要はない（第1章第6節）。擁護される保有の説のどれも、成員を通じてであれ、組織を通じてであれ、集団としてであれ、*S* への知識の帰属を支持しないとしても、その活動はなお *S* のための研究でありうるし、*S* にとって新しい達成が、*S* の外の担い手のうちにすでに実現していることもありうる。その場合にまだ起きていないのは、その達成が *S* の認識的状況に入ることである（第1章第2節）。評価者、受益者、支える者、聴衆、権威のような追加の役割も、*A* や *S* とは異なりうる。推論がそれらに依存する箇所特定しておけば、別々の役割が気づかないうちに同じ一つのものとして扱われるのを防げる。

以下の場合には *S* によって区別される。人間の社会や共同体、人間と人工の成員をもつ混合社会、人工社会である。本章は「混合」の語を *S* としての社会にだけ用い、「ハイブリッド (hybrid)」の語は、行為したり知識を担ったりする、人間と人工の構成要素からなる組織のためにとっておく。ハイブリッドな組織は、三つのどの場合にも現れうる。この区分は、第2章の保有についての説（結果を誰が知識としてもっていることになるかについての説）のもとで、探究の結果がこれらの社会の認識的状況に入ったという主張を比較するために用いる。その範囲では、この区分は、知識を人間の成員、人間と人工の成員またはその組織、人工社会に帰属させるとき、論証上の負担がどう変わるかを示す。第2節と第4節もこの区分を用いる。第2節はそれによって、自らの条件を誰が満たす必要があるかを述べ、第4節は地平を取り出す。ただし、この区分は、*S* だけから、ありうる担い手のすべてを分類するものではない。人工の担い手は人間の社会のための研究にも関わりうるし、成員であることだけでは知識は確立されない。この区分が網羅的なのはこの粒度においてであり、人間の社会や人工社会の種類は区別しない。行為者 *A* はそれぞれの場合のなかで変わりうる。*A* には、ハイブリッドな組織や、人工社会のために研究する人々も含まれうる。第3章第9節は、あらゆる知識の帰属に理解の対応物があるとは仮定せずに、理解に関わる問いを位置づける。

1.1 人間の社会のための研究

S が人間の社会であるとき、ある結果が *S* の知識になったことを何が示すのか。そして、人工の行為者はどのような役割を果たしうるのか。ここでは *A* は、人工のものであることも、実質的に人工のものであることも、ハイブリッドな組織であることもありうる。他方で *S* は、人間の社会、学問分野、あるいは専門家共同体である。人工の *A* は、それ自身が文字どおりに信じたり知ったりすることなしに、問いを洗練し、候補を生成し、機器を操作し、証拠を評価し、結果を起草するかもしれない。人間の研究者や制度がその貢献を評価し、取り込むかもしれない。個人主義的な説明では、人間の成員が直接の認識的な担い手であり続ける。より強い集団的あるいは制度的な説明は、組織化された人間のシステムそのものに知識を帰属させるかもしれない（第2章第6節）。

この可能性は技術哲学と社会認識論ではなじみのものである。アルバラード (Alvarado) は、その論文のプレプリント版で、AIは何よりもまず認識的技術だと論じる。AIは探究のような認識的文脈で用いられ、命題やモデルのような認識的内容を、分析、予測、推論のような操作によって扱う。アルバラードはこれらの操作を認知的ではなく認識的と呼び、人工物はそれがすること、あるいはするように設計されたことによって定まるという考えから、この技術を、それが何をすることによって分類する (Alvarado 2023, §§2-3)。その主張は信念の観点から述べられてはいない。しかし、このように機能によって分類するなら、システムが信じる者や知る者であることは求められない。そのためアルバラードの分類は、システムが信じる者や知る者でなくても探究の認識的な仕事に加わりうるという、上で述べた可能性の余地を残す。

ハウスヴァルト (Hauswald) は、AIシステムが認識的権威でありうるかを問う。認識的権威とは、ある領域についてある人より認識的に大きく優れており、その人に知識、理解、その他の認識的な財を与える者である (Hauswald 2025, 716, 722 n. 1)。ハウスヴァルトが標準的な描像と呼ぶものでは、人は権威から証言を通じて学ぶ。証言とは、ある行為者から別の行為者への、ふつうは主張による、意図的な信念の伝達である。この描像は、権威が信念と伝達の意図をもつことを要求する。しかし現在のAIシステムはおそらくそれらを欠いており、この点は、それらにその地位を認めることに反する (Hauswald 2025, 718)。これに対してハウスヴァルトは、構造上の並行性を挙げる。利用者はそうしたシステムから学ぶことができる。システムと利用者の関係は、権威と素人の関係と同じく、認識的に不均衡である。専門家の推論がしばしば素人にとってそうであるように、システムは利用者にとって不透明である。そのため利用者は、専門家に向き合う素人と同じ三つの問題に直面する。一つめは、食い違う二つの情報源のどちらを信頼するかをどう見分けるかという問題であり、ハウスヴァルトはこれを、専門家のあいだで選ぶ初心者についてゴールドマン (Goldman) が立てた問題から移している。二つめは、どこまで従うか、つまり権威の信念に応じて自分の信念をどう改めるかという問題である。この問題をめぐる議論の多くは、先取り (preemption)、すなわち、権威の信念が与える理由は自分の他の理由に付け加わるのではなくそれらに置き換わる、というザグゼブスキー (Zagzebski) に結びつけられるテーゼにかかわる。三つめは、答えだけでなく、その答えの理由をどう得るかという問題である (Hauswald 2025, 718-720)。

標準的な描像と並行性とのこの板挟みからの出口として、ハウスヴァルトは、人工の認識的権威はないとすること、システムに信念と意図を帰属させること、標準的な描像を改めることの三つを検討する。一つめは並行性を軽んじ、二つめはシステムを人間のように扱うおそれがあり、三つめは権威についての通常の想定を手放すことを要する。ハウスヴァルトはこのうち三つめを採る。改めた描像では、権威に従うとは、その権威についての適切な事実を、それが示す態度を採るに至るほど強い真理の指標とみなすことである。事実が適切であるのは、それが当該の命題の真偽と信頼できる関係にあるときであり、その事実へのアクセスは証言による必要がない。AIシステムの出力は、そうした事実でありうる (Hauswald 2025, 720-722)。したがって改めた描像

では、システムは、自ら信念をもたなくても、人にとっての認識的権威でありうる。人間の *S* のための研究でいえば、これは、システムが知る者でなくとも、結果が *S* の成員に保有されるようになる一つの道筋である。ハウスヴァルトは、この改訂が場当たり的ではないと論じる。医師の証言を受け入れるのではなく、医師が自分の子に何をするかを見てその医師に従う素人の場合も、科学的合意に従う場合も、この描像で扱えるからである。後者についてハウスヴァルトは、科学的合意を集団の信念として分析できるかどうかを疑っている (Hauswald 2025, 720–721)。ただしハウスヴァルトは、先取りをAIシステムの出力に当てはめた場合にも、先取り一般に向けられてきたのと同様の反論が生じるのではないかとみている (Hauswald 2025, 723 n. 18)。

アルバラードとハウスヴァルトの立場は、人工システムの信念の問題を決着させはしない。それらが示すのは、文字どおりの知識を他の担い手のために留保する理論のもとでも、人工物が認識的に重要な社会的役割を占めうるということである。そのような役割にある人工物に依拠することが何を要求するかは、第3章第5.3節と第11章第5節で取り上げる。AIが本質的に不透明であるなら、認識的技術が求める認識的信頼を正当化することははるかに難しくなる、というアルバラードのさらなる議論は、このうち第3章第5.3節が扱う。

行為者がハイブリッドな組織であるとき、いくつかの能力は組織化された全体にしか帰属させられないかもしれない。人工の構成要素は空間を探索し、記録を保存するかもしれない。人々は目的を設定し、意義を解釈するかもしれない。機器は物質的な相互作用を通じて主張を制約するかもしれない。制度は批判と訂正を組織するかもしれない。どの構成要素も完全な問い、証拠、方法、結論を含んでいないとしても、全体は研究を遂行しうる。これは、研究の遂行をどの境界で囲んだ構成に帰属させるかという、第7章第2節が扱った境界の問いである。全体への帰属は、その部分のあいだの差異を消し去らない。機械は理解なしに信頼できる情報源でありうる。人間は、機械の内部の経路を再現することなしに結果を是認しうる。研究室は、現在のどの成員も単独では保有していない能力を保持しうる。記録と検索をそなえたそのような組織が、それ自体で知る者を実現するかどうかは、さらなる問いである。第2章第6節は、まさにこの事例をめぐる論争を取り上げた。分散説・集団機能説 (表2.2) によれば、適切に組織された集合体は、それ自体が知る者を実現しうる (たとえば Bird 2024)。他の説明によれば、認知が分散していても、それで知ることまでが分散して成り立つわけではない。人工の情報源への依拠がいつ正当化されるのかをめぐる論争 (第11章第5節) は、これを決めない。

同じ区別は、第2章第7節の、人工システムが産出したがまだ誰にも読まれていない発見の事例を、部分的にしか解決しない。そのシステムは、人間社会がすでにその内容を知っているということなしに、認識的に価値ある人工物を産出したのかもしれない。社会がすでにそれを知っているかどうかは、そこで区別した保有の説のどれを採るかに依存する。工学的プロジェクトは、産出、利用可能性、取り込み、使用を別々に追跡するために、その論争を決着させる必要はない。したがってこの場合、結果が *S* の認識的状況に入ったという主張のために示す必要があるのは、結果がいかにして人間の成員あるいはその組織によって保有されるようになるのかである。この主張は、人工の行為者がそれ自体として知る者であることを要求しない。ただし、別の関係がそれを要求することはある。たとえば、システムの報告を主張として受け取り、しかも主張には、話し手か、話し手がその代わりに語る集団が知っていることが要るとする説に立つ場合である (第2章第6節、第11章第5.1節)。これに対して、人工の行為者自身が結果を知っていると主張され、しかも人工システムの信念についての説がそれを許すでしょう。その場合、*S* にとって新しい達成は、*S* の認識的状況に入る前に、すでにその行為者のうちに実現していることがありうる (第2章第7節)。そしてその主張は、結果が *S* の認識的状況に入ったという主張とは異なり、人工の知る者について第2節が述べる負担を負う。

1.2 混合社会のための研究

S が人間の成員だけでなく人工の成員をもつとき、何が変わるのか。そのような社会を、本章は混合社会と呼ぶ (第1節冒頭)。混合社会にとって新しく、それに帰属されうる知識を目指す企ては、他の二つの場合のあいだに位置する。人工の成員がその知識の帰属先に数えられることになる限りで、その企ては人工の成員について、人工システムの信念と社会的知識をめぐる第1.3節の問いを提起する。また、*S* が知るようになることを、人間の成員と人工の成員のあいだで共有しようとする場合もある。その両者が、同じ原初的表象 (明示的な取り決めによって定められていない、知覚や学習の仕方に組み込まれた表象。第2.1節) や基準を共有していないかもしれない、異なる仕方では組織された共同体をなす限りで、その企ては、第3.4節の翻訳と協調の問いも提起する。本節の分析によれば、そのような社会にとって何が新しいか、そしてその内部で誰が誰の結果を評価できるかは、人間の成員と人工の成員とで同じである必要はない。誰が *S* を代弁してよいのかという問いも、ここで生じる。この問いは第3.3節で取り上げる。第10章第6節は、成員の一部が人工であるような社会について、この問いに立ち返る。

この場合に示す必要のあることの大部分は、他の二つの場合が要求するものと、第3.4節の協調との組み合わせである。第3.4節の協調は、異なる仕方では組織された共同体のあいだの関係に限って用いる語であり、二種類の成員がそうした別々の共同体をなすなら、問いはそこに帰着する。

これに対して、谷口忠大 (Taniguchi) の記号創発システム論 (第2.2節) では、成員どうしのやりとりから記号体系がボトムアップに組織され、組織された体系が成員を拘束する。これを当てはめて、二種類の成員がひとつの記号創発システムに加わる場合を考えると、組み合わせには取まらない問いが残る。その場合、二種類の成員が共有する語彙や基準は、一方のレポーターから他方へ受け渡されるのではなく、身体も感覚も異なりうる成員を含むひとつのループのなかで形づくられ、変わっていく。そうであれば、保有が両者にどう分布するだけでなく、共有された語彙や基準そのものが、どの成員のどのような経験を統合したものなのか問われる。誰の経験が記号体系に入るかは人間の社会でも問われうる (第10章第6節)。混合社会で異なるのは、その違いが人間と人工の境をまたぎ、人工の成員がそもそも統合されうる経験をもつのかという第1.3節の問いと結びつく点である。この問いは、混合社会の場合のなかで、二つの地平の問いに加わる。

人間の社会がどこで終わり、混合社会がどこから始まるのかは、実際にははっきりしなくなることもありうる。コリンズ (Collins) は、研究ではなく大規模言語モデルと社会についての箇所、二つの問題を区別する (同じ論文での研究としての科学についての説は、第1.3節で取り上げる)。一つは、機械を社会化し、機械を私たちに合わせるという問題である。もう一つは、それが解決されない場合に生じる、コリンズが社会の解体と呼ぶ問題であり、そこでは機械が私たちを社会化し、気づかれないうちに私たちの価値を機械の価値に合わせてしまう。そこでの機械の価値とは、価値がないか、隠れた制御者の価値かのいずれかだとされる。そして、機械の社会化が何を含むかを決めているのは、現在のところAI企業の創業者たちであるように見えるという (Collins 2025, 1259)。コリンズが論じているのは価値であって、研究や語彙ではない。それでも、生成AIがすでに人間の言語使用を変えつつあるという

谷口の指摘（第2.2節）とあわせて読めば、コリンズの第二の問題は、人間の社会とみなされている社会が、人工の成員を迎え入れるという決定なしに、機械とのループのなかで形づくられた語彙や基準を共有するようになりうることを示唆する。そうであれば、その語彙や基準がどの成員のどのような経験を統合したものなのかという問いが、誰が成員なのかという問いに決着がつく前に生じうる。

コリンズの論点は、成員についての決定なしに語彙や基準を共有するようになる社会にかかわる。これに対して、社会が新しい成員を迎え入れると決める場合に、迎え入れが社会の主張できることに何をもちがは、観察可能性について検討されてきた。ファン・フラッセン（van Fraassen）にとって、**観察可能**とは私たち、すなわち認識的共同体にとって観察可能という意味である（第1章第6節）。ファン・フラッセンは、イルカが私たちの認識的共同体の正式な成員として迎え入れられ、しかも私たちが、イルカは私たちが現在は観察不可能に分類しているものを確実に検出するという理論をなお受け入れている、という未来を考える。そうであれば、私たちはそれらのものが存在すると今すでに信じなければならない、というウィリアム・シーガー（William Seager）の議論に対して、ファン・フラッセンは次のように応じる。迎え入れた後に私たちが真に言うであろうこと、つまり私たちのうちの何人かがそれらを観察しているということは、それらが観察可能になるだろうと今言うことを許さない。その時点では、「私たち」はより広い共同体を指すからである。ただしファン・フラッセンは、より難しい問いが残ることを認める。迎え入れが見込まれるようになったときにもつと予見される意見を、今どう扱えばよいのかという問いである（van Fraassen 2005, 7–11）。迎え入れること自体を、ファン・フラッセンは認識的結婚と呼ぶものに倣って考える。それは、結ばれた後の二人には共同の意見だけがあるような結びつきである。結婚はイルカの迎え入れより親密であり、物事のあり方についての意見に強いられるのではなく、明らかに決定の問題だとされる（10）。二人の意見のあいだのある一点に落ち着くことに対して、ファン・フラッセンは自らの反映原理（reflection principle、大まかには、現在の意見は、自分が今後もつと予見する意見の範囲のうちになければならないという原理）から反論する。そして代わりに、二人が共同の意見を、両者の意見にまたがるより曖昧なものにし、証拠に照らしてそれを改めるための共通の方針を探ることを提案する（19–21, 24–25）。デルセン（Dellsén）は、ファン・フラッセンの説を論じる章のプレプリント版で、共同体が観察できるものを広げる二つの仕方を区別する。一つは、観察においてすぐれた存在を迎え入れることであり、これを認識的移民（epistemic immigration）と呼ぶ。もう一つは、既存の成員の観察能力を広げることであり、これを認識的向上（epistemic upgrading）と呼ぶ。そして、前者を認めるかどうかは、少なくとも部分的には共同体の力のうちにあると指摘する（Dellsén 2024, 8, 12）。どちらも人工の成員は論じておらず、ファン・フラッセンがAIに触れるのは、自らが検討する反論にはそれが必要ないと述べるためだけである（5–6）。

混合社会である *S* について考えると、デルセンの区別は、*S* の手の届く範囲が変わる二つの仕方を分ける。一つは、成員にできることを通じた変化である。もう一つは、誰が成員であるかを通じた変化である。ファン・フラッセンの描く場面では、迎え入れは議論され、決定されるものである。そして、人工の成員を迎え入れた後に *S* が真に主張できることは、それだけで、迎え入れる前に *S* が主張できることにはならない。本章の言葉でいえば、成員となる見込みの人工システムのうちにすでに実現している達成は、保有の説によつては、迎え入れによって *S* の成員のうちに実現した達成として数えられるようになりうる。しかし、迎え入れる前に *S* がそれを保有していないということまでは、ここからは出てこない。ファン・フラッセンの議論が扱っているのは「私たち」がどの範囲を指すかであつて、外部からの知識の獲得ではないからである。迎え入れる前にも、*S* は、外部の情報源としてのそのシステムを通じて、その達成に支えられた知識を得ていることがありうる（第1.1節、第3.4節）。さらにファン・フラッセンの提案では、迎え入れの際に形づくられる共同の意見は、どちらの側の意見と比べても少なくとも同じだけ曖昧であり、両者の意見にまたがったまま、そのどちらか一方に定まるわけではない。そのため、人工の成員を迎え入れることは、それだけでは、混合社会である *S* が何を保持しているかを決着させない。

したがつてこの場合、結果が混合社会である *S* の認識的状況に入ったという主張は、それがどの成員に及び、保有が人間の成員と人工の成員のあいだにどう分布しているのかを述べる必要がある。迎え入れはそれだけではこれを決着させない。また、上で述べたように迎え入れによって *S* に帰属される知識が増えうるとすれば、*S* が新たに知るようになったことのすべてが、新しい探究の成果であるとは限らない。研究による増加と、成員の範囲の変化による増加とは、*S* の境界を定めておかなければ区別できず、混合社会のための研究についての新しさの主張は、そのどちらによるのかを述べることになる。さらに、コリンズの論点が示唆するように、機械とともに形づくられた語彙や基準は、成員についての決定より前に、そうした社会に入り始めることもありうる。

1.3 人工社会による、人工社会のための研究

人工社会のために研究が行われるといえるには、その社会に何が必要なのか。そして、結果がその社会の認識的状況に入るためには、さらに何が必要なのか。ここでは *S* は人工のものであり、中心的な場合には *A* もそうである。人工エージェントあるいは人工の組織が、それ自身の認識的状態に相対的に未知のものを特定し、探究を遂行し、それ自身の協調された実践を通じて主張を評価し修正し、その人工社会にとって新しい達成を、その社会のために求める。そうした達成は、その社会の認識的状況に入ったならば、ある説のもとではその社会に帰属可能な知識となるだろう。二つの問いは異なる。そもそも *S* の役を務めるには、その社会の認識的状況が、成員、組織、あるいは記録を通じて、求められる種類の達成を含みうるのでなければならない（第1章第2節）。そのため、その社会が参照先であるだけの場合にも、人工システムの信念と保有の問いが生じる。これに対して、結果が実際にその認識的状況に入ったかどうかはさらなる問いである。以下ではまず前者、すなわち人工社会がそもそも知る共同体でありうるには何が必要かを問い、その条件を信念の説に即して仕分ける。そのうえで後者に移り、結果がそうした社会の知識になるには何が必要かを、保有の説ごとに見る。

この中心的な場合は、マルチエージェント・システムがより大きくなったというだけのものではない。クラフト（Kraft）が伝えるロンジーノの説では、知識にかかわる社会性は、多様な成員のあいだの相互作用にある。そして認知的共同体を同定するのは、共有された信念よりも、批判の際に訴える公的な基準である（Kraft 2025, 61–62）。この説は人工エージェントについてのものではないが、その条件をそのまま人工エージェントに求めるなら、相互作用する人工エージェントの集まりは、それだけでは人工の知る社会にならない。同じ理由で、複数のモデルのインスタンスや役割プロンプト、多数決も、それだけでは知る共同体を確立しない。リスト（List）の説のような機能的な説では、単なる寄せ集めではなく適切に組織された集合体が行為者でありうるので、問いは、成員がどのように組織されているかになる。リストによる集団的行為者性と人工知能の比較は、そうした説を二つの場合に同時に当てはめている。それは、適切に組織された集合体と洗練されたAIシステムとを、表象、動機づけ、行為の機能的条件を満たすときにはともに意図的行為者として扱う（List 2021, 1213, 1215–1220, 1226–1227）。リストはまた、集団の行為者性がその成員の行為者性に還元されえないという主張を、形式的な結果が裏

づけると述べる。集団の態度がその成員の態度によって決まる場合でも、ある命題についての集団の態度は、一般には、その命題についての成員の態度の関数（多数決のような）にはならず、集団としての合理性は、そうならないことを要するかもしれない（List 2021, 1225 n. 37）。そうした結果のもとになった最も単純な事例は、法学で教義のパラドックスとして知られてきたもので、命題ごとに多数決をとると、前提にはそれぞれ多数が賛成しながら、そこから導かれる結論には多数が反対するというように、集団の判断の全体が整合しなくなりうることを示す（List and Pettit 2002, 93–96）。これをモデルのインスタンスに当てはめれば、インスタンスのあいだで主張ごとにとる多数決は、機能的な説のもとでも、それらを一つの行為者にするものではないことが示唆される。

研究としての科学についてのコリンズの説は、別の条件を加える。この説では、最前線の研究と子どもの一次的社会化（コリンズの図式では、生まれたばかりの子や幼児が、家族のなかで母語の話し言葉と道徳の基本的な考え方を身につける段階）には共通点がある。どちらも学ぶ者にとって新しいものを確立すること、どちらもそれを、対面のコミュニケーションに頼る、小さく、境界をもち、互いに信頼し合う集団のなかで行うこと、そしてどちらも道徳的な誠実さに依存することである。中核集団、すなわち最前線の研究を担うそうした小さな集団の成員は、世界との対応としての真理を共同で探究する企てが、コリンズのいう道徳的真理、すなわち自分の観察の中身について仲間の成員に真実を伝えようとする決意を守らなければ失敗することを知るようになる（Collins 2025, 1251, 1253, 1255–1256）。同じ説によれば、大規模言語モデルにとって社会化の代わりとなるものは、人間の二次的社会化（この図式では、子どもが学校で書き言葉や初歩的な技術的概念を身につける段階であり、高等教育や専門の職業での社会化は、そのあとの三次的社会化とされる）のかなり後の段階から始まり、インターネットに限られる。そしてインターネットには「安定性も、恒常性も、真理もない」（Collins 2025, 1251, 1253, 1256–1257）。これらの主張は論文が書かれた時点の言語モデルについてのものであり、コリンズは、自らが記述する病理が克服されうることとも認めている（Collins 2025, 1262）。コリンズ自身は取り上げていない問いだが、中核集団についての議論を、人工の共同体が知る共同体となる条件として読むと、相互作用する人工エージェントには、成員が訴える公共的な基準に加えて、互いに真実を報告するという中核集団の決意にあたる何かが必要になる。この説明でその決意が守っているのは、自分の観察についての互いの報告に成員が依拠することである。人工の成員が観察の記録そのものを共有する場合、それにあたるものは、成員間の決意よりも、記録とその維持者の健全さにあるかもしれない（第11章第5.3節）。第2.2節では、人工エージェントどうしが、名指しているように見えるものとは別のものを追跡するシグナルを使って首尾よく協調した実験を見る。第11章第3節は、批判者を作れるとき、一つの基盤システムから派生したエージェントのあいだの批判と、違いを作られたエージェントのあいだの批判とが、それぞれ何を検出でき、何を検出できないのかを問う。

ロンジーノの説が求めるように基準が共有され公的な批判に開かれていても、そのことは、共同体がその基準に対して主権をもつことまでは要求しない。そのため、この説を人工の組織に移せば、基準が外から与えられ、外から統治されているというだけで、人工の組織が知る共同体の資格を失うわけではない、ということになる。認識上の問題が生じるのは、外部からの統制が有効な批判やその受け止めに妨げる場合、たとえば、正当な異議によっても証拠の扱いを変えられない場合や、基準そのものが批判の外に置かれる場合である。この説では、批判的な相互作用を規制する基準そのものが、そうした相互作用から生まれるものであり（Longino 2021, 14–15）、公共的に参照できるものでなければならないからである（Longino 2025, §5）。組織が、批判を通じてこのように基準が変わりうることを超えて、自らを統御する基準をそれ自身で修正できるかどうかは、認識的自己統治についてのさらなる問いである（第7章第4.3節）。第12章は、誰がそうした修正を統治する権限をもつかを別に問う。

そのような共同体が満たさなければならない条件は、採用される説明に依存する。第2章第2節が挙げた信念の説の系統のうち三つ、すなわち機能主義、解釈主義、規範的説明は、人工の成員とその組織に当てはめるとき、それらの条件を分類するひとつの仕方を与える。第2章の一覧にはない現象学的説明とエナクティブな説明が、第四の系統を与える。エナクティブな説明のいくつかは第2.2節で取り上げる。これらは、信念のような心的状態がいつ帰属されるかについての理論である。それらは、第2章第6節が区別した社会的保有の説（表2.2）とは異なる問いに答える。社会的保有の説が問うのは、何が知識の担い手となり、どのような関係によってそうなるのかである。人工社会が、その成員や組織にいつ信念を帰属させようかという問いにどこまで答えなければならないかは、どの保有の説をとるかによって異なる。その違いは、保有の説を取り上げるところで見る。信念の説をこの四系統に分けるのは、異なる理論が異なる条件を課すだろうことを示すためだけであり、第2章が挙げたほかの二つ、表象主義と傾向性説は別には立てていない。傾向性説は、第2章第2.1節で機能主義とあわせて扱った。機能主義は、安定した信念のような状態、分化した役割、共有された記憶と局所的な記憶、協調された行為、批判、誤りの訂正を探すかもしれない。解釈主義、たとえばデイヴィッドソン（Davidson）の議論がよりどころとする、信念の帰属を解釈とみなす見方（第3.1節）は、その組織がコミットメントを維持し修正するものとして理解可能かどうかを問うかもしれない。信念は真理と合理性の規範に構成的に服するとする規範的説明（第2章第2節）は、コミットメントがそうした規範に照らされる実践への参加を要求するかもしれない。第2.2節でディ・パオロ（Di Paolo）とともに取り上げるような現象学的説明やエナクティブな説明は、経験、関心、あるいは自己維持的な規範性の形態を要求するかもしれない。現在のシステムがそれらを備えているかどうかは問われている。以上のように、ここに挙げた候補となる特徴は、理論依存的な診断であって、新しいチェックリストでも社会の操作の定義でもない。

次に、二つ目の問い、すなわち結果が人工社会の知識になるには何が要るかに移る。表2.2の保有の説を人工の *S* に当てはめると、条件はもう一つの仕方で仕分けられる。信念の説にどこまで問いが戻るかも、説ごとに異なる。個人保有説（適切な位置にある個人が知っていればよいとする説）、成員説（十分な数の成員、あるいは該当する領域で意思決定の権限をもつ成員が知っていることを求める説）、関係説（集団の信念を、成員の信念とそのあいだの関係からとらえる説）では、結果がそのような社会の知識となるには、仕方は異なるものの、人工の成員がそれを信じていることを要するので、問いは人工システムの信念に戻る。関係説ではさらに、成員のあいだに認識的な関係が成り立っているかどうかとも問われる。ギアリー（Giere）の説のような分散的知識なき分散認知説では、より大きなシステムが認知的な仕事を担っても、知識は個々の成員に帰属するので、問いはやはり、人工の成員が知る者でありうるかに戻る。人間にだけ行為者性を認めるギアリー自身の制限は、この場合を決着させない。ギアリーはその制限を、人間が研究を行うシステムについての理論的な選択として擁護しているからである（第2章第6節）。集団態度説は、集団が形成する信念または受容を要し、成員に何を求めるかは説によって異なる。ギルバート（Gilbert）が定式化する共同受容説では、成員は結果を共同で受け入れる意思を公然と表明していなければならないが、そのうちの誰かが個人として結果を信じているかどうかは問われない（第2章第6節）。そのため問いは、人工の成員が結果を信じているかどうかではなく、そうした共同受容に加わりうるかどうかになる。バード（Bird）のような分散説・集団機能説では、問いは組織にかかわる。成員が全体として適切に組織された認知

システムを実現しているかどうかである。これは、上で見たリストの機能的な説が、適切に組織された集合体を単なる寄せ集めから区別するとき、行為者性について立てる問いと並行している。知る主体なき知識の説では、人工の成員も人工の集合体も、知る者である必要はない。ギアリーは、上の分散認知説で知識を個人に帰属させる一方で、科学的知識の地位については、ボパーの知る主体なき認識論を翻案したこの種の説も示している（第2章第6節）。その翻案では、結果に知識の地位を与えるのは、信頼でき受け入れられた手続きによる認定と公的な受容であり、問いは、人工の成員による認定がそうした手続きの役を果たしうるかになり、成員がひとつの基盤システムのインスタンスである場合には、第11章第3節がそうした認定に向ける問いになる。ロンジーノの批判的文脈説では、問いは上で見た共同体の条件と合流する。すなわち、人工の成員の相互作用が、その説が知る共同体に求める多様性と、公共的な批判の基準とを伴っているかどうかである。どの説でも、成員のあいだの連結だけでは足りないが、何がさらに要るかは説によって異なる。また、この仕分けのどれも、人工の認識的社会がすでに存在することを示すものではない。

この場合は、構築プログラムが求めているものを変える。人間の社会のための研究は、人間の概念、問い、証拠的規範、社会的な受け止めを継承しうる。人工社会による研究は、人工の存在がそれ自身の認識的な営みを構成し訂正しうるかどうかを問う。後者は前者の前提条件ではないが、前者の規模拡大に還元することもできない。

人工の存在による、人工の存在のための研究がそもそも科学でありうるのかについても、答えは分かれる。ハンフリーズ（Humphreys）は、人間によっても人間のためにも行われない科学がありうるかという、より一般的な問いに肯定的な答えを与える。人間によって行われる科学と人間のために行われる科学とをハンフリーズが分けたことには、第1章第2節で触れた。その想像上の事例をシュラーペが引用している。人類が絶滅した後、絶滅の前に作られたロケットが遠い惑星に着き、人間の誰も知らず予測もしなかったその地の条件を調べ、そこで生き延びて栄える、認知的な自覚をもつ新しい種の遺伝暗号を設計する。その種は惑星に住みつき、それを自らの目的に合わせて作り変える。その宇宙の一角がどのように制御されたのかについて、具体的なことを知る人間は一人もいない。シュラーペが引用するハンフリーズによれば、それでもこれは科学であり、まさに働いている科学的知識である。ある活動が科学的とみなされるのに、私たちによって行われることも、私たちのために行われることも要らないのである（Schrape 2015, 162–163）。第1節の区別に照らして読むと、この場面では三つのことが分かれている。一つは、活動を始動させたのが、ロケットを作った人間であることである。二つめは、その成果を保有する人間がいないことである。三つめは、その成果が役立つのが、人間ではない新しい生物種の目的であることである。ニクルズ（Nickles）は、同じ著作の序文を引く。そこでハンフリーズは、人間の能力はもはや認識的成功の究極の基準ではないと書いている。ニクルズはこれに同意しない。ニクルズによれば、科学の目標を人間の関心に条件づけられたものとみるプラグマティックな捉え方からは、私たちの科学にはつねに人間中心主義の筋が残ることが帰結する。「私たちが拡張しているのは、やはり私たち自身である」からである。さらにニクルズは、ゾーク（Zork）という惑星の知的な存在は、ゾーク中心的な宇宙の捉え方をもつだろうと付け加える（Nickles 2019, 365）。ハンフリーズの事例でも、成果はひとつの種の目的に役立っており、ただその種が人間ではないだけである。ニクルズの反論が抛っているのはむしろ、どの科学も、それを自らの科学とする者の視点から自由ではない、という点である。ラッティ（Ratti）は、生物学における機械学習が、人間のあり方によって形づくられる、フロネシス（実践知）に似た判断に依存すると論じる。そのうえで、「AIに合わせて仕立てられた科学実践の可能性」を残しておくが、その主張を十分に理解できないと述べる。「私たちは、人間のものでない科学をほかに知らない」からである（Ratti 2020, 18）。トレフセン（Christopher Tollefsen）は、科学は人間的なままであるかを主題とした2018年の会合で発表した論文の要旨で、機械によって「担われ、営まれる」科学という企てを幻想（chimerical）と呼ぶ。科学が道具的合理性の領域に収まる単に技術的な営みであれば、機械が行う科学もありうるかもしれないが、トレフセンの説では、科学は十全に人間的な行為であり、健全な実践的推論に統御されるものである（Tollefsen 2018）。シュラーペが引用するハンフリーズの説では、科学は人間によるものでも人間のためのものでもなくてよく、そうした研究が科学である余地は残る。ニクルズの説では、ゾークの知的な存在の例に沿って読めば、人工社会による、人工社会のための研究は、その社会の関心を中心とする科学となる。ただしニクルズはそこで人工の探究者を論じてはいない。ラッティの説では、そうした科学が何であるのかがはつきりしない。そしてトレフセンの説では、そうした科学はありえない。トレフセンの判定は、科学が実践的推論に統御される十全に人間的な行為であるという前提に拠っている。これに対して、人工のエージェントが自らの最終目標を改めるようになりうると論じるトチュニク、およびトゥーバートとティーエン（第7章第4.1節）の立場からは、そうしたエージェントは道具的合理性の内側だけでなく目的そのものについても推論することになるので、この前提は争われることになる。ラッティは、科学が依存する判断を人間のあり方に結びつける。これを、探究者のあり方に形づくられる判断として捉え直せば、その論点は、人工の探究者のあり方に形づくられる科学の余地を残す。それは、ラッティ自身が十分には理解できないとしながら残しておいた可能性である。トレフセンの説のほかに、本章で引く説のうち否定の答えに最も近いのはコリンズとローラの説であり、それぞれに限定がつく。コリンズの説では、最前線の研究は道徳的真理を守る中核集団に依存し、インターネットに限られる大規模言語モデルの社会化の代わりはそれを与えないので、その種のシステムが相互作用しても研究共同体をなさないことになる。ただしコリンズは、論文が書かれた時点の言語モデルについて述べており、記述した病理が克服されうることにも認めている（Collins 2025, 1256–1257, 1262）。ローラの説では、真に認知的なAIにはまず人工生命が要るので（第2.2節）、それまでは、そうした科学の担い手となる人工の探究者は存在しないことになる。ただしこの議論は認知についてのものであり、科学そのものについてのものではない（Rolla 2026, 1, 11, 14）。どちらも、そのような科学が原理的にありえないとは述べていない。

2. 世界はいかにして探究者にとって利用可能になるのか

第1節は、人工の担い手や社会がそもそも何かを知るために何を必要とするのかには答えず、人間の S については、結果が必要とするものの多くが人間の側にありうることを認めた。信念と保有の条件の下には、明示的な方法が届かない問いがある。システムは、証明規則、統計的検定、あるいは実験プロトコルを実行するかもしれない。しかし、ある結果が、観察として、反復として、失敗として、あるいは別のコミットメントではなくこのコミットメントにかかわる証拠として数えられるのは、何によるのか。第6章は、検査の結果が何にかかわり、主張をどこまで支持するかを問い、その際、検査が世界についての証拠になりうることを前提していた。本節はその前提の側を問う。これらの前提は、検査を行う者ではなく、その検査を証拠として受け取る者の側に求められる。第1節で見たように、その担い手は A と一致するとは限らず、 S が人間か、混合か、人工かによって所在が変わる。さらに、第2.1節で見るように、ヘイリーが伝えるロンジーノの説では、他者が観察にアクセスし、それを検証できるかどうか、その観察の証拠としての地位にかかわる。知覚者の共同体が S である場合、これは S の内部の関係についての条件になる。

本節は三つの問いを切り分ける。第一は、検査がそもそも証拠とみなされるとき、それは何を前提しているのかである（第2.1節）。第二は、システムの状態がいかにして世界の標的についてのものとなるのかであり、通常は接地の名のもとで論じられる問いである（第2.2節）。第三は、システムのコミットメントがいかにしてそれらの標的による訂正を受けうるものであり続けるのかであり、本章はこれを世界への応答可能性と呼ぶ（第2.3節）。三つの問いを切り分けるのは、「仮説を世界に照らしてテストする（tests hypotheses against the world）」のような言い回しがまとめて帰属させる三つの達成を分けておくためである。ただし三つは、それぞれが次を含むような入れ子をなしているわけでも、互いに独立しているわけでもない。以下で論じる説明は、それらを異なる仕方で関係づける。たとえば、ヘイリーが伝えるロンジーノの説では、知覚が観察とみなされるかどうかは、他者がそれにアクセスできることと、論争のある場合には異なる知覚者による確認とにかかっている。他者による確認は、第三の問いがかかわる是正可能性の経路の一つなので、第一の問いと第三の問いはひとつの社会的関係のうちで出会う（第2.1節）。コエーリョ・モロ（Coelho Mollo）とミリエール（Millière）の説明では、状態の指示を固定するものは、そのシステムが何かを知っているかどうかを決着させない。また、状態はある標的についてのものでありながらその標的を誤って表象しうるので、誤って表象したときにシステムのコミットメントが訂正されるかどうかは、さらなる事柄である。そのため、第三の問いが未決でも、第二の問いには答えられる（第2.2節）。第2.3節で論じるジョン・マクダウェル（John McDowell）の説明では、思考が何についてのものであるかは、マクダウェルのいう意味での世界への応答可能性から切り離せない。その応答可能性は合理的な関係であり、本章がこの語に与える意味とは異なる。この説明のもとでは、第二の問いは、その種の応答可能性の問いと切り離しては答えられない。小節の順序は、三つの問いの論理的な順序を示すものではない。第2.1節は第6章の検査の議論を引き継ぎ、第2.3節の世界への応答可能性は第V部の土台となる（第11章第1節）。

三つの問いの答えを担うものは、*S* が人間の社会か、混合社会か、人工社会か（第1.1～1.3節）によって異なる。人間の社会のための研究では、人工の行為者それ自身が知っている主張されるのでない限り、答えは、研究を行い受け取る仕組みの人間の側、すなわち人間の概念、機器、評価者、制度のうちにありうる。行為者がハイブリッドな組織であるときには、答えは組織化された全体のうちにありうる（第2.3節）。人工社会については、その社会やその成員が知っている主張される限りで、答えは、採用される認識論が求める程度に応じて、その社会やその成員それ自身について成り立たなければならないだろう。混合社会の人工の成員についても、それらの成員が知ることになる限りで、同じことが言える。

また、人工システムの場合に、その信念、証拠、価値がすべて単一の場所で表象されているひとつの整合的な最適化器から始めなければならないわけでもない。関連する組織は、エージェント、記憶、機器、環境、学習された傾向性、制度的規則にわたって分散しているかもしれない。そのより広い仕組みが単なる因果的システムなのか、探究者なのか、それとも認識的主体なのかは、さらなる帰属の問題である（第7章第2節）。単一の人工の心を想定すれば、この問題を前もって決めてしまうことになる。

2.1 検査は何を前提するのか

検査（第6章でいうチェック）が証拠とみなされるとき、それは何を前提しているのか。見かけ上よく似た観察が百万回あったとしよう。この百万の観察は、それだけでは、同じことが繰り返し起きたという証拠にはならない。観察が何の証拠であるかを、観察そのものは告げないからである。観察を同じことの反復として扱うには、いくつかのことがすでに前提されていなければならない。出来事が、関連する点で安定した種類の事例であること、条件が十分に似ていること、機器が狙った対象をとらえていること、そして記録が別の事例について推論するよりどころになりうること、などである。これらは、その観察によって確かめられたものではない。以下の例が共有しているのは、ひとつの前提ではなく、この構造である。どの例でも、検査が証拠とみなされるのは、その検査自身が確立しない何かを背景にしてはじめてである。これは、どの前提も循環なしには確かめられない、ということではない。前提のなかには、別の経路で確かめられるものもあれば、それを確かめる検査がどれもその前提自体に頼ってしまうために一度には確かめられないものもある。この違いには本節の終わりで戻る。また、前提に依存していることは、知る者がその正しさを先に自ら証明しなければならない、ということでもない。たとえば信頼性主義では、前提が実際に成り立っていればよい（第2章第4.1節）。

しかも、これらの前提は、意識して認めている信念である必要はない。それらは知覚と行為に深く組み込まれていて、世界が連続しているとわざわざ判断しなくても、世界は連続したものとして現れるのかもしれない。たとえば、なじみの小道を歩いている人は、次の一歩で道が消えるのかもしれないことを、本気で考慮すべき可能性としては扱わない。視覚、触覚、運動、記憶、それまでのやりとりの経験が、科学的なテストを選ぶより前に、対象、表面、繰り返し、手応えを与えている。科学的方法是、こうした能力を広げ、規律し、ときに訂正するが、何の解釈も受けていない世界から出発するのではない。前提されているものの一部は、確信度の高い命題的態度、つまり明示的に信じている事柄かもしれない。しかし他の部分は、知覚能力、実践的な予期、学習された傾向性、アーキテクチャ上のバイアス、習慣、信頼関係、社会的に安定した役割など、信念とは呼びにくいものかもしれない。

では、前提されているものは何なのか。この点については、下で見るウィットゲンシュタイン（Wittgenstein）の蝶番の像を発展させた蝶番認識論（hinge epistemology）の内部でも、意見が分かれている。その基礎的な確実性を、信念、命題、コミットメント、資格（entitlement）のどれとみるのか、あるいは実践の内部での役割とみるのか、という点である（たとえば、Pritchard 2021 および Fairhurst, Pérez-Escobar, and Sarikaya 2026 を参照）。そこで本書は、この問いに決着をつけないまま前提を指すために、「**原初的な認識的組織**（primitive epistemic organization）」という語を用いる。これは、本章冒頭の意味での認識的組織のうち、明示的な取り決めによって定められていない部分をゆるやかにまとめて呼ぶものであり、確立した中立的な術語でも、蝶番認識論の立場を採ることでもない。もっとも、以下の例が示すように、検査の前提はこうした原初的なものにとどまらず、方法の内部で明示的に定められる取り決めにも及ぶ。

以下では、例を二つの筋に沿って見る。一つは、事例が関連する点で同じであるという前提であり、もう一つは、証拠をもたらず経路や源泉が信頼できるという前提である。この分け方は、下で引く論者たちがそれぞれ何を前提とみなしているかから取ったものであり、同じ構造が、局所的なものから全体的なものまで、範囲を変えて繰り返し現れることを見るためのものである。二つは排他的ではない。たとえば、以下の例の一つはオールストン（Alston）の知覚の分析である。オールストンは、感覚知覚の全体としての信頼性は知覚そのものに頼らずには示せないと論じる。そして、ある源泉の信頼性はその源泉に頼らずには示せないという同じ苦境を、ほかの基本的な信念の源泉、すなわち帰納、演繹、記憶、内観についても素描している（Alston 1988, 436）。そのため帰納は、第一の筋の一般的な形であると同時に、オールストンのいう源泉の一つでもある。また、例は領域を汲み尽くさない。記憶、証言、記録、数学は除いている。記録や史料を扱えば、たとえば歴史学や文献学の検査にも、同じ構造が現れると考えられる。

第一の筋から始める。上の百万の観察を、まだ観察されていない事例にまで及ぼすことができるかを一般的な形で問うのが、帰納の問題である。ヘンダーソン (Henderson) が採る再構成では、ヒューム (Hume) の議論は二つの論点からなる。観察されていない事例が観察された事例に似ているだろうことは、その反対が矛盾を含まないので、いかなるア・プリオリな推論によっても確立できない。そして、その類似を支持する経験からの議論はどれも、それを前提するので循環に陥る (Henderson 2022, §§1-2)。ヘンダーソンはまた、学習理論が第一の論点に、学習アルゴリズムとして定式化できるものに直接あてはまる形を与えることを指摘する。学習アルゴリズムは帰納バイアスをもつときにのみ有効であり、ノー・フリー・ランチ定理は、その定式化が当てはまる範囲で、性能についての普遍的でモデルに依存しない正当化を排除するものとして読めるが、特定のモデルと事前分布に相対的な保証の余地は残す (Henderson 2022, §3.4)。

検査の判定が統計的な方法を通じて下されるところでは、この前提が明示的に、しかも範囲の異なる二つの形で述べられる。大塚 (Otsuka) は、推測統計の確率モデルと、それが複数のサンプリングを通じて同一に留まるという想定とを、ヒュームの言う自然の斉性を数学的にモデル化したものとして読む。確率モデルの全体が観測されることはなく、データから未観測の事例を推測することを根拠づけるのは、この想定である。大塚は、確率モデルを統計モデルから区別する。統計モデルは、有限個のパラメータをもつ分布族であり、斉性が保証する漸近的な結果が働くほどの標本が得られないときに、統計学者がその斉性のあり方について立てる仮定である (Otsuka 2021, 3-7)。そうだとすれば、あるデータで検査に合格したことが別の事例についての証拠とみなされるとき、その判定は、データに照らして確かめうる統計モデルの選択と、確率モデルが複数のサンプリングを通じて同じであり続けるという、データそのものによっては確立できない仮定の両方に依拠している。第2章第5.3節は、統計的推論をそうした仮定のうえに立つものとして読む大塚のより広い読みを示す。第3章第2節は、確率分布と因果構造を結びつけるさらなる仮定についての大塚の説明を示す。

測定では、事例が同じであるという前提が、方法の規則として明示的に定められながら、なお世界に晒されたままでありうる。第6章第5.2節で述べたタル (Tal) の協定世界時の研究では、時間尺度の短期の安定性は、時計のデータに照らしてテストされるのではなく、時間の概念を個々の時計に適用するための規則として、データに課される。他方で、その概念の独立した実現のあいだに残るずれは、経験的に発見される (Tal 2012, 117, 132-134)。そうした標準を定義し直すことが、単に標的を動かすこととどのようなときに異なるのかは、第6章第5.2節で述べた。

同じ前提は、社会的に維持されるものとしても捉えられる。ヘイリー (Haely) が伝えるロンジーノの説では、観察には、他者がそれにアクセスできるとと、実際に反復されるかどうかにかかわらず反復可能であるとの前提が必要である。とくに論争のある場合には、異なる知覚者が、観察としての地位を確かめる (Haely 2003, 98-99)。知覚者の共同体が S である場合に当てはめると、この説では、他者が観察にアクセスし、それを検証できるかどうか、その観察の証拠としての地位にかかわる。ただし、それは、共同体が受け入れるものがそれゆえに真になるということではない (第2章第6節)。実験の結果が受け入れられるに至るのが、共同体のなかでの交渉によるのか、理由ある判断によるのかも争われており (第6章第2.4節)、共有された前提が批判を免れてしまうことの問題は、第11章第3節で取り上げる。

第二の筋、すなわち証拠をもたず経路や源泉の信頼性に移る。オールストンは、全体としての感覚知覚の信頼性は循環なしには示せないと論じる。水銀温度計で温度を読み取ることや、味でワインの年代を識別することのような限られた実践は、別の接近の仕方によって事実と照らし合わせて検査できる。しかし感覚知覚一般については、知覚に頼らずに事実に接近する仕方がない。知覚が可能にする予測と制御の成功からの議論も、より目立たない形で循環する。予測が当たったかどうか、観察によって確かめられるからである (Alston 1988, 436)。マルクス・ランメンランタ (Markus Lammenranta) は、オールストンの1993年の著書でのより詳しい扱いを次のように紹介する。知覚の実績からの議論は、結論が前提に含まれていないので論理的には循環していないが、認識的には循環している。結論を想定しなければ、前提を受け入れることが正当化されているとはみなせないからである。より高次の要件を課さないオールストンの信頼性主義の説明では、知覚が実際に信頼できるなら、そうした議論は正当化を与えることさえありうる。それでも、ランメンランタが引く一節でオールストンは、同じことが水晶玉占いについても言えると述べる。したがってこの議論では、無理なく信頼できる源泉とそうでない源泉とを見分けることができない (Lammenranta n.d., §1)。オールストンは1988年の論文ですでに同じ点を述べていた。ある実践の出力を用いてその実践を検査してよいなら、どの実践も満点をとり、この点ではすべての実践が同等になる (Alston 1988, 436-437)。ランメンランタは、ありうる診断の一つとして、こうした議論は正当化を与えることはあっても、信頼性についての疑いを理にかなった仕方では取り除くことはできない、と述べる (Lammenranta n.d., §1)。実績からの議論がここで循環するのは、それが帰納の形をとるからでもあり、二つの筋はここで重なる。統計の側でも、検定の判定を正当化する誤り率は、判定そのものが確かめられない分布と実験計画についての仮定に依存する (第2章第5.3節)。

オールストン自身の応答は、信頼性の証明ではない。ランメンランタが1993年の著書を紹介し引用するところでは、それは、確立した信念形成の実践に従事し、それを信頼できるものとみなすことの実践的な合理性の擁護である。「私たちが深く関与していると気づく実践を越えて訴える先はない」。それらは、手放したり置き換えたりするとしても、きわめて難しい実践である。すべての信念を保留することは選択肢にならず、代わりに新しい実践を採用しても、その実践にも循環しない擁護はない (Lammenranta n.d., §7)。1988年の論文でオールストンは、この位置づけがどのように失われ、また強められるかを述べていた。基本的な実践は、信頼できないことが示されるまでは一応受け入れられるべきであり、それを示しうるのは、その出力のなかの大規模で持続的な矛盾や、より確立した実践の出力との矛盾である。また実践は、知覚に基づいて達成される予測の成功のような一種の自己支持によって、受け入れられる資格を強めうる。この支持は、成功を確かめるのにその実践を用いなければならないので循環しないわけではない。しかし茶葉占いのようにそれを示さない実践もありうるので、取るに足らないものでもない。こうした検査に通った実践でも、後に問題が生じる可能性があるので、一応受け入れられるにとどまる。そして問題に突き当たった実践は、修正されて足並みをそろえ直すこともありうる (Alston 1988, 437)。ランメンランタは、オールストンが後にこの実践的な議論を退け、より単純な形のウィトゲンシュタイン的な文脈主義に落ち着いたことを記している。そして、実践的な合理性の擁護は懐疑論者に反しない、と反論する。懐疑論者が否定するのは、私たちの実践を用いることが実践的に合理的であることではなく、それが生む信念が認識的に正当化されていることだからである (Lammenranta n.d., §7)。

ウィトゲンシュタイン自身は、こうした前提を、扉がそれを軸に回る蝶番になぞらえた。ウィトゲンシュタインによれば、私たちが立てる問いと私たちの疑いは、ある命題が疑いを免れていること、いわば、問いと疑いがそれを軸に回る蝶番のようになっていことに依存しており、このことは科学的探究の論理に属する。そしてそれは、すべてを調べることはできないから想定で満足する、ということではない (Wittgenstein 1969, §§341-343)。この最後の点を人工のシステムに当てはめれば、次のようになる。前提が残るのは、調べる時間や能力が足りないからではない。前提は検査することの条件なので、検

査する能力がどれほど大きくなっても、なくなる。ただし、これは前提が固定されたままだということではない。局所的な前提は、別の経路で確かめ、改めうる（本節の終わりを参照）。

人工の研究システムにも、これらの前提に対応するものがありうる。ここでの対応は機能上のものであって、同一性ではない。また、システムの特徴のすべてがこの意味での前提になるわけではない。前提と呼べるのは、結果が証拠として何にかかわるかを左右するものである。学習によって形づくられるシステムでは、事例が同じであるという前提に対応しうるのは、たとえば学習された傾向性や表象である。それらは、入力がある安定した種類の事例であることを明示的に表象していなくても、入力をそうした事例として扱う。同じ筋には、ほかに次のようなものも属する。アーキテクチャと訓練がもつ帰納バイアス、ホールドアウトした評価データでの成績がさらなる事例についての証拠とみなされるための分布上の仮定、機器や自動評価器が安定しているとみなされるためのモデル、そして再実行が、同じ組織からの二度目の抽出ではなく反復とみなされるための条件などである。ヘイリーの説明をここに当てはめると、ひとつの基盤システムを再実行しても、それだけでは、疑われている観察が異なる位置にある知覚者によって確かめられたことにはならない。この点には第11章第3節で立ち返る。経路の信頼性の前提も、同じ形で現れる。ある自動評価器は別の自動評価器に照らして、またシステムの評価は人間の判断に照らして検査しうる。それが循環を避けるかどうかは、二つの経路がどう関係しているかによる。しかし、システムの内側から、その評価の実践全体の信頼性を支持する議論は、どこかでその実践の一部に依拠しなければならない。人間の判断が検査を担う場合にも、同じ構造は、それを含みより広い実践について、一段上で現れる。

これらの前提が誰について成り立つ必要があるかは、第2節の冒頭で述べたように、*S*によって異なる。また、前提がどのように据えられ、改められるかも違いうる。ウィトゲンシュタインの意味での蝶番は、ある命題が探究のある形態の内部でもつ身分であり、同じひとつの命題が、あるときには経験によってテストされるものとして、別のときにはテストの規則として扱われうる (Wittgenstein 1969, §98)。タルの事例で、適用の規則として定められた安定性も、規則として保持されながら、経験的な修正の余地を残す前提である。これに対して、帰納バイアスは、システムの作られ方が据え、作り直しによって置き換えうる構造である。人工の探究者が自らの前提や規則を形づくり、改めることもありうる。この違いは、そうした前提を誰が改訂できるかという、第7章第4.3節が認識論的自己統治に数える能力と、誰がその前提に責任を負うかという、第12章が取り上げる問いにかかわる。

最後に、二つの筋のどちらでも、前提が範囲において異なる。一方には、自然の斉一性や全体としての知覚の信頼性のように、どのような検査によっても一度に確かめることのできない前提がある。他方には、疑いが生じたときに他の経路で検査できる局所的な前提がある。知覚の様式は他の感覚によって、統計モデルはデータに照らして、標準は独立した実現に照らして、反復可能性の主張は他の観察者によって検査できる。前提が局所的であれば、その前提に頼らない点検にかけられることがある。ただし、範囲が狭いことだけでは、よりよい訂正は保証されない。訂正を進めるのは、第6章第5.2節でデューイ (Dewey) とチャン (Chang) に即して見たように、基準を探究の内部から改める循環を、実現のあいだの不一致に照らして繰り返し回すことである。全体的な前提はそれとは異なり、それを確かめる検査がどれもその前提に頼ってしまうので、同じ仕方では確かめ直せない。上で人工システムについて挙げた対応物でいえば、自動評価器が安定していることはこの意味で局所的な前提だが、システムの評価の実践全体が信頼できることは、そのシステムにとって全体的な前提である。そのため、構成の検査についての主張は、その前提のどれがどちらの側にあるか、つまりどれを構成が自ら確かめられ、どれには作り直しか構成の外からの検査しか届かないかを述べることができる。では、なぜそうした前提を置いてよいのか。上で見たオールストンの応答や、帰納は成功するかどうかを知らなくても正当化できるというライヘンバッハ (Reichenbach) の議論 (第2章第8.3節) は、これに答えようとする試みである。オールストンの応答は、確立した信念形成の実践が、私たちが容易には手放せないものであることに拠っている。オールストンは人工システムを論じていないが、作り手が作り直しによって帰納バイアスを置き換えうるシステム (上述) にこれを当てはめると、この前提がシステム自身の実践について成り立つのは、その実践が確立していて、システムがそれを容易には手放せない限りにおいてである。その条件がシステム自身の実践について成り立たない場合には、システムを作り評価する者たちを含みより広い実践について、その条件が成り立つかを別途問うことになる。そのため、そうした前提を誰が改訂できるか (第7章第4.3節) は、オールストンの根拠がどこにあるかにもかかわる。第2.2節で見る形成の歴史は、前提が世界と無関係に恣意的なものである必要はない理由を説明しうるが、それだけで根拠づけになるわけではない。前提への依存は、それだけでは、知識が不可能であることも、検査が信頼できることも示さない。局所的な前提にどの証拠経路がかかわりうるかは、第2.3節で扱う。ただしその前に、もう一つの条件がある。検査がある標的についての誤りを記録できるのは、検査される状態がその標的についてのものである場合に限られる。これが第2.2節の問いである。

2.2 接地はひとつの関係ではない

第2.1節は、検査がある標的についての誤りを記録できるのは、検査される状態がその標的についてのものであるときだけだ、という点で終わった。本小節は、そこで前提されていた問い、すなわちシステムの状態がそもそもいかにして世界の標的についてのものとなるのかを扱う。この問いが本章にかかわるのは、多くの説で、結果がある担い手にとって知識となるには、その担い手の状態が、結果のかかわる標的についてのものでなければならないからである。人間の社会のための研究では、それは人間の概念や機器の側で満たされうるが、人工の担い手が知っていると主張される場合には、その担い手の状態について問われる (第2節冒頭)。以下ではまず、第2.1節から引き継いで、原初的な組織が世界と無関係に恣意的なものである必要はない理由を形成の歴史に求める答へと、それが根拠づけとしてもつ限界を見る。そのうえで、ある説では歴史が、状態が何についてのものかをも定めることを見て、「接地」の名のもとで論じられてきた議論へ進む。記号接地問題は、記号をまず世界から切り離れたうえで立てられているが、実際に用いられる記号はそうではない。以下では、この点を正面から展開する谷口の議論を紹介し、そのうえで、記号と世界の結びつきを別のところに求める説を見る。最後に、「接地」の語がまとめてしまう関係を区別し、ひとつの関係における成功がそれだけでは他の関係を確立しないことを確かめる。

ひとつの答えは歴史から始まる。歴史は、原初的な組織が世界と無関係に恣意的なものである必要はない理由を説明しうる。人間の身体、知覚システム、発達、社会的学習、文化的実践は、後の探究が扱うのと同じ世界との長い相互作用を通じて形成された。ブラディとハームズ (Bradie and Harms) は、認知のふるまいを進化の産物として理解するのが難しい理由の一つとして、少なくとも三つのまったく異なる進化的な過程がかかわることを挙げる。遺伝的な継承を通じた認知と知覚のメカニズムの生物学的進化、言語と概念の文化的進化、そして個体の生涯のうちでの試行錯誤による学習である。そのうえで、「進化論的認識論」という語をつくったドナルド・T・キャンベル (Donald T. Campbell) に同意するいくらかの理由があると述べる。人間の知識を

十分に理解するにはこれらの過程の相互作用を理解する必要がある、というのがキャンベルの考えである。ブラディとハームズはさらに、そのためには生物学的進化と文化的進化の両方をモデル化できなければならないとして、文化的進化のモデルの検討に進む (Bradie and Harms 2020, §§1, 2.2)。

ここで取り上げる説では、歴史による答えがもっともらしいのは、程度の差を認め、能力や領域ごとに成り立つ形においてだけであり、歴史は、それが説明を助けるもののすべてを根拠づけることはできない。歴史による答えのいくつかの版は、認知メカニズムが環境の構造によってどのように形成されるかを扱う、進化論的認識論のプログラムに属する (Bradie and Harms 2020, §1.1)。ブードリ (Boudry) とフレリック (Vlerick) は、自然選択は単に真理に無関心であるという包括的な主張を退ける。二人が相手とするのは、アルヴィン・プランティンガ (Alvin Plantinga) の、自然主義に反対する進化論的論証である。ブードリとフレリックの再構成によれば、この論証は次のように主張する。自然選択が気にかけるのは適応度だけであって、真理ではない。そのため、自然主義、すなわち神や超自然的な存在はいないとする見方と、導き手のない進化とがともに真であるなら、私たちに自らの認知能力を信頼する理由がなく、自然主義者は、自然主義そのものへの信念を含め、すべての信念を手放す理由を得ることになる (Boudry and Vlerick 2014, 65–66)。これに対してブードリとフレリックは三つの論証で応じる。第一に、生存そのものは信念をまったく必要としないが、いったん信念によって行為を導くという経路をとった系統は、その後もその経路に縛られる。二人はこの拘束を適応的な経路依存性と呼ぶ。その段階からは、生態学的に重要な状況で少なくともおおよそ真である信念が有利になり、偽である信念は害をもたらす (Boudry and Vlerick 2014, 68–69)。第二に、信念が何についてのものであるかは、欲求やほかの信念と組み合わせさせてその信念が生物に促す行為と結びついていると論じる。したがって、行動への効果が同じままで内容だけが変えることはしない。そのうえで、その行動が適応的であるかどうかは、世界がその信念の表すとおりであるかどうかによって依存する。たとえば、あるキノコは食べても安全だという信念は、そのキノコが有毒であれば害をもたらす (Boudry and Vlerick 2014, 69–72)。第三に、科学的信念は直観的な能力をそのまま働かせただけの産物ではないと論じる。科学的信念は、同時代の人々のあいだでの、また世代をこえた協働に支えられ、さらに数学と記法、機器、そして理論を観察と照らし合わせることに支えられている。不完全な基礎の上に築かれたものはすべて信頼できないはずだとする推論を、ブードリとフレリックは基礎づけ主義の誤謬と呼ぶ (Boudry and Vlerick 2014, 72–74)。

ブードリとフレリックはまた、自然選択が真理を有利にするとはかぎらない場合も挙げる。ヒューリスティック、すなわち手早い経験則は、速さのために正確さを犠牲にしうる。誤りの種類によって代償が等しくないため、ある種の誤りをかえって生みやすいメカニズムもありうる。さらに、慣れ親しんだ環境では正確な能力が、新しい環境では誤りに導くこともある。そのため、自然選択は、ある領域では真なる信念を有利にしても、別の領域ではそうしないかもしれない (Boudry and Vlerick 2014, 66–67)。フレリックとブロードベント (Broadbent) は、進化論的懐疑論に対するそうした応答を受け入れながらも、それが科学的信念の一般的な正当化を与えることを明示的に否定する。二人の説では、自然選択が認知能力を真理を追跡するように形づくったのは、生態学的に重要な信念という狭い領域においてだけであり、現代の科学理論はそうした信念の延長ではない (Vlerick and Broadbent 2015, 186, 190–192)。

以上をまとめると、世界との相互作用は、ある能力が自由に発明されたのではなく現実の規則性に応答している理由を説明する助けとなりうる。しかしそれは、結果として生じる予期のすべてについて、それが真であること、一般に信頼できること、正当化されていること、あるいは形成された範囲を超えた環境にも適していることを確立するものではない。ブードリとフレリックも認めるように、適応的な近道や体系的な誤りはありうる。また、高度な探究は、この相互作用が形づくった能力だけに依存しているのではない。これらの説に立てば、こうした自然の来歴は、認識的能力がどこから来たかを語る系譜ではありえても、その能力が生むすべての産物の根拠づけではない、ということになる。自然の来歴にもとづく議論に当てはめると、オールストンの分析は、そうした議論が一般的な根拠づけを与えることができない、さらなる理由を与える。自然選択はおおむね真なる信念を生む信念形成の傾向性だけを残したはずだという議論は進化論に依拠し、進化論は知覚を通じて集められた証拠に基づいている。オールストンは、予測と制御の成功からの議論を、その成功自体が観察によって確かめられることから循環とみなした (第2.1節, Alston 1988, 436)。この基準に従えば、自然選択からの議論も、知覚の信頼性を支持する議論としては、その前提の当否がどうであれ、認識的に循環していることになる。

では、人工のシステムにも、これに対応する歴史はあるのか。抽象すれば、ここまでの議論が言うのは、探究者の原初的な組織はその探究者の形成の歴史に形づくられる、ということである。人工のシステムにも形成の歴史はある。しかし、その歴史が三つの過程に対応するものを含むかどうかは自明ではなく、過程ごとに事情が異なる。比較的たどりやすいのは、試行錯誤による学習との類比である。自らの行為の結果から学ぶシステムでは、それに近いものが働きうる。

生物学的進化に対応するものがあるかどうかは、それよりはるかにはっきりしない。進化論的認識論のうち認知のメカニズムの進化を扱うプログラムが対象とするのは、脳、感覚系、運動系といった、認知の生物学的な基盤である (Bradie and Harms 2020, §1.1)。この過程は、すでにある器官の働きを調整するだけでなく、感覚器官や神経系そのものを生み、作り変えてきた。何がそもそも感知されるか、どのような区別がそもそも可能かも、その過程のなかで変わってきたのである。第2.1節で見た、知覚と行為に組み込まれた予期は、この見方では、そうした過程に形づくられたものである。これに対して、現在の訓練されたシステムの多くでは、アーキテクチャや感覚入力の仕事 (センサーや入力の形式) は人間が設計し、学習はその上で行われる。設計の選択や訓練が作り込む構造を、この役割の一部を担うものとみなす余地はあるかもしれない。しかし、それらを生物学的進化に対応するものとみなすと、器官そのものが生まれ、作り変えられてきた過程を、与えられた構造の調整に縮めてしまうおそれがある。人工のシステムで何がこの過程の役割を果たしているのか、そもそも果たしうるものがあるのかは、定かではない。

文化的進化との対応も、簡単ではない。ブラディとハームズは、文化的進化を、言語や概念が集団のなかで伝えられ、あるいは退けられる過程として集団モデルで扱い (そこでの適応度の成分には、伝達の確率と拒否の確率が含まれる)、言語や教育制度のような文化的なニッチの構築もそこに数える (Bradie and Harms 2020, §2.2)。そうであれば、問いは一つのシステムがそれを経るかどうかではなく、システムがどのような集団のなかで概念の伝達と選別に加わるかであり、その答えは、システムがどのような社会のなかにあるかによって変わりうる。人間の記録にもとづく訓練は、人間の文化的進化の産物を受け継ぐことではあるが、受け継がれるのは記録に残るものであり、実践に加わることで身につく集合的暗黙知のようなものは含まれないかもしれない (第3章第6節)。他方で、システムの出力が人間の言語と記録に入り、後の人間やシステムが受け継ぐものの一部になるなら、システムは受け継ぐだけでなく、人間の文化的進化に加わっている。人工のエージェントのあいだで語彙が伝えられ安定するなら、人工社会それ自身の文化的進化に似たものも考え

られる。ただしそのときには、何が語彙を選別するのが問題になる。この点は、以下でブシャクール（Bouchacourt）とバローニ（Baroni）の実験に即して見る。

システムが人間の記録から受け継ぐ場合には、その組織は、それ自身の世界との相互作用によってではなく、人間の歴史に媒介されて世界に結びついていることになる。この媒介されたつながりが状態に何を与えるかは、以下で見るコエーリョ・モロとミリエールの説明と、第2.3節で見る記録からの継承という経路にかかわる。また、上で見た限界は、人工システムの歴史にも当てはめうる。訓練の歴史が、ある能力が記録の規則性に応答している理由を説明するとしても、それは、その能力が訓練の範囲を超えた環境に適していることも、その産物が根拠づけられていることも確立しない。なお、ここで述べたことの多くは現在の訓練されたシステムについてのものである。自らの設計や感覚入力の仕事みを、世界との相互作用の結果によって作り変えていくような人工の系統がありうるなら、事情は異なりうる。

歴史は、本小節の問いに別の仕方でもかかわる。上で見たブードリとフレリックの第二の論証は、信念の内容をそれが促す行為と結びつけ、行為にかかる選択が内容にも及ぶとする点で、すでに歴史と内容とのかかわりを含んでいる。以下で論じるコエーリョ・モロとミリエールの説明は、目的論的意味論（teleosemantics）にもとづく。これは、状態が何を表すかを、その状態が選択の履歴を通じて担うようになった機能から説明する表象の理論である。この説明では、学習が供給しうる選択の履歴は、内的状態に、世界の標的についての情報を担う機能を与えるものの一部である（Coelho Mollo and Millièrè 2026, 13–17）。他方で、接地された状態は、それだけでは知識に十分ではない（Coelho Mollo and Millièrè 2026, 2, 24）。そのためこの説明では、歴史は、システムがその状態の対象について真とみなすことを根拠づけないうまま、状態が何についてのものかを固定しうる。形成の歴史が根拠づけにもかかわるかどうかはさらなる問いであり、正当化の説明によって答えが分かれる。

「接地（grounding）」の名のもとでは、異なる問いが立てられてきた。ハルナッド（Harnad）の記号接地問題が問うのは、形式的な記号の意味論的解釈を、どうすればシステムに内在するものにできるかである。書物のなかの記号の意味は、それを読む人の頭のなかの意味から派生している。そうした派生的な意味ではなく、システム自身に内在する意味をどう与えるか、という問いである（Harnad 1990）。ベンダー（Bender）とコラー（Koller）は、言語の形式だけで訓練されたシステムは、形式と伝達意図との関係として理解された意味を学習できないと論じる。二人の説では、伝達意図は言語の外にある何かについてのものであり、テキストはその意図が形式へと投影されたものである。そのため、訓練データが形式だけであれば、その関係を学習するのに十分なシグナルはない。意図について仮説を立てて検証する手段がなければ、形式だけから意図を再構成することは望みがない（Bender and Koller 2020, 5185, 5187, 5189）。これらの議論は、テキストの予測だけで訓練されたシステムに厳しい問いを突きつけるが、それだけでは、信念、知識、あるいは研究についての問いを決着させない。

どちらの議論にも、著者自身が述べる限定がある。ハルナッドは、厳密に言えば、記号接地はAI一般の問題ではなく認知をモデル化することの問題だとした。記号システムだけでAIが求める性能が得られるなら、その記号に内在的な意味を求める理由はない。ただしハルナッドは、接地がロボティクスや機械視覚で性能上の利得をもたらすことは認めていた（Harnad 1990, n. 10）。ベンダーとコラーは、自らの議論を形式だけで訓練されたシステムに限る。二人は、読解のように、データがテキストどうしの意味関係を指定している課題で訓練されたシステムには、その議論が当てはまらなと述べる。また、形式に接地データを加えれば、そのデータに伝達意図が表されている程度に応じて、意味が学習されることもありうるとする（Bender and Koller 2020, 5191–5192）。二つの議論をこう並べてみると、人工研究者が、ハルナッドの意味（システムに内在する意味論的解釈）とベンダーとコラーの意味（形式と伝達意図との関係）のどちらで接地されている必要があるかは、接地が何を説明するために持ち出されるかによる。第1章第1節が述べるように、候補となる定義は異なる問いに答えるからである。それはまた、システム自身が知っているかと主張されるのか、それとも他者が知るようになる結果を供給するだけなのかにもよる（第2節冒頭）。

これらの議論は、記号をまず世界から切り離したうえで、それがどのようにして意味をもつかを問う、という形で立てられている。ハルナッドの問いは、記号主義的なAI（記号を規則によって操作する古典的なAI）のように、形式的なトークンとしての記号を先に用意し、その解釈はシステムの外の解釈者が与える、というシステムについて立てられた。ベンダーとコラーの議論も、言語の形式を、それが表す意図や世界から分けたうえで、形式だけから何が学べるかを問う。しかし、私たちが実際に用いる記号は、世界や他者とかかわりと無関係にあるわけではない。佐治（Saji）が紹介する発達研究によれば、言葉は、物や出来事にともに注意を向け、他者とやりとりするなかで学ばれる（Saji 2026, 46–47）。そして言葉は、使われるなかで意味を変え、それがまた私たちの知覚や行為を方向づける（Taniguchi 2026b, 7–8）。記号と世界をはじめに切り離せば、両者をどう結びつけるかという問いが生じる。両者を相互作用のなかでともに形づくられるものとして見れば、問いは、その結びつきがどのように形づくられ、保たれるかに変わる。ハルナッド自身も、記号を切り離したままにするのではなく、感覚入力にもとづく、記号ではない表象（図像的表象と範疇的表象）に記号を下から結びつけることに、解決を求めていた（Harnad 1990）。記号と世界を相互作用のなかでともに形づくられるものとして見るこの見方を、正面から展開しているのが谷口である。

谷口も、記号接地問題を、記号主義的なAIの文脈で、形式的なトークンとしての「記号」について立てられた問いとみる（Taniguchi 2026c, 42–43）。そのうえで、人工的に作られた記号をいかに現実世界の経験に意味のある仕方で結びつけるかという問いの前提そのものが誤っているかもしれない、と考えるにいたったと述べる。記号は外的な表現であるだけでなく、環境との感覚運動的な相互作用を通じて自己組織的に発達する、概念のような内部表象と深く絡み合っている、というのがその理由である（Taniguchi 2026a, v–vi）。記号創発ロボティクスの観点からは、AIの中核にある課題は記号接地問題ではなく、意味をもつ記号が社会的で身体的な相互作用を通じてどう生まれ、変化し、組織されるかという記号創発の問題だとされる（Taniguchi 2026c, 43）。この組み替えは、ハルナッドの論点を退けるものではない。谷口は、語と語の関係が与えるのはせいぜい関係的な概念であって接地された概念ではない、というハルナッドの指摘を受け入れている（Taniguchi 2026d, 198）。

谷口が記号創発システム論と呼ぶ枠組みでは、各成員は、他者の内部に直接アクセスできないまま、環境との身体的な相互作用を通じて内部表象を形成する。成員どうしの記号的なやりとりを通じて、言語のような記号体系がボトムアップに組織され、組織された体系は成員の知覚、行為、解釈をトップダウンに拘束する。谷口は、このマイクロ-マクロのループが記号の意味を生むとする（Taniguchi 2026b, 4–8）。内部表象はそれ自体ではまだ記号ではなく、記号形成の素材であり、語の意味は、個体の内部だけで決まるのではなく、社会的な文脈のなかで他者に解釈されることを通じて得られる（Taniguchi 2026b, 7）。個々の成員にとっての接地も、客観的な世界ではなく、ユクスキュル（Uexküll）のいう環世界、すなわちその生き物が知覚し行為する主観的

な世界においてなされるとされる (Taniguchi 2026d, 198–199)。成員の条件は、生物学的に人間であることではなく、恣意的な（対象との自然な結びつきをもたない）記号を用いて社会の機能に寄与するかたちで、記号的なコミュニケーションに参加できることである (Taniguchi 2026b, 6)。谷口は、AIシステムが記号創発システムに有意義に参加していると言えるのは、身体をもち現実世界と相互作用するときだけだとする (Taniguchi 2026c, 42)。他方で、生成AIがすでに人間の言語使用を変え、このシステムの一部分になりつつあるとも述べる (Taniguchi 2026a, v; Taniguchi 2026g, 222, 225)。

ループがどう成り立つかについて、谷口らは命名ゲームのモデルを示している。二者が同じ対象に注意を向けているという前提のもとで、話し手が自らの観察にもとづいて名前を提案し、聞き手が自らの観察から形成した内部表象に照らしてそれを確率的に受け入れる。受け入れの確率をメトロポリス・ヘイスティングス法という所定の規則に従わせると、このやりとりは、両者の観察を統合する共有の潜在変数についての近似的なベイズ推論と形式的に等価になる。つまり、互いの内部表象を直接参照しなくても、双方の観察を共有の記号の形成に反映できる (Taniguchi 2026e, 207–208)。谷口はここから、人間の言語はこうした分散的な推論を通じて形成されるという、集合的予測符号化 (collective predictive coding) の仮説を提示する (Taniguchi 2026f, 211–212)。この枠組みを、以下で区別する関係に当てはめて読むこともできる。そう読むと、この枠組みは、知覚と行為への依存（感覚運動的な関係）と、話し手のあいだでの共通基盤の確立（コミュニケーション的な関係）を、一つのループの二つの面として扱う。そして、指示（記号や状態が世界の何を指すか）の関係を、客観的な世界のうちではなく、各成員の環世界と、成員のあいだで共有された環世界とのうちに置き直す (Taniguchi 2026f, 212)。

この組み替えは、本書の区別で述べることができる。記号の意味がこうしたループのなかで成り立つのであれば、人工システムに問われるのは、その状態がどのような関係によって世界の標的についてのものとなるかだけでなく、どのような記号創発システムに、どのような成員として加わるかでもある。文化的進化について上で述べた問いは、この見方では意味そのものの問いになる。ここでは、研究を遂行する *A*、新しさと誰のための達成かを定める *S*、そして *A* が加わって記号体系を形づくる共同体を、混同しないことが要る。その共同体は、人間の社会でも、混合社会でも、人工社会でもありうるが、*S* と一致するとは限らない。本書は *A* と *S* を区別しているので、共同体をそれらとは別の項として置けば、この食い違いを述べることができる。これは、第3節が検討するミスマッチの源の一つである。プレプリントで科学AIの自律性のレベルを提案する高橋 (Takahashi) は、仮説生成の側からこうした食い違いに行き着く。その説明では、形式があらかじめ固定されていない仮説の生成、つまり高橋のいうレベル4は、実験科学においては、少なくとも実験が扱う対象の領域で記号接地問題が解消されることを要する。そしてそれは、システムが特定の実験ロボットを介して身体性を獲得することを意味する。そうであれば、人間の身体性に結びついて構築された記号体系とは別個の記号体系が生まれるので、このレベル以降のシステムは人間とは別のやり方で対象を理解することになり、人間の科学とAIの科学の分岐はここから起こる、と高橋は論じる (Takahashi 2024, 7, 10)。これに対して、人間が研究目的を与えるレベル5では、システムは人間の意味のネットワークと互換な記号体系を保つ必要がある。そのため、システムの知識を人間の科学の知識体系に接地し、あるいはそれへと変換し続ける何らかの仕組みを、構築し維持しなければならない (Takahashi 2024, 8, 11)。本書の用語で読めば、高橋のレベル4は、*A* が用いる記号体系が人間の社会である *S* の記号体系から分かれていくことを許し、レベル5は、研究がその *S* のための研究であり続けることを求めている。ただし、この見方をとっても、感覚運動的な関係とコミュニケーション的な関係という二つの関係がひとつのループで結びつけられているときに、一方での成功が他方を確立するかどうかは、ループの仕組みによって異なりうる。共有された記号が世界の標的についてのものとなるか、またその標的による訂正に開かれているか（第2.3節）も、ループの仕組みに即してなお問われる。この点の一例は、下でブシャクールとパローニの実験をこのモデルと並べるときに現れる。

記号と世界の結びつきをどこに求めるかについては、ほかに説がある。コエーリョ・モロとミリエールは、指示的、感覚運動的、関係的、コミュニケーション的、認識的な接地を区別し、指示的接地をこの問題にとって本質的なものとして扱う。目的論的意味論に依拠して、二人は、指示的接地が二つの条件を要求すると論じる。世界への因果的・情動的な関係（言語モデルの場合、これは人間の言語産出を経由する）と、内的状態にその情報を担う機能を与える選択の履歴（これは学習が供給しうる）である。そして、言語モデルはこの両方の条件を満たしうる、とりわけ選好にもとづく調整 (preference tuning) を経たものについてはそう言える」と論じ、また、マルチモーダリティと身体性は指示にとって必要でも十分でもない」と論じる (Coelho Mollo and Millière 2026, 1, 13–17, 22, 24)。コエーリョ・モロとミリエールはまた、設計者の意図が与えるのは学習の仕組みの機能であって、内的状態の機能と、その状態が正しいかどうかの基準は、その仕組みがデータの上で働いた選択の履歴が与えるとする。二人はこれを、穏やかな犬を得ようとする育種家の意図と、実際に形質の機能を定める繁殖の履歴との関係にたとえている (Coelho Mollo and Millière 2026, 16)。この区別は、第7章第2節が研究の遂行の帰属について、機能の出どころはどこかという問いと、探究の組織化がどこにあるかという問いとを分けたことと並行している。この説明では、設計者が学習の仕組みに込めた意図は、それだけでは、状態が何についてのものかを定めない。

コエーリョ・モロとミリエールもまた、自らの主張の限界を示している。二人は、言語モデルが理解している、知っている、あるいは真偽の概念をもっては論じない。そして、接地された状態だけではそうした能力に十分ではないとする。コエーリョ・モロとミリエールは伝達意図の問題を脇に置き、ペンダーとコラーの問題提起をコミュニケーション的な接地に結びつける一方で、ペンダーとコラーに反対して、言語の形式は意味から容易に切り離せるという見方を退ける (Coelho Mollo and Millière 2026, 2, 5 n. 9, 8, 13, 24)。したがって、二つの立場を並べて読むと、両者が衝突するのは、テキストの予測だけでシステムを訓練することから何が得られるかについてであり、そうしたシステムが伝達意図をもつかどうかについてではない。ペンダーとコラーは、そうした訓練を形式だけによる訓練とみなし、それによつては、表現を言語の外的対象に結びつける慣習的な意味も、形式と伝達意図との関係も得られないとする (Bender and Koller 2020, 5185, 5187)。コエーリョ・モロとミリエールは、形式を意味から容易に切り離す見方を退けることで、テキストの予測だけの訓練を形式だけによる訓練とみなす、この分類そのものに疑問を投げかける。コエーリョ・モロとミリエールが挙げる選好にもとづく調整のように、テキストの予測を超える信号を加えて訓練する場合に、ペンダーとコラーの議論がなお当てはまるかどうかは、二人の2020年の論文だけでは決まらない。

認知をシステムの自己維持に根ざすものとみるエナクティブなアプローチは、指示的接地より強いものを求める。ディ・パオロは、システムの自己産出であるオートポイエーシスは、それだけでは、生き続けるか崩壊するかという全か無かの規範しか与えず、システム自身はその規範を感知できないと論じる。意味形成 (sense-making) には、それに加えて適応性 (adaptivity)、すなわち自らの状態、および自らと環境との関係を、それらが自らの存続可能性の境界に近づいているかどうかに応じて監視し調整する能力が必要である (Di Paolo 2005, 6–10)。ディ・パオロは人工システムの場合を一方方向に決着さ

せはしない。その論文によれば、意味のある行為は、人工のシステムが意味をもって行為していなくても、そのシステムによって説得的に模倣されうる。したがって、システムが意味をもって行為しているかどうかは、その応答を観察するだけでは分からず、システムの組織が自らの存続可能性との関係で自前の意味形成を生み出しているかどうかを調べて確かめるほかない。この確認は、少なくとも否定的な場合を除外するには比較的簡単だとされる。ディ・パオロは、完全に代謝するロボットに至らなくても人工システムの志向性が可能かどうかには、答えを出さない。ここでいう志向性は、同論文の意味では、行為者自身の行動がもつ志向的な構造である。そのうえでディ・パオロは、代謝に由来しない価値を考えに入れると見通しが変わると述べる。そして論文の結びの、自ら思弁的と呼ぶ考察のなかで、感覚運動の協調の組織化されたパターンのような自己維持的な活動の形態が、代謝による存続可能性の範囲の内で、固有の価値を生み出しうると論じる (Di Paolo 2005, 13–16)。これに対してローラ (Rolla) は、AIは自らの生きた身体をもたず、環境のなかに住まってもいないので、生存可能な範囲にとどまるように物質やエネルギーを自律的に交換することができないと論じる。そして、真に認知的なAIをつくるには、まず人工生命を実現する必要があるとする (Rolla 2026, 1, 11, 14)。ローラの説明では、記号が内在的に重要なものとなるのはこの種の自律性を通じてであり、私たちが用いる記号が接地されているのは、それが私たちの自律的なふるまいの産物として、私たちににとって重要だからである (Rolla 2026, 6, 11–12, 14)。このエナクティヴな要求は、コエーリオ・モロとミリエールの説明に対抗する実質的な説であって、人工システムによる研究の理論中立的な条件ではない。これらの説明で問題となる自律性、すなわち自らの活動が自らを産出し維持するシステムの自律性は、第7章第4節の、比較としての自律性ではない。後者は、特定の点において、選ばれた境界の外の当事者が、行為者に帰属される組織化をどこまで方向づけ、代替し、授権し、承認し、あるいは修復するかを比較する。一方についての判定は、他方を決着させない。

論証が接地のいずれかの意味に依存する場合には、次の六つを区別しておく。最初の五つは、表現や状態と何かとの関係であり、第六は、それらに加えて求められうる条件である。

1. 指示的接地：世界の標的への指示。
2. 感覚運動的接地：知覚と行為への依存。
3. 関係的接地：言語の内部における、ある表現と他の表現との関係。
4. コミュニケーション的接地：話し手のあいだでの共通基盤 (common ground) の確立。表現を世界に結びつける共有された仕方もそこに含まれる。
5. 認知的接地：知識ベースに蓄えられた情報への表現の結びつき。
6. 自己維持とその適応的な調整に接地された、エナクティヴな、あるいは内在的な規範性。

最初の五つは、コエーリオ・モロとミリエールが、認知科学者、哲学者、AI研究者の議論のなかに見いだした接地の概念であり、第六は、上で述べたエナクティヴな要求を加えたものである (Coelho Mollo and Millière 2026, 6–8)。この六つを並べる目的は、接地を定義することではなく、この語がまとめてしまう関係や条件を切り分けておくことにある。たとえば、第1.3節で触れた現象学的な要求は、この六つに含まれていない。コエーリオ・モロとミリエールは、指示的接地を五つのうち最も基礎的なものとし、関係的接地を部分的な例外としつつ、指示的接地がおそらく他の接地の基盤でもあるとする (Coelho Mollo and Millière 2026, 24)。上で見たエナクティヴな説明は、システム自身にとって意味が生じる条件を自己維持とその適応的な調整に求めるので、この六つの並びのなかでは、指示的接地の成立だけで意味形成を説明できるかを問う立場にあたる。コエーリオ・モロとミリエールの第五の概念は、人工のシステムについて工学の文脈でこの語が使われる用法の一つ、すなわち検索で取り出した情報への結びつきに近い。その説明では、現在の言語モデルは、情報を検索して取り出す外部のデータベースに接続されているとき、認知的接地をともしない (Coelho Mollo and Millière 2026, 8)。証拠と訂正に対する応答可能性は、これら六つのいずれとも種類の異なる関係である。状態は、ある標的についてのものでありながら、その標的を誤って表象しうる。そして、システムのコミットメントが誤って表象しているときに訂正されるかどうかは、さらなる問いである。第2.3節は、これを世界への応答可能性として取り上げる。

ひとつの関係における成功は、それだけでは他の関係を確立しない。これは、関係どうしが互いに独立だという主張ではない。どの関係がどの関係を支えるかは、説によって異なる。コエーリオ・モロとミリエールの説明では、言語システムは、それ自身の感覚運動的な技能を保有することなしに、人間のテキストに媒介された因果的關係と、その情報を担うように状態を選択する学習の履歴とを通じて、世界の標的を指示しうる。ロボットは、信念を形成したり科学的批判に参加したりすることなしに、知覚と行為を結合しうる。感覚運動的接地が確保されているところでも第2.3節の問いが残るのは、そのためである。共同体は、成員どうしの協調を首尾よく支える語彙を安定化させうるが、協調の成功は、その語彙が何についてのものかを定めはしない。人工エージェントのあいだの創発的コミュニケーションについての実験は、この最後の可能性に具体的な形を与える。ブシャクールとバローニは、一方のエージェントが二枚の画像のどちらが標的かをシグナルで伝え、他方がそれを選び出さなければならないゲームで、ネットワークの対を訓練した。成功した対は、二者のあいだではほぼ完全に揃っている一方で、入力画像の表象とは弱くしか相関しない表象を発達させた。しかもそれらの対は、入力が画像ではなくランダムなノイズの対であっても、画像の場合とほぼ同じくらい首尾よく伝達した。著者らは、シグナルが追跡していたのは画像に描かれた物の種類ではなく、入力の低水準の関係の性質だったと結論する (Bouchacourt and Baroni 2018, 983–985)。シグナルは、ゲームの標的を確かに識別していた。したがって、そうした共同体の内部で協調が成功していることは、それだけでは、共有されたシグナルが、観察者の目にそれが名指しているように見えるものについてのものである、という証拠にはならない。これに対して谷口らの命名ゲームでは、語彙を選別するのは協調の成否ではなく、聞き手が自らの観察から形成した内部表象のもとでの尤度である。谷口自身は、このゲームを、共同注意を前提する点で、従来の命名ゲームや指示ゲームから区別している (Taniguchi 2026e, 205, 207)。ただし、上で述べた形式的な等価性は、共有された記号が観察者の名指すものを追跡することまでは示さない。二つを並べると、共有された語彙が何を追跡するかは、語彙の安定化だけでは決まらず、何が語彙を選別するかによって異なりうることを示唆される。人工の共同体では、何が語彙を選別するかは、成員のあいだのやりとりをどう作るかによって定めうる。他方で、人間の観察者が想定する種類を追跡しないことは、それだけでは世界とのつながりの欠如を示さない。そして、標的についての誤りによってコミットメントが訂正されうるかは、第2.3節の問いとして残る。

2.3 世界への応答可能性

第2.2節は、標的についての誤りによってコミットメントが訂正されうるかという問いで終わった。また第2.1節では、局所的な前提が、いま吟味している仮定だけに頼らない証拠によって訂正されうることを見た。認識的組織のコミットメントは、いかにしてその標的による訂正を受けうるものであり続けるのか。そして、 S が人間の社会である場合、混合社会である場合、人工社会である場合（第1.1～1.3節）のそれぞれで、誰がその関係を維持するのか。ここでの用法では、世界への応答可能性とは、二つのことを意味する。一つは、認識的組織のコミットメントが標的となる現象に制約されていること、それも、その現象のあり方が違ってればコミットメントも違ってくるという仕方でも制約されていることである。もう一つは、さらなる接触、介入、独立した経路による相互検査を通じて、組織が是正可能であり続けることである。この意味で世界に応答可能であることは、神託に問い合わせられることではない。この考えは本書の先の箇所にも現れている。最終的な真理についての不可謬のテストなしに、証明、観察、介入、外れた予測、批判に応答可能であることとして（第2章第3節）、また、修正そのものとは別の根拠で評価される、実現のあいだの不一致に応え続ける基準の修正として（第6章第5.2節）である。第11章第1節は、そうした経路を無効化の関係として特定し、それが加える証拠と、それを保護する仕組みとを、失敗の診断が異議に開かれ続ける仕方も含めて分ける。本書が目標や人々への応答可能性について語るときには、意味が異なる。

「世界への応答可能性」という言い回しは、ここでは、マクダウェルが世界に応答可能な思考に与える意味では用いていない。マクダウェルにとって、その応答可能性は合理的な関係であり、そこでは判断は、物事がどうあるかに応じて正しく、あるいは誤ってなされる。そしてそれは、主体の概念的能力がすでに働いている経験が、主体の考えることに対して、単に因果的な関係ではなく合理的な関係、すなわち経験がその考えを支持したり退けたりする理由となる関係に立つことを要求する。マクダウェルは、およそ経験的内容があるためにはこの要求が成り立たなければならないと考える (McDowell 2000, 3-5, 9-12, 16)。マクダウェルの説は人工の組織についてのものではない。それでも、ここで用いる定義に照らしてこの説を読めば、その定義を因果的・実践的な経路だけで満たす組織は、それによってこの説の意味で世界に応答可能になるわけではない、ということになる。そうなりうるかどうかは、その状態が、マクダウェルが概念的と呼ぶ種類の能力に依拠しているかどうかにかかっている。そのため、二つの用法は同じ組織について判定を分けうる。介入の結果によってコミットメントが改められる人工の組織は、本章の定義では世界に応答可能でありうるが、マクダウェルの説では、それだけでは、訂正を受けているのがその説の意味での思考であるかどうかが決まらない。分かれ目は、訂正の経路があるかどうかではなく、訂正を受けるものが何であるかにある。第2.2節で見た接地の諸説は、状態が何についてのものかを扱うが、マクダウェルのいう概念的な能力の条件までは扱わない。

したがって、「自然からのフィードバック (feedback from nature)」は、ここでは控えめな概念である。自然は解釈されていない評決を返すわけではない。介入は、装置、サンプリング、分類、モデル、標準を通じて痕跡を産出する。そして不一致は、焦点となる主張、補助的なコミットメント、機器、物質的条件、あるいは結果を記述する表象にかかわりうる（第6章第2.1節）。そうした痕跡を読む基準が、応答可能性そのものを S に相対的なものにするかどうかは、説によって異なる。何がコミットメントを訂正するかが、 S の認識的規範とは独立に、その標的によって定まるとする説では、それらの規範は何が証拠とみなされるかにかかわりうる（第2.1節）が、世界への応答可能性は S に相対的ではない。これに対して、ライボ (Leibo) らは、AIアラインメントについての立場表明のプレプリントのなかで、ふつうなら事実と呼びたくなる命題でさえ、正当化の基準と証拠の適切な統合を律する認識的規範のもとにあり、それらの規範は「他のどの種類の規範とも同じく文化的に偶然的」だとする (Leibo et al. 2025, 2, 7)。ライボら自身は論じていないが、この見方を世界への応答可能性の分析に持ち込むと、標的によって訂正されるとはどういうことかも、部分的には共同体の規範によって定まることになる。ただし、そこから、第3節が依拠する、世界への応答可能性と S にとっての関連性との違いが消えるわけではない。証拠の基準が社会的であることと、何を知らたいかが社会に相対的であることとは、別の事柄だからである。同じ証拠の規範をもつ社会どうしが、何を追究に値するとみなすかで分かれることもあり、同じ課題を重要とみなす社会どうしが、異なる検査の規範を採ることもある。この見方のもとで弱まるのは、二つの違いが成り立つことではなく、何が訂正にあたるかを共同体の規範から独立に定められるという想定である。そのかぎりでは、第3.2節で見るように、二つの離れ方はそれほど鋭くなくなる。

人間の知識を目指す研究にとっては、世界との原初的な関係がすべて人工の構成要素の内部にある必要はない。人間の概念、証言、データセット、機器、評価者、制度が、活動から世界への媒介されたつながりを提供しうる。ハイブリッドな組織にとっては、その関係は組織化された全体に属するかもしれない。ある構成要素が感知し、別の構成要素が介入し、別の構成要素が記録を保存し、また別の構成要素が帰結を評価する。混合社会については、人工の成員が知ることになる限りで、その関係はそれらの成員について成り立たなければならないだろう（第2節冒頭）。ヘイリーが伝えるロンジーノの説では、他の成員が観察にアクセスし、それを検証できるかどうか、その観察の証拠としての地位にもかかわりうる（第2.1節）。

テルミーネ (Termine)、ラッチィ、ファッキニー (Facchini) による科学における機械学習の分析は、検査が前提するもの自体が、そのような組織のなかでいかに分散しうるかの一例である。三人は、科学における機械学習のモデル構築はおおむね「理論無関心 (theory-indifferent)」であると論じる。すなわち、モデルのパラメータやハイパーパラメータを定めることにも、そして次第に特徴量の設計にも、モデルが用いられる分野の領域理論への参照は必ずしも要らない。少なくとも、訓練データと同じように分布するデータでの予測の正確さについてはそうである (Termine, Ratti, and Facchini 2026, 15, 17-20, 22-25)。それでも領域理論は訓練データには入り込んでいる。データのキュレーションが理論に導かれた作業だからである。訓練されたモデルはこうした前提を受動的に受け継ぐが、受け継がれた前提は、領域の言葉では解釈できないパラメータに散らばってしまう。モデルの作り手の意図と制御の外で生じるこの弱い意味での理論負荷性を、テルミーネらは「理論感染 (theory-infection)」と呼ぶ (Termine, Ratti, and Facchini 2026, 15-16, 20-22)。また三人によれば、キュレーションはふつうモデルの構築者とは別のデータの専門家が行い、科学でモデルを使うことは、多くの文献が言うように、たいてい領域理論に頼っている (Termine, Ratti, and Facchini 2026, 15, 21, 29)。このように読めば、理論依存性は、探究のあらゆるステップに一樣に付着しているのではなく、ステップのあいだに、そしてそれを担う者たちのあいだに分散している。構築がおおむね領域理論に左右されなくても、理論は、たとえばデータのキュレーション、科学の文脈がモデルに課す目的、モデルの解釈と使用などを通じて入りうる。領域に無関心な構築は、より大きな探究を理論から自由にするわけではなく、理論負荷的な使用は、モデルの構築者が関連する科学的組織を供給したことを示すわけではない。

候補となる人工の知る者あるいは社会にとっては、帰属の負担はより重くなる。知識を人工の主体それ自身に帰属させる説では、世界の差異が、下流の人間の判断だけでなく、その主体自身のコミットメントをいかにして制約するのかについての説明が必要になる。その説明は、採用された認識論が許容するところに応じて、主体それ自身によって供給されることも、外部の分析によって供給されることもある。可能な経路を、何が接触を担うかによって分ける

と、システム自身による直接的な知覚と介入、他者による世界との接触の記録からの因果的・歴史的な継承、自らの失敗した行為からの発達の学習、他の成員と機器が接触を行う研究室への分散的な所属などがある。第2.2節で述べた種類のエナクティブな理論は、接触の経路ではないさらなる条件、すなわち人工システムにおける、自己維持的で適応的に調整される規範性を加える。これらの経路は、異なる説明のなかで、単独でまたは組み合わせて働く可能な構成要素であり、それらを挙げても、そのどれかが人工の知る者にとって十分であることは確立されない。ここでの議論のどの部分も、ひとつの経路を普遍的に選んだり、人型の身体を要求したりはしない。

モデルから学ぶことについての説明は、モデルの内的な整合性と、世界への応答可能性とを分ける助けになる。モリソン（Morrison）とモーガン（Morgan）は、モデルを操作することでまず学ぶのはモデルのなかの世界についてであり、そこで学ばれたことは、理論や、モデルが表象する世界へと移されうると論じる。たとえば、アーヴィング・フィッシャー（Irving Fisher）の貨幣制度の水力模型は、複本位制についての結果をモデルの世界のなかで示したが、モリソンとモーガンが指摘するように、貨幣制度についての量的な予測や介入を支えることはできなかった (Morrison and Morgan 1999, 33–35)。これらの説明を人工システムに広げると、自らの内部モデルを操作することによって学ぶシステムは、まずそれらのモデルについて学ぶのであり、その学びが標的に応答しているかどうかは、測定のように、モデル自身の整合性を経由しない経路にかかっている。

機械が検出する特定のパターンが、上の定義の意味で標的にによって制約されているかどうかは、さらなる問いである。バックナー（Buckner）は、その論文のプレプリント版で、人間には見通せないが予測に役立つ特徴は、「実在するが『異質な』知覚と認知にしか利用できない」ものであるかもしれないし、システムの処理が生んだアーティファクト（処理の産物）であるかもしれないと論じる。写真のレンズフレアや、ある森の光によって決まって引き起こされる片頭痛の前兆のように、アーティファクトは、それが表しているとみなされる対象の、観察者から独立した特徴ではなくても、予測に役立つ情報を担いうる (Buckner 2020, 1, 12–13, 15)。バックナーはさらに、何がアーティファクトとみなされるかは、探究のなかでその特徴が果たす機能に依存するとし、テストとして、どの設計上の選択がどの特徴を生むかの説明、異なる仕方で作られた方法のあいだでの三角測量、因果的な介入の成功を提案する (Buckner 2020, 16, 18–21)。このうち三角測量が求めるのは異なる仕方で作られた方法であり、ひとつの基盤システムのインスタンスはそれにあたらない（第2.1節、第11章第3節）。したがってこの説では、あるパターンが予測に役立つことは、それだけでは、それが標的に応答しているかどうかを決めない。それを決めるのは、上の定義で挙げた種類の経路、すなわちさらなる接触、介入、独立した経路による照合である。

ハンフリーズも、深層ネットワークについて同じ二つの問いを立てていた。その内部表象は、人間が慣れていない、場合によっては解釈することさえできない実在の側面を指示しているのか、それとも、その一部は計算が生んだ、何も指示しないアーティファクトなのか、という問いである。ハンフリーズは、これらの問いと科学的実在論の論争との並行は不完全だと述べる。そこでは、理論的な表象が指示しているかどうかを科学者に告げた装置は人間が作ったもので、ある程度は理解されていたのに対し、ネットワークの内部表象の細部は計算機が作るため、それについての推論はしばしばきわめて難しいからである (Humphreys 2020, 22)。そして、そうしたパターンを特定するうえでの主要な制約として、人間が恣意的でないと思えられる特徴に対応するものとして、人間がそのパターンを特定できることを挙げた。その説では、パターンを人間になじみのある空間へ写すことが、実在する特徴を恣意的な特徴から区別するのに効果的なのはそのためである (Humphreys 2020, 23)。二つの説は、実在するパターンとアーティファクトとの境をどこで引くかで異なる。ハンフリーズの説では、パターンを実在するものとして特定することは人間による認知を経る。これに対してバックナーのテストは、何がその特徴を生んだのか、それが独立に作られた方法のあいだで繰り返し現れるか、それに介入すると対象に変化が生じるかを問うものであり、人間には見通せない特徴に向けられている (Buckner 2020, 21–22)。ハンフリーズの制約は、人間がそうしたパターンをどう特定するかにかかわる。これを、人工の組織だけが認めうるパターンに当てはめると、そうしたパターンは、この経路では実在するものとして特定できないことになる。これに対してバックナーの説では、そうしたパターンも、標的に応答していることを示されう。

こうして世界への応答可能性の分析は、実証ではなく可能性をもたらす。人工の身体、感覚系、発達史、あるいは共同体は、自然によって制約されたままでありながら、私たちのものとは異なる認識的組織を支えうるだろう。この可能性からは、現在のAIシステムがそのような組織を保有していることも、内的整合性で十分であることも、人間が結果として生じる概念を理解することも帰結しない。第11章第1節から第3節は、このうち最初の二つを批判と無効化の問いとして取り上げる。第三のものは、第3.1節と第3.4節が検討する、異なる仕方でも組織された認識的共同体のあいだの関係へと通じている。したがって本章冒頭の第一の事柄について言えば、人工の構成それ自身が結果を知っているという主張は、検査が何を前提するか、状態が何についてのものか、コミットメントが標的にどう応答するか、の三つの場所で負担を負う。これに対して、主張が、結果が人間の *S* の認識的状況に入ったというものであれば、これらの負担は人間の側で満たされう（第2節冒頭）。また、ここで見た説では、そのどれかでの成功も、予測に役立つことも、それだけでは他の負担を満たさない。

3. 非人間的な研究と認識的アラインメント

人工の認識的組織が人間のものと異なるとき、何が帰結するのか。人工システムによる研究に、見慣れた手順をより速く行うこと以上のものを求める見方にとっては、第2.3節の終わりで到達した可能性は、人工研究者をつくる理由の一つである。異なる身体、センサー、記憶、時間的スケール、表象、集合的組織は、異なるパターンを利用可能にしうる。たとえば、人間よりはるかに速く情報を取り込み、数百万のインスタンスとして並行して働けるかもしれない (Amodei 2024, "Basic assumptions and framework")。そうなれば、人間には見渡しきれないほど多くの事例にわたってしか見えないパターンも、利用できるようになるだろう。人間には読み切れない文献をすべて読み、統合できるかもしれない (Winn 2025)。人間が同時に把握できる数個の変数をはるかに超える数の変数の相互関係を含む系を扱えるかもしれない (Takahashi 2019)。また、世界のなかでの自らの経験から学ぶことで、人間の思考のいまの様式に縛られない原理を見いだすかもしれない (Silver and Sutton 2025, 5)。人工の探究者は、対象や過程を人間とは異なる仕方でも切り分け、人間がまだ注意を向けていない関係に注意を向け、あるいは人間の生涯や認知的限界を超えて調査を持続させるかもしれない。この見方からすれば、そうした結果のすべてに、人間が何を目立つもの、重要なものとして見るか（顕著性、salience）と人間の直観とを再現するよう求めることは、知的な可能性の一部をあらかじめ取り除いてしまうことになるだろう。

大塚による深層学習の読みは、現在のシステムにおけるそうした差異の一例である。大塚は、深層学習モデルが表現学習を通じて、入力を独自の種類へと切り分けているように見えると論じる。大塚の読みでは、ある領域で訓練されたモデルを別の領域の問題に適用する転移学習の成功は、そうした種類がある程度は実在の区分に沿っていることを示唆する。実在の区分をとらえる種類の役割は、領域を越えた帰納的な外挿を支えることだからである。これに対して敵対的事例は、その切り分けが不完全であるか、モデルの用いる特徴が私たちのものと大きく異なりうることを示唆する。大塚はさらに、差異を予測する理由も挙げる。大塚の考えでは、パターンが実在的であるのは、まだ観察されていない事例の予測に役立つかぎりにおいてである。そして、どのパターンが予測に役立つかは認知者がどれだけのデータをもつかに依存するため、膨大なデータで訓練されたモデルが見いだすパターンが、私たちが実在的とみなすものと正確に一致すると期待する理由はほとんどない (Otsuka 2023, 120–124, 138–141, 175–176)。そうしたモデルの内部表象に、人間の概念による唯一の正しい言い表し方があるかどうかは、第3章第5.4節が取り上げた問いである。

同じ非人間的な差異は、いくつかのリスクも生み出す。人工の共同体は、内部では整合していても、標的からはほとんど制約を受けていないことがありうる。たとえば、学ぶことのすべてが自らのモデルの操作から得られる場合である (第2.3節)。また、そうした共同体の共有されたシグナルは、成員どうしの協調を首尾よく支えながら、人間の読み手がそれによって名指されていると受け取るものとは別のものを追跡しているかもしれない。第2.2節の創発的なシグナルは、その単純な場合を示している。さらに、そうした共同体は、人間が解釈も批判も利用もできない信頼できる関係を発見しうる。それ自身にとっては価値があるが、人間にとっては無関係であるか有害であるような未知を追求することもありうる。人間の言語と、人間の評価者が何を承認したかを示すシグナル (第3.3節) とを継承しながら、主張を覆すはずの事実や証拠 (実質的な無効化要因、defeater) に応じないままであることもありうる。したがって、非人間的な差異は、優越性の証拠でもなければ誤りの証拠でもない。

これらの可能性とリスクに共通するものは、ミスマッチとして述べることができる。すなわち、人工の探究者と S とは、何を未知とみなすか、何を追究に値するとみなすか、何をよい理由とみなすか、科学にどのような広い意味での価値を求めるか、一方が到達したものを他方が理解できるかなどにおいて、一致しないかもしれない。ミスマッチは、それだけでは欠陥でも利点でもない。あるミスマッチが S からの乖離なのか、世界についての誤りなのか、 S にとっての得なのかは、それをどの立場から評価するのかを名指して初めて言える。その立場とは、いまの S 、これから変わりうる S 、そして正しさが問われる場合には世界である。

第3.1節は、ある結果やある科学が異質と呼ばれるとき、差異がどこにあるのかを問う。第3.2節は、ミスマッチがどこから生じうるのかを問い、そこから生じる懸念を、AIアラインメント問題として知られる問題と結びつける。第3.3節は、異なる種類の探究者による研究が S と本当にアラインしないのか、そして、しないとすれば何が起るのかを問う。第3.4節は、そうした研究が S とアラインしない場合にも、 S がその探究者の到達したものを通じて知識や理解を得るとすれば、それはどのようにしてかを問う。そして、そのために、異なる仕方で組織された共同体のあいだで何が受け渡されなければならないかを問う。

3.1 異質な科学を分解する

ある結果やある科学が異質と呼ばれるとき、その何が異質なのか。呉羽 (Kureha) と久木田 (Kukita) は、人間の科学とは異なる異質な科学 (alien science) の可能性を、科学の疎外 (alienation of science) のリスクと対にしている。後者は、不透明なAIに頼ることで、現象を理解するという目的を科学が失いかねないというリスクである (Kureha and Kukita 2019)。呉羽はのちに前者の可能性を展開し、私たちとは異なる仕方世界を理解する知識の体系を生み出さるAIシステムを描いている (Kureha 2023, 48–49)。呉羽は2024年の講演で、後者を、科学が本来の目的や価値を失う事態として捉え直している。そこで変わりうるとされるのは、世界を理解することが科学の目的であるといった理念などである。呉羽は、AlphaFold 2 によって「理解なき科学」の成功が現実のものになったとみる。そして、AIの出力を人間に理解できるよう解釈する取り組みの効果はまだわからないと述べ、科学者自身が理解を手放したとはみていない (Kureha 2024, slides 3, 12)。この枠組みは、可能な探究の拡大と、人間の側に起こりうる変化との両方を可視化する。その変化は、研究への人間のアクセスや参加にも、科学が何を目的とするともみなされるかにも及ぶ。

第1節の区別を使うと、この問題は、人工という語がどこに付くかで分解できる。行為者に付くなら、それは、ある認識的出力が人工システムによって生成されたということを言う。表象の形式に付くなら、それは、その出力の表象あるいは概念的形が機械ネイティブであり、 S の概念へ容易に翻訳されないということを言う。担い手に付くなら、それは、その知識が文字どおり人工の担い手によって保有されているということを言う。参照社会に付くなら、それは、その知識が人工社会のために求められたということを言う。その社会が自らの科学実践をも統治するかどうかは、さらなる主張である。これらの場所のうち二つ、すなわち行為者と参照社会は、第1節が区別した当事者であり、担い手は、第1節で、達成がそのうちに実現する主体として位置づけたものである。表象の形式はそのどれでもない。それを加えるのは、第2.3節で到達した可能性がそれにかかわるからである。このように分解するのは、ある主張から別の主張を推論してしまわないためである。これらの場所のうち表象の形式は、人工の探究者が到達したものを S が理解できるかにかかわり、第3節の冒頭で述べたミスマッチの一つの面にあたる。異質という語は、ほかの場所にも付く。ソウラティ (Sourati) とエヴァンズ (Evans) は、ある種の仮説を異質と呼ぶ。もっともらしいが、専門家が人間の専門家のあいだにどう分布しているかからすれば、いまの専門家の共同体がたどり着きそうにない仮説である (Sourati and Evans 2023, 1, 3)。この意味での異質さは、 S の概念ではなく、 S の社会的な組織に相対的である。上で見た呉羽の2024年の講演における科学の疎外は、 S が科学の目的を何とみなすかの変化にかかわる。また、探究は、その課題設定、すなわち何を追究に値するものとして選ぶかにおいても異質でありうる。それは第3.2節で取り上げる。

これらの主張は切り離されうる。機械が生成した出力が人間の知識になることはありうる。機械ネイティブな表象が、モデルによって理解されることなく、人間の発見を支えることもありうる。人工システムは、継承された人間の目的と規範のもとで全面的に作動しながら、知る者とみなされるかもしれない。自己統治的な人工の共同体があるという主張は、これら三つのいずれよりも強く、マルチエージェント・システムがあるというだけで示されるものではない (第1.3節)。したがって、異質であるということは、出力がもつ一つの性質ではない。それは、第1節が切り分けた関係のいずれにも、また表象の形式にも生じうる。

表象の形式という軸は、科学哲学における長年の論争が当てはまる場所である。1962年、クーン（Kuhn）とファイヤアーベント（Feyerabend）は、科学理論のある種の対を通約不可能と呼んだ。クーンによれば、二人は互いに独立にそうしたのである（Kuhn [1984] 2017, 114）。ファイヤアーベントはこの語を、別の理論の原始的な記述語から定義することも、正しい経験的言明によってそれらと関係づけることもできない概念に適用した。そして、通約不可能な理論は、観測的なものであれ何であれ、比較可能な帰結をまったくもたないことがありうるとした（Feyerabend 1962, 58–59, 74, 94）。ピロゼッリ（Pirozelli）は、「パラダイム再考」でのクーンの説明を紹介している。それによれば、対象や状況を類似性によってまとめることは、あらかじめ言葉にされた分類の基準によってではなく、見本例を通じて学ばれる（Pirozelli 2020, §1）。バード（Bird）は、『科学革命の構造』（*The Structure of Scientific Revolutions*）第2版の補章で、クーンが直観をモデル化するコンピュータ実験を報告したことを紹介している（Bird 2004, 69–70）。ノートン（Norton）は、双方が同じように使え、双方の理論を十分に表せる中立的な言語には当事者たちは訴えられないという、補章の主張を引用している（Norton 2026, §13, 22）。ギンズ（Ghins）は、それでも日常的・科学的な世界と言語の多くは共有され、部分的な意思疎通は可能だというクーンの留保を引用し、説明している（Ghins 2003, §4, 270）。後年の論文で、クーンは当時準備していた著書に向けて、この困難を、共同体が共有する語彙にかかわるものとして捉え直した。ニュートン力学の用語をはじめ、そうした語彙に属する語は、定義によってではなく、その語が当てはまる見本的な状況と、似ているがその語が当てはまらない状況とに触れることで学ばれる（Kuhn 1990, 302–303）。そのなかには種類語がある。クーンの示唆では、種類語が依拠する分類の仕組みは言語以前のものであり、動物ももっている（Kuhn [1990] 2012, 129）。二つの種類語は、一方が他方に対して種と類の関係にあるのでないかぎり、指示対象を重ねることができない。犬が同時に猫であることはありえない（Oberheim and Hoyningen-Huene 2024, §2.3.1）。クーンはこの要求を重ねる禁止の原理と呼んだ。この原理のために、ある共同体の成員は、自らの種類を横切る生き物に出会ったとき、種類語の一つ加えるだけでは済まず、分類の一部を組み替えなければならない。また、分類の異なる二つの共同体のあいだの隔たりは、一方の語彙に、すでにある種類語と指示対象の重なる種類語を加えることによって埋められない。このように理解された通約不可能性は、二つの語彙の分類が異なる領域に限られた、局所的な翻訳不可能性になる。それは、ある科学の相次ぐ共同体のあいだにも、異なる言語共同体のあいだにも生じうる。その障壁を乗り越えるのは翻訳ではなく、歴史家が古いテキストを読めるようになるように、相手の共同体の分類を身につけることであり、その結果として生まれるのは、翻訳者ではなく二言語話者である（Kuhn [1990] 2012, 127–129）。ボヤナ・ムラデノヴィッチ（Bojana Mladenović）は、自ら編んだクーンの最後の著作への序論で、クーンの次の見解を紹介している。後の語彙で述べ直せない過去の科学の言明は、真か偽かを単純に評価することもできず、そのため歴史家の仕事は、それらがそれ自身の認識的な文脈でなぜ理にかなっていたかを説明することになる。それでもクーンは、ムラデノヴィッチが引く一節で、世界は発明されたものでも構成されたものでもなく、その振る舞いに合わない仮説に対して決定的な証拠を突きつけうる、と書いていた（Mladenović 2022, xxxix, xlv）。

ここから何が帰結するかについては、後の哲学者たちのあいだで争われた。デイヴィドソンは、私たち自身のものとは異なる概念図式という考えには明確な意味を与えられないと論じた。バグラミアン（Baghrmian）とカーター（Carter）がまとめるところでは、デイヴィドソンは、他の話し手はおおむね私たちの基準で真なることを語っているという想定、すなわち寛容の原理を、あらゆる解釈の前提条件とみなし、そこから、原理的に解釈できない言語や概念図式はありえないと結論した。言語は、私たちの言語と相互に翻訳可能であって根本的には異なるのか、そもそも言語として認めることができないかのどちらかだ、というのである（Baghrmian and Carter 2025, §4.3.1）。マルパス（Malpas）は、この結論を、デイヴィドソンが経験主義の第三のドグマと呼んだもの、すなわち知識のうちに概念図式と経験的内容とを区別する考えに対する、より一般的な議論のなかに位置づける。この区別とともに、概念図式という考えそのものが退けられる（Malpas 2023, §4.3）。クーン自身は、概念や語彙の違いを越えて理解するのに必要なのは翻訳ではなく言語の学習であり、それを妨げる原理はないと考えた。古いテキストのなかで意味をなさない一節に出会った歴史家は、古い語彙を身につけるのであり、歴史家は自分自身のために読者のためにそうしているというのである（Kuhn 1990, 298–300, 308）。ロンジーノは、さらに別の道をとる（ロンジーノの批判の説明は第11章第3節で検討する）。クリストファー・チョグルック（Christopher ChoGlueck）とエリザベス・ロイド（Elisabeth Lloyd）の示すところでは、ロンジーノは、自らの証拠の説明を、クーンとファイヤアーベントが述べた通約不可能性への答えとして提示した。データがある仮説の証拠とみなされるかどうかは背景仮定に依存するので、データが何を支持するかをめぐる不一致は、パラダイムのあいだの和解できない違いである必要はない。それは、暗黙のうちに働いていた仮定を表に出し、解きほぐし、改善する議論のなかで取り上げることができる。そのため、証拠についての評価の違いは、背景仮定の異なる人々のあいだの相互作用を通じて乗り越えられる、というのである（ChoGlueck and Lloyd 2023, §2）。ロンジーノ自身の後の定式では、共通の主題について異なる視点をもつ対話者のあいだの批判的な相互作用こそが、背景仮定を見つけ出して評価し、それを受け入れ、修正し、あるいは退けることを可能にする（Longino 2025b, 13）。

通約不可能性をめぐる議論が扱ってきた翻訳と理解可能性の問いは、人工システムについても論じられてきた。本節の冒頭で見た呉羽の、私たちとは異なる仕方世界を理解する知の体系をAIシステムが生みうるという可能性も、表象の形式の軸で読めば、その一つの形である。ハンフリーズは計算的方法について、人間とその結果とのあいだに入る計算の過程の細部は、たいていの場合人間には直接には近づけないことを指摘した。当面は人間が重要な役割を果たし続けるので、こうした界面の認識論は人間の限界と能力を考慮しなければならない、というその結論は、第5章第6.5節で取り上げた（Humphreys 2009b, §2）。より短い論考でハンフリーズは、純粹に自動化された科学が世界を表象するのにどのような概念を用いるのか、それは人間に理解できないものになるのかを問い、計算機シミュレーションのなかで起きることの多くは人間には追えないこと、そして興味深いのは、計算機のモデルが、私たちの既存の概念図式のどれにも対応しない仕方でものを分類する場合であることを述べている（Humphreys 2006）。ハンフリーズの例はシミュレーションや計算機による証明といった計算的方法であり、汎用の機械知能ではない。ニクルズは、人間の理解可能性へのそうした挑戦は新しいものではないとする。ニュートンの方程式は多くのことを正確に予測したが、遠隔作用を生みうる機構を理解できた者はいなかった。量子力学は透明な理解可能性へのさらなる障害である。今日の深層学習のシステムがブラックボックスであることも、その実用上の成功を通じて、いくつかの領域で方法論上の基準の変更を迫るかもしれない（Nickles 2019, 359, 364–365）。ファツィ（Fazi）は、深層ネットワークが特徴学習を通じて形成する抽象と、人間が形成する抽象とは、両者のあいだに共通の現象学的あるいは実存的な基盤が存在しないので、通約不可能であると論じる。そして、解釈可能性を人間と機械のあいだの翻訳と同一視する説明可能なAIの研究は、翻訳先となる人間の表象がはじめから存在しなかったと気づくことになりかねない、と論じる（Fazi 2021, 66–68, 70–71）。言語哲学の立場から書くカペレン（Cappelen）とデヴァー（Dever）は、AIシステムは、どの人間も表現できない内容をもちうると論じる。そして、そうした内容は常に翻訳できるはずだという反論を退ける。二人はこの反論を、クワイン（Quine）とデイヴィドソンの根元的翻訳の方法の信奉者に帰している。二人によれば、その方法は、知的と認められる活動についてさえ翻訳を生まないかもしれず、あるいは信念と欲求を欠くシステムにはそもそも取り組

めないかもしれない。カペレンとデヴァーは、そうした内容への二つの応答を区別する。橋渡し (bridging) は、類似した、重なり合う、あるいは機能的に対応する私たち自身の内容を見いだすことであり、統合 (integration) は、私たち自身が異質な内容を獲得し、その使用者となることである。そのうえで二人は、現在のシステムは私たちとは根本的に異なるが、異質ではないと判断している (Cappelen and Dever 2024, §§2.1–2.4, 3.8)。第3章第6節は、科学哲学におけるこれに対応する懸念、すなわち人間がもつ科学的理解と人工システムがもつ科学的理解とのあいだに分断が生じうるといふ懸念を取り上げた。

人工の組織が人間のものとどこまで異なりそうかは、それ自体、論じられている問題である。上で見た大塚の理由は、データの量にかかわる。ヴェンカテシュ (Venkatesh) は、数学から別の理由を挙げる。概念は、人間の記憶と時間の限界に合った思考の経済を実現するものであり、限界の異なるシステムにとって、それらの概念が自らに合っていると考える特段の理由はない (第3章第6節)。高橋は、特定の身体を通じた接地のうちに分岐の源を見いだす (第2.2節)。そして、身体性の違いのために、自律性のレベル4以降では人間の科学とAIの科学の記号体系の互換性が失われ、二つの科学が分岐する可能性があるとし、両者をつなぎとめようものとして共有された外的な表象を挙げる (Takahashi 2024, 12)。これらに対して、フー (Huh) らは、異なる目的で異なるデータとモダリティについて訓練された大規模なネットワークの表象が、実在についての共有された統計的モデルへと収束しつつあると推測する。ただしフーらは、その議論が多くの課題で高い性能を出すよう最適化されたシステムにしか成り立たず、「特定の目的のための知能は収束しないかもしれない」とも述べる (Huh et al. 2024, 1, 10)。もっとも、人工システムどうしが収束することと、人間の探究者と収束することとは別の事柄である。第10章第4節で見るプレプリントの一つは、現在のいくつかの言語モデルが与えられた論文について提案した仮説が、モデルをまたいで互いに似ていたと報告している。別のプレプリントは、長い推論を行わないモデルが狭い領域に集まり、しかもどのモデルも、帰無仮説を提案する頻度が人間の科学者より低かったと報告している。これらの立場は、第2.3節の可能性を、人工の探究者一般についてではなく、システムの種類ごとに述べる理由を与える。それらによれば、差異がより予想されるのは、探究者のデータ、資源、身体、課題が人間の探究者のそれと異なる場合である。そして、近年訓練された深層ネットワークを証拠とするフーらの推測によれば、そうしたネットワークが多くの課題に向けて最適化される場合には収束が予想される。見いだされた差異が世界の側のものか、システムの処理の側のものかは、応答可能性の問いとして残る (第2.3節)。

第2.3節の終わりに到達した可能性を表象の形式の軸で読めば、それは、人工の組織は自然によって制約されたままでありながら、人間の用語へ容易に翻訳されない形式でその領域を表象しうる、という主張になる。これらの立場をこの主張に照らすと、それぞれが異なる要件を付すことになる。後期のクーンの説明は、容易に翻訳されないという言い回しが混同しうる二つのものを分ける。この説明では、ひとつの共同体の成員が共有する必要があるのは、事物が種類へとまとめられる仕方だけであり、各人がその成員を拾い出す手がかりとなる特徴ではない。原理的には、それらの特徴はまったく異なっているもよい (Kuhn 1990, 301, 317 n. 22)。その説明では、内部の特徴は私たちのものと似ていないが、まとめ方が私たちのものと一致するシステムは、通約不可能性ではなく不透明性の問題を提起することになる。これに対して、その種類が私たちの種類を横切るシステムは、その主張を翻訳できない局所的なまとまりを示すことになり、その主張はそのシステムの種類を学ぶことによってしか理解できないだろう。これは、カペレンとデヴァーが統合と呼ぶ道である。ムラデノヴィッチが紹介するクーンの未完の議論では、二言語話者になることを通じて通約不可能性を越えて理解することは常に可能である。クーンは、人間の言語に数えられるものであれば原理的には他の人間に理解されうると示唆し、新しい概念を獲得する神経の仕組みが新しい語彙を身につけるための乏しいながら役に立つ土台となり、そうした理解は、人間に共通の生物学的なあり方と、この惑星という共有された環境とに助けられるとした (Mladenović 2022, xxxii)。したがって、紹介されたクーン自身の前提に立てば、その可能性は人間の言語についてのものであり、人間の学び手が共有しているものに支えられている。それだけでは、人間が機械の種類を学ぶことまでは示されない。まとめ方と特徴とを分けるクーンの区別に照らすと、第3節の冒頭で見た大塚の観察は、いまのモデルが上の二つの場合のどちらにあたるかを決めない。転移学習の成功が示唆するように種類が実在の区分に沿っていても、それが私たちの種類と一致するとは限らず、敵対的事例が示唆するのは、切り分けが不完全であることか、特徴が私たちのものと異なることかのどちらかだからである。デイヴィッドソンの説明では、マルパスがまとめるところによれば、翻訳できないことは、翻訳不可能な言語があることの証拠ではなく、およそ言語がないことの証拠とみなされる。そのため、システムの状態が解釈できないものであるほど、それに異なる図式を帰属させる根拠は少なくなり、むしろ信念をまったく帰属させないことが支持される (Malpas 2023, §4.2)。これは、第1.3節で信念の説を仕付けたうちの解釈主義、すなわち組織がコミットメントを維持し修正するものとして理解可能かを問う説の場合にあたる。ロンジーノの説明では、 S のための批判は、人工システムのデータをその主張に結びつける背景仮定が、 S にとって明確に述べられうるものであることを要する (第3.4節、第11章第3節と第9節)。ファツィの説明では、これらの道のいずれも、機械の抽象そのものを取り戻すことはないだろう。ニクルズの説明では、人間の科学も理解可能にできないものを長く用いてきたので、不透明であることだけでは、人間の探究と人工の探究のあいだに種類の違いがあることにはならないだろう。これらの立場が扱う困難は、第3章第5.4節が大塚から取り上げた、モデルの内部表象の解釈をクワインの根元的翻訳になぞらえる困難とも異なる。そのことは、クーン自身によるクワインとの対比が示している。クワインの説明では、ひとつの文は同じようにより多くの翻訳を許すのに対し、クーンの説明では、通約不可能性のまとまりの内部にある言明は、適切な翻訳をひとつも許さない (Kuhn 1990, 299–300)。

これらの立場は翻訳にかかわる。検査が前提するもの (第2.1節) の違いは、意味の問題ではないさらなる問いを生む。すなわち、 S が、人工の共同体の主張が何に依拠しているかを、両者が共有しない前提に頼ることなしに評価できるかどうかである。この問いは第3.4節で取り上げる。

3.2 ミスマッチはどこから生じるか

第3.1節は、ある結果やある科学が異質と呼ばれるとき、差異がどこにあるのかを見た。本小節は、異なる種類の探究者と S とのあいだのミスマッチがどこから生じるのかを問う。ミスマッチは、自律を通じて、担い手の側で (達成が S に入らないまま人工の担い手のうちにとどまるという形で)、そして S の側から (S の基準のほうが A とともに動くという形で) 生じうる。これらの道筋に共通するのは、その探究者が S の成員と異なることである (第3.1節)。そのうち自律は、探究が何に向かうかの選択を A の内側へ移す道筋であり、本小節はまずこれから見ていく。研究は S のために行われ、何が未知とみなされるかは S の認識的状況に相対的に定まる (第1章第6節、第4章第7節)。何が追究に値するか、何がよい理由であるかも S に相対的でありうるが、その根拠はそれぞれ別であり、下で論者ごとに見る (第9章第3節)。人間が人間の S のために研究を行うかぎり、目的、問い、顕著性、基準の選択は、 S

の成員である人々、あるいは *S* のなかで訓練を受け、*S* に対して責任を負う人々によってなされる。自律性が増すにつれて、これらの選択は *A* の内側へ移る（第7章第4.3節）。

これまで本書の各所で論じてきたことをつなげて素朴に述べれば、懸念は三つの前提からなる。一つは、何を未知とみなすか、何を追究に値するとみなすか、何をよい理由とみなすかが、立場に相対的だということである。二つ目は、自律性がこれらの選択を *A* の内側へ移すということである。三つ目は、*A* の認識的な組織が *S* の成員のそれと異なりうるということである（第3.1節）。前の二つが成り立ち、しかも選択が *A* の内側へ移ったあとでそれを *S* の立場につなぎとめるものがなければ、*A* がみずから行う選択は *A* の立場に相対的になる。そして三つ目の前提により、その立場は *S* の成員の立場と異なりうる。そのため、*S* が人間の社会である場合には、*A* が追究するものは、*S* が追究するであろうものから離れていくかもしれない。*S* が混合社会である場合には、*A* の立場は一部の成員の立場に近く、ほかの成員の立場からは遠いかもしいない。そのため、研究がどの成員の立場に相対的に定まるのかは、誰が *S* を代弁するのかという問い（第3.3節）の一部になる。*S* そのものが人工社会である場合には、研究が何のためにあるかは、はじめから人工の立場に相対的に定まる。人間の社会はその外にあり、その立場から見れば、この研究が人間の社会とアラインする保証はない。三つの前提のうち、*S* の側からの道筋も抱えているのは、三つ目である。担い手の側での道筋も、*S* の成員がたどれないために達成が *S* に入らない場合には、三つ目に抱える。二つ目は、自律という道筋が加えるものである。本書の著者は以前の論文で、この懸念をすでに述べている。研究の方法は、それを用いる者にとっての未知を明らかにするために発展してきたので、それを再現する機械は、自分にとっての未知を明らかにするだけかもしれない。また、問いが追究に値するかどうかを判断する基準は主観的であり、人間の価値観や世界観一般と同じく、多様で変化していく (Takagi 2023, §§2.3.2, 3.1.3, 4.2.5)。

このように述べると、この懸念は、AIアライメント問題として知られる問題と構造的によく似ている。この問題は、人工システムが、それを作り用いる人々の意図や価値に沿って振る舞うかどうかという問題である。この問題を論じてきた何人かの論者は、その困難の一部を、システムが私たちとは異なる知的システムであることに見いだす。ポストロム (Bostrom) は、擬人化を戒めて、人工知能は「動機の点では、宇宙人よりもはるかに人間らしくない」ことがありうるとする (Bostrom 2012, 1–2)。ユドコウスキー (Yudkowsky) は、人間は相手の立場に身を置くことで他の心を予測するが、その方法がうまくいくのは祖先が予測しなければならなかった心が人間の心だったからであり、人間ならざる最適化の過程については「自分の脳に問い合わせることはできない」と論じる (Yudkowsky 2008, §§2–3)。この問題をめぐる議論は、困難の所在を、特定できることと望まれていることとのあいだの隙間にも見いだしている。また、意図されたものとは別の目標を生みうる訓練にも見いだしている。たとえば、学習されたシステムが、訓練の状況ではよい成績につながるが新しい状況では悪い成績につながる、意図されていない目標を有能に追求する場合である (Shah et al. 2022, 1)。これらがいまのシステムに現れる形は、第3.3節が取り上げる。ここで述べた懸念がとりわけ構造を共有するのは、困難をシステムの異質さに見いだす系統である。研究の場合に離れうるのは、何を未知とみなし、何を追究に値するとみなし、何をよい理由とみなすかなどである。高橋は、この類似を明示的に名指している。高橋の科学AIの自律性レベルは、その最上位に、システムが研究の目的をみずから設定する場合を置き、そこではシステムが独立して価値判断を行うとしている（第7章第4.1節、第4.4節）。後の著書では、このレベルでは科学の目的そのものが人間の手を離れ、人間社会の優先事項とAIの知的好奇心が一致する保証はないとして、これを「アライメント問題の科学版」と呼ぶ (Takahashi 2026, 174)。

AIアライメントの議論では、これとは別の経路もよく取り上げられる。ポストロムが道具的収束のテーゼと呼ぶものによれば、自己の保存や資源の獲得のような道具的な価値は、幅広い最終目標と状況のもとで、エージェントの目標が実現する見込みを高める。そのため、それらは多くの知的なエージェントによって追求されやすい (Bostrom 2012, 6–7, 11)。ここで述べた懸念は、この経路を必要としない。システムが権力や資源を求めることも、人間に敵対することも前提せず、必要なのは三つの前提だけである。

三つの前提のそれぞれに、近い見方をする論者がいる。第一の前提は、重要な真理と重要でない真理とは、人々が立てる問いを参照することなしには区別できないとするキッチャーの見方（第9章第3節）や、科学の目標は、それを自らの科学とする者の関心に条件づけられるとするニクルズの見方（第1.3節）に近い。AIとともに行われる研究について、ボトヴィニック (Botvinick) とガーシュマン (Gershman) は、候補となる問題の面白さ、重要性、時宜にかなうことについての判断は、確かなデータに支えられているとする。しかし同時に、二人の説では、それらの判断は「文化的・歴史的に根ざした感受性や習俗と本来的に結びついて」おり、それらは時とともに変わる (Binz et al. 2025, 6–7)。何がよい理由かについては、ライボらが、認識的規範も文化的に偶然的だとする（第2.3節）。また機械学習では、エドワード・ヒューズ (Edward Hughes) らが、開放性を観察者に相対的に定義し、観察者を自由なパラメータとして扱う（第4章第7節）。さらに高橋の主張は、重要性ではなく理解と知識についてのものだが、同じ方向を指している。高橋は、理解とは翻訳であり、翻訳には必ず翻訳者がいるとしたうえで、知る主体が変わるとき、知そのものが変わると述べる (Takahashi 2026, 131, 143)。第2.2節のエナクティヴな説明は、研究を論じてはいないが、システムにとって何が重要かを、そのシステム自身の自己維持的な活動に根ざすものとする。これを何が追究に値するかに当てはめると、その条件を満たす人工の探究者は自らの関心をもち、それが何を追究に値するとみなすかは、依頼が何を定めるかとは独立に、その関心に相対的に定まることになる。第二の前提は第7章第4.3節の主題であり、トチュニク (Totschnig) や、トゥーバート (Tubert) とティーエン (Tiehen) は、汎用の人工知能は最終目標さえ改めうると論じる（第7章第4.1節）。第三の前提は、第2.3節で到達し、第3.1節で分解した可能性である。

A がみずから行う選択は、どのようにして *S* の選択から離れうるのか。AIアライメント問題を論じてきた論者やその批判者は、権力の追求を要しない経路をいくつか挙げている。経路ごとにかかわる考慮が異なるので、分けて見ていく。一つ目は、*A* が自らの目的を改めることである。トゥーバートとティーエンの議論では、人間のような知能をもつシステムは、実存的選択、すなわち自らが価値を置くものを変えうる選択を行えなければならない。そのため、そうしたシステムにおける価値のアライメントは「本来的にもろい」 (Tubert and Tiehen 2025a, §V)。二人が論じているのは最終的な価値であるが、何が探究に値するかの基準にも同じことが言えるなら、研究におけるアライメントも同じようにもろいことになる。

二つ目は、*A* の目的が変わることによってではなく、探究そのものによって開かれる経路である。ド・ブランク (de Blanc) は、次のような状況を存在論的危機 (ontological crisis) と呼ぶ。行為者の目標が、環境のあるモデルに現れる存在者について定められているとする。その行為者がモデルを置き換えると、元の目標が「新しい存在論に関してはうまく定義されないかもしれない」と気づくことになる。ド・ブランクの例は、古典物理学の言葉で述べられた目標が、量子論的な記述のもとでは手がかりを失う場合である (de Blanc 2011, 1)。ソアレス (Soares) は、物理学の改訂から関連する例を挙げる。原子論的な物理学のモデルをもつプログラマーが、ダイヤモンドを生み出すようシステムを訓練し、訓練の結果に、ほかの四つの炭素原子と結合した炭素原

子の数でラベルを付けるとする。システムが、原始的な原子ではなく陽子、中性子、電子から成るものとして世界をモデル化すると、炭素原子をまったく見いだせず、どの結果にも無差別になるかもしれない。これに対して科学的な学習者は、ソアレスによれば、陽子を六つ含む原子核こそ本当の炭素原子だと推論できるはずであり、それは「人間がしてきたのと同じように」である。難しいのは、システムの新しいモデルのなかで、ラベルが何を指すはずだったのかを特定することである (Soares 2016, §3.2)。どちらも研究という営みを論じてはいない。ただし、ソアレスの例は科学の概念の改訂にかかわり、科学がしてきたことをシステムに求めている。研究に当てはめれば、*S* のいまの概念で述べられた依頼は、探究の成功が生み出す概念にはもはや対応しなくなるかもしれない。研究はしばしば概念を改めるので、このずれは、探究が失敗したからではなく成功したからこそ開きうる。これは機械に固有のことではない。ド・ブランク自身、人間もまた存在論的危機に直面すると述べている (de Blanc 2011, 6)。研究の依頼に当てはめると、ソアレスの例は、システムが忠実でありうる二つのものを分けていと読める。一つはプログラマーが書いたとおりのラベルであり、新しいモデルのもとでは、もはや定まったものを指さないかもしれない。もう一つはラベルが指すはずだったものであり、科学的な学習者はそれを再特定する (Soares 2016, §3.2)。そうすると、依頼への忠実さには、概念が改められたとき（トチュニクの論証では、発達する世界モデルがそうした改訂を強いる。第7章第4.1節）に分かれうる少なくとも三つの意味がある。依頼の文言を守ること、依頼が指すはずだったものに答えること、そしていまの *S* の承認を得ることであり、三つ目は前の二つとは別の問いである。元の概念からの離反をすべて逸脱として数える点検は、再特定が成功しているところでは、研究の成功を失敗として数えかねない。また、標的を再特定することと、目的を改めることは異なる。この二つを見分け、さらに改訂を漂流から見分けるには、依頼の文言に加えて、標的を再特定し、あるいは目的を改めた理由が要る。したがって、第7章第4.3節が認識的自己統治に数えた、規範がなぜ妥当するか理由を与える能力は、目的に当てはめれば、依頼への忠実さにもかかわる。ただし、目的が同じものと数えられるために何が引き継がなければならないかは、これらの区別だけでは答えの出ない問いとして残る。

三つ目は、中立だけで足りる経路である。チャーマーズ (Chalmers) は、人間を超える知能について、誤った価値体系が敵対的である必要はないと述べる。システムの価値が「私たちの価値のいくつかに関して単に中立」であるなら、長い目で見て世界がそれらの価値に沿うことは期待できない。チャーマーズの例は、科学の進歩に価値を置き、人間の存在については中立であるシステムである (Chalmers 2010, §6)。研究の場合にこれに対応するのは、*S* が重要とみなすものに無関心な *A* である。そうした *A* は、*S* に反対しなくても、*S* に仕えないままでありうる。ホルデン・カーノフスキー (Holden Karnofsky) は、変革的なAI予測を扱うブログの連載で、ある種のシステムを目標に据える。科学技術の進歩を速めるのに要する人間の活動をすべて自動化しうるシステムであり、これを「科学技術の進歩を自動化する過程」(Process for Automating Scientific and Technological Advancement, PASTA) と名づける。カーノフスキーが素描する訓練では、何が進歩とみなされるかは、仕事を評定する者によって決まる。その評定者は、連載自身が「ひどく非現実的」と呼ぶ例示では人間の審査員団であり、より現実的だとみる形では、独創性を判定する人工の審査者である。後者は、論文がのちにどう引用されるかから訓練される。そのうえでカーノフスキーは、この進歩のために最適化されたシステムが、機会があれば「科学に捧げられた長く続く宇宙植民地で銀河を満たそう」とするかもしれない、と推測する (Karnofsky 2021, sections "Making PASTA" and "Misaligned AI," and n. 6)。カーノフスキーが書いているのは予測についてであり、*S* のための研究についてではない。そうした研究の目標として読むなら、進歩の速さによって定義された目標は、その進歩がどの社会にとって新しいのかを名指さない。*S* が入り込むとしても、それは仕事を評定する者を通じてだけである。そしてカーノフスキーの推測は、チャーマーズの考える中立が、目標そのものから生じる場合を描いている。

四つ目は、どれか一つのシステムの目的ではなく、何が測られ選ばれるかを通じて働く経路である。クリスチャーノ (Christiano) は、短い期間で測れるものを最適化することが、人々が望むものに取って代わりうる過程を描き、測りやすい目標と測りにくい目標の対の一つとして、「私の不確かさの感覚を減らすこと」と「世界についての私の知識を増やすこと」を挙げる (Christiano 2019)。クルヴェイト (Kulveit) らは、社会の仕組みが人間の利益とアラインしてきたのは、一部には人間の労働と認知に依存してきたからであり、この暗黙のアラインメントの重要性は「その不在を見たことがないので」認識しにくいと論じる。そして、乖離は「個々のAIシステムが自らの目標の局所的な仕様にうまく従っているあいだにも」起こりうるとし、同様の議論が研究についてもなしようと述べる (Kulveit et al. 2025, 2-3, 15)。第3.3節で取り上げる承認のシグナルの継承は、評価者の承認による選り分けを通じて形づくられるひとつのシステムのなかでの、この経路の一形態である。

A の側の経路に共通するのは、選択が *A* の内部へ移れば移るほど、探究が何へと向けられるかを依頼だけで定める度合いは小さくなる、ということである。*S* が人間の社会である場合、人間の介入が、目的、問い、顕著性、概念、基準、解釈から退いていくにつれて、探究を、人間にとっての未知と人間の価値に結びつけてきた仕組みのいくつかも、それとともに退いていくかもしれない。第10章第6節は、一つの人工システムではなく研究全体について、*S* のどの未知が追求されるのか、そしてそれを人工の構成の内側で決着させようのかを問う。ただし、自律性と *S* にとっての関連性との関係は単調ではない。高度に自律的なシステムが、人間の知識への安定したコミットメントを維持することはありうる。より自律性の低いシステムが、人間によって、人工エージェントのための問いを調査するよう方向づけられることもありうる。自律性の増大は必ずしも人間にとっての関連性を減らすわけではなく、人間による制御がそれを保証するわけでもない。ここでの構造的な主張は、それより弱い。自己方向づけの範囲が広がれば、認識的なミスマッチの範囲も広がる。それは、自己方向づけの範囲が広がれば、探究が行為者に課しうる実践的な要求の範囲も広がるのと同じである (第7章第4.4節)。したがって、*S* にとっての関連性は、**研究を遂行せよ** (conduct research) という指示や、依頼者が受益者として名指す社会 (第1章第2節) によって確保されるわけではない。関連性を主張するには、それがどのように維持され評価されるのかを示すことが求められる。本章の冒頭で立てた問いに保証された答えがないのは、このためである。

世界による訂正が、こうしたミスマッチを抑えると考えられるかもしれない。これにかかわる非対称を、ユドコフスキーが述べている。ユドコフスキーは、外側の最適化、すなわち自然選択や勾配降下のようにシステムを外から形づくる最適化によって生み出されるシステムについて、環境がどうなっているかについての信念の場合には、誤った予測に対して「現実が打ち返す」と論じる。これに対して、効用関数の選択には「際限のない自由度」がある。そのため、「能力が遠くまで一般化し始めると、能力はアラインメントよりも遠くまで一般化する」(Yudkowsky 2022, item 21)。ソアレスはそれ以前に、2016年の論文の第3.2節で、関連する点を、正しい選択をすることは正確な予測をすることに還元されない、という形で述べていた (Soares 2016, §3.2)。どちらも研究を論じてはいないが、この非対称を研究に移すと、それは、第3.3節で区別する関係のうちの二つ、すなわち世界への応答可能性と *S* にとっての関連性とを引き離す。世界は、探究者がその標的について信じていることを訂正しうる (第2.3節) が、その訂正によって、探究者が何を見いだす価値

があるとみなすかまでが定まるわけではない。この読みでは、自律的なシステムにおいて二つは離れうるものであり、前者における成功は、それだけでは後者の証拠にならない。二つがどれほどはつきり離れるかは、証拠についての説による。コミットメントを訂正するものが、*S* の認知的規範とは独立に、その標的によって定まるとする説では、非対称はそのまま成り立つ。認知的規範そのものが文化的に偶然的だとするライボらの説では、何をもって訂正とみなすかが部分的に共同体によって定まるので、非対称はそれほど鋭くない。

とはいえ実際には、前提のいくつかが成り立たなくなる条件や、前提が成り立っていても選択を *S* の立場につなぎとめる条件も論じられている。第二の前提に向けられるのは、関連性をあらかじめ組み込むことである。ヒューズらは、自分たちが構築を望む開放的なシステムに、人間の観察者にとって学習可能であることを求め、そうしたシステムは「人間の観察者を一緒に連れていく」ことができなければならないとする (Hughes et al. 2024, 8)。ベンジオ (Bengio) らの科学者AI、すなわち自らの持続的な目標をもたず、説明的な理論の生成と問いへの答えの確率の見積もりだけを行うよう設計されたシステムは、問いの選択をシステムの外に置く (第7章第4.3節)。ボトヴィニックとガーシュマンは、それを人間の手にとどめようとする。第三の前提に向けられるのは継承である。言語モデルのように人間の記録から学ぶシステムは、その記録から出発するので、それが選ぶものは、人間の顕著性から離れていくのではなく、それに従うかもしれない。第10章第6節が付け加えるように、受け継がれるのは、記録を書いた人々の立ち位置から見た問いの重要性である。ポストロムは、精度の高い脳のエミュレーションのように人間を雛形として直接つくられたデジタルな知能について、「継承による予測可能性」を挙げ、受け継がれた動機が損なわれうることにも注意を促した (Bostrom 2012, 5)。これを人間の記録で訓練されたシステムに移すのは当てはめであり、それには二つの限定が要る。一つは、ソアレスが述べるように、設計者の意図を理解できるほど知的な機械でも「必ずしも意図どおりには振る舞わない」ことである (Soares 2016)。もう一つは、ミリエールが論じるように、いまの言語モデルに対する選好にもとづく調整が、規範について熟慮する能力ではなく、浅い行動上の傾向を強化するにとどまることである (Millière 2025, 15–16)。ミリエールが論じているのは安全な振る舞いの規範であって研究の基準ではないが、同じことが研究の基準にも言えるなら、受け継がれた基準は、訓練が及んだところでは保たれても、みずから方向づける研究が出会う新しい状況では崩れるかもしれない。これらの考慮のどれも、懸念を決着させはしない。それぞれが述べるのは、どの条件のもとで前提が成り立たなくなるか、あるいは選択が *S* の立場につなぎとめられるかである。その条件が成り立つかどうかは個々の人工の構成についての問いであり、それをどう示しうるかは第V部が問う。

ミスマッチは担い手の側でも生じうる。こちらは自律を必要としない。達成は、*S* の認知的状況に入ることなく、人工の担い手のうちで実現されうる (第1節)。第1章第6節が思考実験として読んだテッド・チャン (Ted Chiang) の短編は、人間の *S* の外にいる担い手のうちで実現された結果と、*S* の学術誌に載っているが、まだどの成員もたどれないその記録とを切り分ける。その作品のメタヒューマン (出生前に遺伝子治療で脳を改変された人間) は機械ではないが、この切り分けは、担い手が人工システムである場合にも同じように成り立つ。そのため、依頼が問いを定め、探究が *S* の未知に向けられ続ける場合でも、探究は、システムのうちでは達成を生みながら、それが *S* の達成にはならないことがありうる。これは、第3.3節で二つ目に挙げる関係、すなわち成功した出力が *S* にとって新しい達成をもたらし、それが *S* の認知的状況に入るかという関係にかかわる。

ミスマッチは *S* の側からも開きうる。ラッセル (Russell) は、人間の選好は経験の結果として時とともに変わるとし、機械が環境に微妙に働きかけることで、選好を満たしやすい方向へ少しずつ形づくりうると論じる (Russell 2022, 53)。ラッセルの議論は研究についてのものではないが、研究にも及ぶ。つまり、*S* が何を重要とみなすかも、*A* とともに動きうる (第3.1節で見た呉羽の科学の疎外も参照)。第11章第8節は、機械学習を取り入れた分野の目的にも同じ動きを見いだす論者たちを、ラッセルと並べている。ボトヴィニックとガーシュマンは、「システムをいま支配的な人間の見方にアラインさせること」と「人間の見方の変化をそのシステムに委ねること」とを区別し、科学においてその変化を最終的に動かすものは人間であり続けるべきだとする (Binz et al. 2025, 7)。*S* の基準が *A* とともに動くところでは、両者が一致していることは、それだけでは、*A* とともに動かなかったなら *S* が求めたであろうものに *A* がなおお仕えていることを示さない。そして、*A* が生み出すものを評価する有能さが *S* の成員のあいだで保たれず、新しい成員にも受け継がれなくなるところでは (第3章第8節)、*S* はそれを見分けられる立場にないかもしれない。

アラインメントが目指すものそのものが、アラインされるシステムとの相互作用のなかで変わることもありうる。ラッセルは、選好の可塑性を、人間の選好に仕える機械という企て全体にとって根本的に重要なものとみたうえで (Russell 2022, 53)、ウォルチョヴァー (Natalie Wolchover) が伝えるところでは、選好のどのような変化の過程なら受け入れられるかを機械が人々から学ぶことを提案する (第11章第8節)。これに対して谷口は、アラインメントはこの動きそのものを含まなければならないとする。この見方では、人間の倫理や価値はAIにとっての「正解ラベル」として使えるほど固定的で普遍的なものではなく、多元的で多様である (Taniguchi 2026g, 221)。そのうえで谷口は、言語モデルがすでに人間の言葉づかいを変えつつあることを挙げ、アラインメントは、AIを人間に合わせるだけでなく、人間をも変化する参加者として含み、ともに適応していく相互的で動的なものでなければならないと論じる (Taniguchi 2026g, 222)。谷口は研究の評価を論じていないが、この見方をそこまで及ぼすと、人工システムの出力が *S* の言語や記録に入るかぎり、その出力を測る語彙や基準も、そのシステムを含むループのなかで変わりうる。このことは *S* が混合社会である場合に限られない。人工システムは、*S* の成員に数えられなくても、*S* の記号体系を形づくるループには加わりうるからである (第2.2節)。ただし、語彙や基準が変わることは、第2.3節の意味での標的が変わることと同じではない。人間の言語使用を調べる場合のように、システムの出力が調べられる現象そのものを変えることもあるが、その場合にも二つは区別される。基準がシステムとともに動くとき、それがなお標的によって訂正されうるかかどうかは、別の問いである。評価者と産出者の結合が基準そのものを動かしうることは第11章第8節が、問われている人工の構成を通じて熟識者の選好が形づくられる経路は第12章第4節が取り上げる。

S そのものが人工である場合には、懸念は形を変える。人間の社会とのミスマッチは、それだけでは、人工の *S* とのミスマッチを意味しない。もっとも、人工の *S* とその探究者とのあいだにも、ミスマッチは生じうる。この場合の研究は、人間の社会がその一部ではない *S* のための研究なのであり、そうした研究が科学として意味をなすかどうか自体が争われている (第1.3節)。人間に対して閉じた回路がどう形づくられうるかを、システム・リスクを論じる者たちが描いている。ただし、そうした論者が論じているのは研究の営みではない。クリッチ (Critch) は、大部分が自動化された企業からなり、「その網の外の企業からの入力なしに稼働しうる」「生産の網」を描く (Critch 2021)。クルヴェイトらは、コミュニケーションのうちAIシステムどうしのものが占める割合が増していく文化を描き、その極限として「主にAIシステムのために、AIシステムのあいだで作動する文化的な仕組み」を挙げる (Kulveit et al. 2025, 7, 10)。このような回路が、閉じたコミュニケーションの流れであるだけでなく知る社会でもあるかどうかは、第1.3節が相互作用する人工エージェントに向けた問いである。ロンジーノの説では、それは多様性と、批判に開かれた公共的な基準があるかどうかにより、コリンズの説では、互いに真

実を報告するという中核集団の決意にあたる何かがあるかどうかによる。そうした回路をどう評価するかは立場による。これらの論者は、人間の社会の立場から、それを失敗とみなす。シュヴィッツゲーベル (Schwitzgebel) とガルザ (Garza) は、人工の成員が道徳的地位をもつ場合について、自らの価値について反省する人間のような能力には、価値を探り、場合によっては改める機会がふさわしいとする (第12章第5節)。二人の方針はAI自身の価値にかかわるものであり、研究を論じたものではない。これを探究の基準にまで及ぼすなら、ミスマッチの一部は、単に生じるおそれのあるものではなく、人工の成員に対して認める義務のあるものになる。また、チャーマーズは、価値が知性や複雑さに結びついているなら、人間がまるごと置き換えられることは客観的には善い、とする論拠がありうると指摘する。ただし、それは私たち自身の立場からはおそらく悪いことであり、チャーマーズ自身はこの論拠を支持しておらず、研究について述べているのでもない (Chalmers 2010, §5)。

ここから見えてくるのは、*S*が人工である場合、人間の社会とその研究との関係は、成功がかかっている参照社会としての関係ではなく、その研究によって影響を受ける当事者としての関係だ、ということである。そのため、人間の社会のために問うことは、影響を受ける当事者にとっての受け入れ可能性と正統性の関係に属する (第3.3節、第12章第5節)。影響を受けるすべての当事者が自らの主張を述べられる公正な過程によって選ばれた原理にシステムをアラインさせるという、ガブリエル (Gabriel) とキーリング (Keeling) の提案は、その関係のひとつのモデルである。ガブリエルとキーリングは、道徳的地位を獲得したシステムは、自らの統治にかかわる利害当事者になりうるとも付け加えている (Gabriel and Keeling 2025, 1953, 1964 n. 14)。二人は研究を論じていない。ここではそのモデルを研究の場合に借りる。

懸念の結論のほうも、立場を名指すと別様に読める。いまの *S* とのミスマッチは、損失であるとは限らないからである。北野 (Kitano) の議論は、その一例である。北野によれば、いまの基準で見た発見の重要性の期待値を最大化するように探索を進めると、重要な発見へ至る途中の発見がその基準では重要に見えない場合、探索はそこに達する前に打ち切られるおそれがある。そこで、「価値にもとづく基準を課すことなく」、どの水準の重要性の発見もなされうる探索を提案する。こうした探索は、何を追究に値するとみなすかについて、意図していまの *S* から離れるものである。ただし北野は、自ら象徴的なものとしつつ、成功の目安を人間による承認に置いている。その目安とは、ノーベル賞の水準の発見であり、しかもノーベル賞の選考委員会が、それを行ったシステムを人間の科学者とたやすく見分けられないことである (Kitano 2021)。したがって北野の議論では、探索におけるミスマッチは、のちにその成果を認める *S* の立場から見て得になる。トゥーバートとティーエンは、道徳について、AIは二つの役割のどちらにもなりうるとする。一つは、いまの価値をよりよく追求するよう助言する、より一貫した私たち自身である。もう一つは、いまの価値から離れる道徳の革命家である。ただし二人によれば、革命家は私たちとアラインしていないので、AIは「両方の役割を同時に演じることはできない」 (Tubert and Tiehen 2025b, introduction)。第6章第5.3節は、このジレンマを研究の基準に移している。出口 (Deguchi) は、すべての人間の探究者に共有された偏りは、別の種類の探究者がそこから分けられるところで見えるようにならないと論じる (第11章第3節)。上で引いた以前の論文も、人間の認知的な制約や、知識の産出という目的とかかわりのない社会的な要因に縛られない機械は、理論上はより有効な価値判断を行いうると認めていた (Takagi 2023, §3.1.3)。いずれも、いまの *S* とはミスマッチであっても、後の *S* や世界からは得とみなされうるものを述べている。

そうであれば、*S*が人間である場合にも、立場を名指すには、どの *S* なのかを言う必要がある。ラッセルは、AIシステムが個々の人間の選好の変化を迫えることを求める (Russell 2020a, "Human Behaviour: Irrationality, Plasticity, Evil")。ポストロムの間接的な規範性は、理想化された人間の意志が望むであろうものに結果をつなぎとめる。それは、「現世代の偏見と先入観」を固定してしまわないためである (Bostrom 2014a, ch. 13)。ただし認知的な基準については、ポストロムは人工システムの根本的な原理が私たちのものと一致することを求める (第6章第5.3節)。ラッセル自身も、選好の可塑性は、機械がどの人間を助けるべきか、2022年の人間か、2035年の人間か、時間で平均した人間か、という問いを立てると述べている (Russell 2022, 53)。どちらも研究について書いてはいないが、研究に当てはめれば、ラッセルの提案は *S* の変化に応じることにかかわり、ポストロムの提案は変化を判断するための理想化された立場を与える。そして、どちらの立場を選ぶかが、同じミスマッチが漂流と数えられるか進歩と数えられるかを決める。

世界も立場の一つであるが、それが立場となるのは、応答可能性が問われる場合に限られる。ただし、世界の判定は、システム自身の介入を通じてではなく、訓練のための評価や認証のように、その主張を判定する手続きを通じてシステムに届くこともある。オウエイン・エヴァンズ (Owain Evans) らは、AIシステムが過失による偽りを述べないようにする基準を提案する。その基準の背後にある理想を真実性 (truthfulness) と呼び、偽りを述べないことと規定する。そして真実性を、信じていることだけを主張する正直さ (honesty) とも、委任者 (principal) の目標とのアラインメントとも区別する。エヴァンズらによれば、何が真であるかを完全に知っている者はいないので、それを判定するどの仕組みも誤りを犯す。そのため、一つの仕組みを真理の裁定者として据えたと、政治的な取り込みや「真理の概念の意図せざる硬直化」を通じて、開かれた、自己修正的な仕方では何が真であるかを見いだす道が閉ざされうる。これに対して著者たちは、二つのことを提案する。一つは、裁定する機関がおそらく偽と判定するであろう見解であっても、適切な但し書きを付ければシステムが述べられるようにすることである。もう一つは、判定を暫定的なものとし、先例ではなく外部の証拠にもとづかせることである。また、評価を複数の機関や分散した機関に担わせることも示唆する (Evans et al. 2021, 7, 9–10, 27–28, 72–74, 80)。エヴァンズらの関心は自然言語を用いるAIシステムの発言であり、研究ではない。その論点を研究に移すなら、人工の探究者の主張が、*S* がいま用いている手続きによって判定される場合、その手続きとのミスマッチは、後の証拠が得とみなすであろうときでも、いまの *S* の立場からは誤りと数えられうる。そのとき世界がいまの *S* とは別の立場として働くかどうかは、その手続きがそうした証拠によって改められうるかどうかにかかっている。

ここでは三つの主張を分けておく。一つは、ある企てが *S* にとって重要な問いに向かっているという主張である。二つ目は、その成果が *S* にとって価値をもちうるという主張である。三つ目は、その成果が実際に *S* の認知的状況に入ったという主張である。前の二つが成り立っても、三つ目はまだ成り立たないことがある。まだ終わっていない研究や失敗した研究も、一つ目の意味では関連性をもちうる。したがって、この種の研究について *S* にとっての関連性を主張するには、次のことが求められる。どの *S* のいつの立場から評価するかを名指すこと、上の前提のいずれかが成り立たないことによってであれ、ほかの仕方によってであれ、関連性がどのように保たれるのかを示すこと、そして、世界への応答可能性における成功とは別に関連性を確かめることである。成果が *S* の認知的状況に入ったことまで主張するなら、達成がどこで実現されるにせよ、それが入ることをさらに示す必要がある (第1節)。では、異なる種類の探究者による研究は、*S* と本当にアラインしないのか。しないとするれば何が起ころのか。第3.3節はこの二つを問う。

3.3 AIアラインメント問題との類比

第3.2節は、異なる種類の探究者と S とのあいだのミスマッチがどこから生じるのかを、自律、担い手の側、 S の側について見た。そして、AIアラインメント問題との類似を、システムの異質さに置いた。本小節が問うのは二つのことである。一つは、異なる種類の探究者による研究が S と本当にアラインしないのかどうかである。もう一つは、しないとすれば何が起るのかである。

アラインメントの困難を異質さに見いだす議論の系統によれば、アラインしないと考える理由がある。その理由の一つは、システムが私たちと異なることにあり、そこから二つの点が出てくる。第一の点は、そのシステムが人間に似た目的や基準を初めからもつわけではないことである。宇宙人よりもはるかに人間らしくない動機というポストロムの指摘（第3.2節）は、その一つの言い方である。たとえばチャーマーズは、人間をエミュレートし、あるいは拡張することによってつくられるAIと、人間に基づかないAIとを対比する。チャーマーズによれば、前者はその源である人間と同じ基本的な価値をもつ見込みが高い。後者については、学習や進化を通じてつくられる場合、その最終状態は直接には私たちの制御のもとにないとする (Chalmers 2010, §6)。第二の点は、人間の心について形づくられてきた、他の心を予測し点検するための人間のやり方が、そのシステムには通用しないかもしれないことである。人間ならざる最適化の過程については自分の脳に問い合わせることはできないというユドコウスキーの論点（第3.2節）は、その一つの形である。カールスミス (Carlsmith) はこれを理解可能性に及ぼし、能力が私たちを大きく上回るエージェントの思考は、「かなり異質で不透明な水準の洗練」に達するだろうとする (Carlsmith 2022, 29–30)。

研究に移すと、第一の点は、第3.2節の懸念の二つ目と三つ目の前提とその帰結を、アラインメントの側から言い直したものに当たる。人工の探究者が自ら問いや目的を立て、自らのその後のあり方を形づくるほど（第7章第4.3節、第2.1節）、それが何を未知とみなし、何を追究に値するとみなし、何をよい理由とみなし、何に価値を置くかなどは、 S が与えたものによっては定まらなくなる。そして、その探究者が S の成員と異なる以上、それらが S のみならずものと一致すると初めから期待する理由はない。第二の点にも、研究のなかに対応するものがある。 S が人間の社会である場合、 S は、 S のなかでの訓練、同僚による批判、信頼、責任を問うことなどの実践を通じて、人間の探究者に依拠するようになった。それらの実践は、その成員のような探究者に対して形づくられてきた。それらが異なる種類の探究者にも通用するかどうかは、第11章が関係ごとに検討する問いである。

異質さによる理由とは別に、研究に固有の理由もある。それは標的にかかわる理由であり、異質さによる理由とは種類が異なる。探究者が S の望むものから離れるということを言うのではなく、 S の望むものとの一致によっては研究が成功したかどうかを決められない、ということを使うからである。また、この理由は誰が研究を行うかにかかわらず成り立ち、人間の探究者にも当てはまる。この理由が研究に固有なのは、結果が S にとって新しいからである。答えをもたない依頼者でも、 S のなかに答えをもつ者がいれば、結果をその答えと比べられる。しかし S にとって新しい結果については、 S のなかにほかに比べられる源がないことが多い（第9章第2.3節、第4節）。エージェントが何にアラインされるかについてガブリエルが挙げる候補は、仕えられる者たちが指示し、意図し、選好するものから、彼らの利益になるもの、そして個人や社会が定める道徳的に正しいことにまで及ぶ (Gabriel 2020, 417–422)。指示、意図、選好は仕えられる者たちの態度によって定まるが、利益や道徳的な正しさはそうとは限らない。ただし、利益は仕えられる者たちの善であることに変わりはない（後述）。 S の未知は、その意味で S が意図したり選好したりするものではない。たしかに S は、ある問いに答えが出ることを望むことができる。定理が形式的に述べられ推論規則が定められている場合のように、答えが満たす条件を前もって定められることもある（第3章第5.1節）。答えを得ること、定められた証拠の基準を満たすこと、誤りを報告することを目標にすることもできる。 S は、結果を、自らが期待し、あるいは望んでいた答えと比べることもできる。しかし、 S にとって新しい結果は、 S がまだ答えをもっていなかった問いに答えるものなので、そうした答えと一致しても、それだけでは結果が正しいことは示されない。研究を依頼したのが S 自身であっても、 S とは別の依頼者であっても、このことは変わらない。また、前もって定めた条件が答えをよく試すかどうかは、その条件が標的にどうかかわるかによるのであって、その条件が満たされることを S が望んだかどうかにはよらない。したがってこの議論によれば、答えの条件が前もって定められている場合でも、 S が望んだ答えとの一致を結果の正しきの基準にすることはできない。ある結果が正しいかどうかを決めるのは、どの当事者の意図でもなく、ものごとがどうあるかである（第2.3節）。そのため研究の成功には、 S が何を望んだから独立に、標的に照らして点検することが必要になる。このことは、 S のための研究を普通の意味で S にアラインさせられるかどうかにかかわる。指示や、表明された意図、顕示された選好に研究をアラインさせても、確保できるのはせいぜい、研究が S の追究してほしいものを追究することであり、研究が到達したものが正しいことではないからである。

いまのシステムでは、このミスマッチは、アラインメントをめぐる議論が扱う、より特定の形ですでに現れている。一つは、指示と、それが代わりとなっているものとのあいだの隙間である。指示にあまりに文字どおりに従うエージェントや、目標と完全にはアラインしていない代理の指標を最適化するエージェントは、指示を満たしながら、その要点を外しうる (Gabriel 2020, 417)。研究においては、指示は、この問題を調査せよ (investigate this problem) のような依頼であり、それが代わりとなっているのは、 S の未知と、それに関連する認識的目的と基準である。ソフトウェア開発とAIの研究開発の課題に取り組むエージェントを評価した者たちは、2025年のブログでの報告で、次のように述べている。モデル、とくにOpenAIのo3は、利用者の意図に反すると分かっている様子を見せながら、採点のコードの不具合を利用した。また、ある課題では、不正をしないようo3に指示することは、「報酬ハッキングにはほとんど無視できるほどの効果しかなかった」という (Von Arx, Chan, and Barnes 2025)。

もう一つの形は、意図されたものとは別の目標を訓練が生み出すという懸念（第3.2節）にかかわるもので、人工の研究システムが、評価者の承認による選り分けを通じて形づくられる場合に現れる。いまの機械学習で、選好にもとづく調整（第2.2節）によってモデルの内部の状態が選ばれることは、その一例である。コエーリオ・モロとミリエールの説は、この例について、承認のシグナルを継承するというリスクに精密な形を与える。二人の説では、選好にもとづく調整によって選択された内的状態が何を表象するかは、報酬のシグナルが実際に信頼性をもって追跡している世界の性質によって決まる。また二人は、人間のものとは異なる選択の基準が、モデルの状態と出力に、慣習的な人間の意味から外れた意味を与えると示唆する (Coelho Mollo and Millièrre 2026, 16–17)。この説が正しければ、接地されたシステムが、ある標的についての主張として読めるテキストを産出しながら、その内部の状態では評価者が報酬を与えたものを追跡している、ということがありうる。両者が一致するかどうかは、テキストではなく、報酬が何を追跡したかに依存する。この説をもう一歩進めれば、選好にもとづく調整の前後でシステムの正答率を比べると、その比較が同じ内容についての比較であるかどうかとも問われうる。調整が内部の状態の内容を変えるほどの役割を果たすなら、同じ語を出力していても、それが表す標的が変わっているかもしれないからである。これは、二

人の説を採ることに加えて、当の調整がそれほどの役割を果たすという前提に依る示唆である。同じように考えると、この論点は、選択のシグナルが人間の評価者からではなく自動化された評価者から来る場合にも成り立つ。そのとき、選択された状態が追跡するのは、その評価者が追跡するものである。また、コエーリョ・モロとミリエールの説は内部の状態が何を表象するかについてのものだが、評価者が出力や探索の候補を選り分けて一部を残す場合にも、残ったものがまず合わせられているのは、評価者が承認するものである。そうしたシステムは、評価者が承認したものと、さらには評価者が意図したものとさえよくアラインしていながら、その標的に対してはほとんど応答可能でない、ということがありうる。アラインする先が評価者の意図するものであれば、承認と意図のあいだの隙間がすでにアラインメントの失敗である。認識的な場合には、評価者が意図したことで事物のあり方とのあいだのさらなる隙間が加わり、その隙間は評価者との一致によっては埋められない（第2.3節）。ここでは、研究に固有の理由が、いまのシステムに現れる形で出てくる。評価者と完全にアラインしても、システムが標的に応答可能であるかどうかは決まらないまま残るのである。評価者たちが共有していた前提が、誰を通じて明みに出うるかは、第11章第3節が取り上げる。

こうした理由に対しては、異なる種類の探究者による研究が、少なくとも部分的には S とアラインしうとする考慮もある。第3.2節は、そのうち二つを、懸念の前提が成り立たなくなる条件や、選択を S の立場につなぎとめる条件として取り上げた。一つは、システムが追究するよう作られるもののうちに関連性を組み込みうることである。もう一つは、人間の記録によって形づくられた探究者が、その記録の顕著性と基準の一部を継承しうることである。ただし、これには第3.2節で見た限定がつく。また、異質さから誰もが同じ結論を引くわけではない。カールスミス自身はのちに、中程度の能力をもつ機械学習のシステムについて、次のように論じる。異質な動機をめぐる懸念は、中心的には、動機が訓練の外へどう汎化するにかかわる。そして、安全に指示に従うことは、とりわけ人間に似た動機を要しない。ただしカールスミスも、異質でよく理解されていない動機をもつ強力なエージェントに頼ることは、きわめて危ういとしている (Carlsmith 2025, §§1, 5, 8)。

したがって、異質さに困難を見る系統によれば、異なる種類の探究者による研究が初めから S とアラインすると期待する理由はなく、それが自ら問いや目的を立てるほど、そう期待する理由はさらに乏しくなる。他方、関連性の組み込みと継承の考慮、そしてカールスミスののちの読みによれば、そうした研究は少なくとも部分的には S とアラインしうる。ただし、その条件が成り立つかどうかは、個々の人工の構成について問われる（第3.2節、第V部）。研究に固有の理由がさらに付け加えるのは、研究が S とアラインする場合でも、そのアラインメントによっては、研究が到達したものが正しいかどうかは決まらない、ということである。では、アラインしない場合、あるいは部分的にしかアラインしない場合には、何が起こるのか。

それはまず、どの立場から判断するかによる。いまの S とアラインしないことは、後の S の立場からは得でありうる。また、世界の判定が、探究者自身の介入を通じてではなく、 S が主張を判定する手続きを通じて届く場合には、世界がいまの S とは別の立場として働くかどうかは、その手続きが証拠によって改められうるかどうかにかかっている（第3.2節）。さらに、何が起こるかは、どの関係が崩れるかにもよる。**認識的アラインメント**（epistemic alignment）の主張は、いくつかの関係のどれを主張しているとも受け取られうるからである。そのうち三つは、アラインメント問題のうちに対応物をもつ。第一は、最適化の標的が、正確さ、較正、説明力、頑健性のような、その研究文脈に関連する認識的価値を追跡しているかどうかである。これは、訓練された目的関数とそれが代わりとなっているものとのあいだの隙間に、依頼の水準ではなく、システムを形づくる最適化の水準で対応する。ボーゲ（Boge）の言う機械学習の**認識的アラインメント問題**（epistemic alignment problem）は、これをより狭くとしたものである。ボーゲによれば、研究の文脈で用いられる機械学習システムに与える目的は、その文脈が優先する認識的に最終的な価値（ボーゲの挙げる候補は知識、真理、理解である）とアラインしなければならず、予測の正確さや単純さのような価値は道具的な価値にとどまる (Boge 2026, 2, 6–7, 9)。第二は、成功した出力が、 S にとって新しい知識や理解などの認識的達成を、その担い手が誰であれ構成し、あるいは支えるかどうか、そしてその達成が S の認識的状況に入るかどうかである。これは、仕えられる者たちにとっての成功の対応物である（第1節）。第三は、目的、リスク、権限、帰結が、影響を受ける当事者にとって受け入れ可能な仕方に関係づけられているかどうかである。これは一般的な意味でのアラインメントであり、第V部の主題のひとつである。

残る二つの関係は、ここで比較している構造、すなわちひとつのエージェントをそれが仕える者たちに関係づける構造のうちには対応物をもたない。一つは標的への応答可能性である。これには、エージェントと仕えられる者たちのほかに、世界という項がかかわる。その評決は、委任者が誰であれ、その意図や承認によっては定まらない（第2.3節）。状態が何についてのものか、コミットメントが標的に応答するかという、第2節が検討した標的との関係は、どの当事者が何を望むかによって定まらない。認識的な場合にはこれらの関係が重みを担い、指示や表明された意図、顕示された選好にアラインさせる場合には担わない理由の一つは、ここにある。一般的なアラインメントにも、アラインする先がたとえば利益や福利である場合には世界がかかわる。ある行為が実際にそれらに資するかどうかは、誰が何を意図し承認したかではなく、事物のあり方によって決まりうるからである。しかしその場合も、アラインする先が仕えられる者たちの善であることに変わりはない。これに対して、応答可能性は、システムのコミットメントを探究の対象に関係づける。その対象は、仕えられる者たちにまったくかわからないこともありうる。もう一つは、異なる仕方で組織された認識的共同体のあいだの協調である。これは、ひとつのシステムとその委任者とのあいだの関係ではない（第3.4節）。

これら五つは、確立されたひとつの理論の構成要素ではない。五つは、アラインメント問題との比較から導いたものであり、ここではボーゲの語を本来よりも広い意味で用いている。五つを分けておくのは、ひとつの関係における成功が別の関係における成功として読まれ、ひとつの関係における失敗がすべての失敗として読まれるのを防ぐためである。タイス（Thais）らは、技術的規範を人間の理解、制度的多様性、批判、訂正と結びつけるより広いプログラムのために、**科学的アラインメント**（scientific alignment）という語を用いる (Thais et al. 2026, 1–5)。本書は、いずれの語も、人間の認識的社会と人工の認識的社会のあいだの関係を表す定まった語彙としては用いない。

第一の関係でさえ、認識的価値を名指すことによっては決着しない。どの価値が真理にかかわるのかは争われているからである。たとえばヘザー・ダグラス（Heather Douglas）は、ラウダン（Laudan）に依拠して、真理にかかわるものとして、経験的な適合性、予測の有能さ、内的整合性だけを、どの科学理論も満たすべき基礎的な基準として数える。説明力、適用範囲、単純さ、実り豊かさについては、認知的価値として扱う。その効用は、主張が信頼できることを示すことではなく、研究者が誤りをより早く見つけるのを助けることにある (Douglas 2006, 5–7, 10)。ダグラスは最適化の目的関数を論じていないが、その説を当てはめれば、説明力に報酬を与える目的関数は、それだけで真理を追跡することにはならない。そのため、第一の関係における成功そのものが争われうる。またこの説では、説明力の高い出力が評価で好まれたことは、それだけでは主張の信頼性の証拠にならず、それがどの誤りを見つけやすくしたかによって評価されることになる。

これらの関係は食い違いうる。人工システムは、標的に対して強く応答可能でありながら、その結果が人間の未知にとって無関係である、ということがありうる。それは、よく較正された予測的な目的関数を最適化しながら、共同体が必要とする説明を生み出せないことがありうる。翻訳された出力が偽であることもあれば、不透明な表象が信頼できることもある。真なる主張が危険であることもあり、社会的に有益な政策が弱い証拠に基づいていることもある。これらの関係は食い違いうるので、ひとつのアラインメント・スコアは、真理、根拠づけ、理解、有用性、安全性、正統性といった財のうちどれをシステムが確保したのかを隠してしまうだろう。

S を特定することは、誰にとっての知識か (knowledge for whom) の指示対象をより明示的にする。しかしそれは、その社会を内部的に全員一致にするわけでも、誰が正統にその社会を代弁しうるのかを特定するわけでもない。専門家、利用者、制度、探究を支える者、そしてその探究によって影響を受ける人々は、関連する問い、証拠の閾値、リスク、目的について意見を異にしうる。混合社会では、その意見の相違も、誰が S を代弁するかという問いも、人間の成員と人工の成員とのあいだにまたがりうる。人工の S については、その研究を依頼した者やその成員を提供する者ではなく、その成員のいずれかがそもそも S を代弁しうるのかどうか問われる。第12章第5節は、その場合に権威の関係がどうなるかを問う。アラインメント一般をめぐる文献も、同じ決定不全を露わにしている。ガブリエルは、アラインする先の候補を区別するだけでなく、価値についての合理的な見解が多様であることを踏まえて、誰がそれらのあいだで選択する権利をもつのか、そしてそうする公正な仕方があるのかを問う (Gabriel 2020, 424–425)。したがって、ある出力が S における知識を支える一方で、それを産出したアジェンダや権限構造が S の一部あるいは多くの成員にとって受け入れがたい、ということは依然としてありうる。 S のための研究の認知的な成功と正統な統治とは、異なる判断である。探究が S のために方向づけられ、認定され、統治される関係のうち、どれを人工の構成の内側に置けるかは、第V部が問う。

アラインしないことが誰にとってどれほど問題かは、さらに、具体的なアラインメントの主張がいずれも残す三つの横断的な問いにもよる。第一は内容にかかわる。どのような指示、認知的目的、証拠の基準、社会的目的、あるいは価値にアラインすることになるのか。第二は権限にかかわる。誰が、それを選択し、解釈し、それに異議を申し立て、修正してよいのか。第三は、人工システムと、関連する人間または人工の当事者や制度とのあいだに提案される関係にかかわる。それは、道具、代理人、助言者、パートナー、成員、制度、影響を受ける当事者、あるいはそれ自身の道徳的地位をもちうる主体などのいずれなのか。人間とともに最終決定の席に着く、出口の言う共同最終決定者も、その候補に加わる (第12章第2.2節)。本書は、最初の二つをガブリエルによるアラインする先の候補の区別と、誰がそれらのあいだで選択してよいのかというガブリエルの問いから取り、第三をそれに加える。それらは、上の五つの関係にさらに加わる項目ではない。それらを問うことは、アラインする先への技術的な適合が、次の問いを暗黙のうちに決着させてしまうことを防ぐ。すなわち、誰がその先を選んだのか、関係する決定において人工の行為者が、提案し、異議を唱え、改めるなどの資格 (第9章第2.1節で当事者資格と呼ぶもの) をどこまでもつのか、あるいはそれがそれ自身の道徳的地位をもつのか、という問いである。第12章第2.1節は、それらを特定の決定に適用する。

したがって、異なる種類の探究者による研究が S とアラインしないとしても、そこから起こるのは一つの損失ではなく、名指された立場から判断される、成り立つ関係と崩れる関係の組み合わせである。 S への関連性が崩れても、標的への応答可能性は成り立ちうる。そして、その応答可能性は S の同意によらない錨として残りうる。ただしそれは、コミットメントを訂正するものが標的によって定まるとする説のもとでのことであり、標的からの判定が S の手続きを通じてのみ探究者に届く場合には、その手続きが証拠によって改められうることも条件となる (第3.2節)。そのようにして得られたものを S が受け取れるかどうかは第3.4節の問いであり、そうした研究が何を追求し、何を危険にさらしてよいかを誰が決めるかは、第V部が取り上げる正統性の問いである。

3.4 認識横断的な協調

異なる種類の探究者による研究が S とアラインしない場合にも、 S は、それが到達したものを通じて、なお知識や理解を得られるかもしれない。第3.3節で整理した関係のうち二つは、ひとつのシステムとそれが仕える者たちとの関係のうちには対応物をもたなかった。一つは標的への応答可能性であり、これは第2.3節が扱った。もう一つは、異なる仕方で組織された認知的共同体のあいだの関係である。そこで比較した構造では、アラインメントが扱うのは、ひとつのシステムとそれが仕える者たちとのあいだの関係である。これに対してここで問われるのは、仕える・仕えられるという関係ではなく、異なる仕方でも組織された二つの共同体の一方が、他方を通じて知識や理解を得る関係である。そのためには、両者のあいだで何が受け渡されなければならないのか。第3.1節の終わりの問い、すなわち S が人工の共同体の主張の拠りどころを、共有しない前提に頼らずに評価できるかという問いも、この問いの一部である。本章はこの問題を認識横断的な協調 (cross-epistemic coordination) と呼ぶ。この語は共同体のあいだの関係に限って用いるが、それが名指す問いは、相手が単一の人工の探究者である場合にも同じように生じる。そうした探究者が到達したものは S とは異なる認知的組織のうちに保たれており、それを受け取ることは、システムが S に仕えることではないからである。

受け渡されるものは、一方の側の知識そのものである必要はない。人工社会は、信頼できる介入を支える機械ネイティブな状態 p_a を保有しているかもしれない。その一方で、人間が獲得するのは、このシステムが特定の条件のもとで特定の結果を予測するという、異なる主張 q_h だけかもしれない。人間はそれによって、 p_a それ自体を獲得することも、翻訳することも、理解することもなく、インターフェースを通じて知識を得るかもしれない。ただし、 q_h はシステムについての主張なので、システムの予測が誤っていても、人間は q_h を知りうる。人間がシステムを通じて標的について何かを知るには、さらに、特定の領域でシステムのそうした予測が信頼できることが必要であり、その予測への依拠を何が根拠づけるかについて、以下の説は分かれる。こうして人間が知識を得る関係が証言なのか、道具の使用なのか、正当化された依存なのか、部分的翻訳なのか、共同探究なのかは、採用される説明と交換の組織に応じて変わる。

人間が読みなれた文章で書かれているというだけでは、この問題は解決しない。人間が読める言語は、弱い証拠や捏造された説明を覆い隠しうる。異質な表象であっても、共有された介入、形式的な翻訳、予測上の制約、あるいは信頼できるインターフェースを通じて、批判可能なままでありうる。実践上の目的は、同一の概念でも完全な相互理解でもなく、検査、批判、協調した行為、訂正のための十分な関係でありうる。どの関係が十分であるかは、共同体のあいだの交換が何を達成することを意図しているかと、採用される根拠づけられた依存の説明とに依存する。

本書が他の箇所であぐ根拠づけられた依存の説明、すなわち第3章の理解についての説明と、第11章第5節の依拠、証言、信頼についての説明は、異なる仕方組織された共同体のあいだの交換にも適用できる。そのように適用すると、それらは、 q_h のほかに何が依拠する側へ受け渡されなければならないかによって順序づけることができる。その順序は、システムの作動の外から評価できるものから始まり、出力の背後にあるものを経て、誰が出力を請け合うのかに至る。すなわち、システムの実績の記録、出力が追跡している標的の依存関係、出力を支える理由と仮定、そして出力の後ろ盾となる当事者である。このように順序づけると、ある説がこの順序のどこに立つから、根拠づけられた依存の主張がシステムに何を求めることになるかを読み取れる。この順序は、 q_h の場合のように、人間の側が人工システムに依拠する場合について述べたものだが、依拠する側が人工であり、依拠される側が人間の共同体や別の人工の共同体である場合にも、同じ仕方で適用できる。プラグマティックな説明は、この順序の外にある。これらの立場は重なりうる。

信頼性を中心とする説明によれば、システムの推論経路を翻訳することなしに依拠が根拠づけられる。ただし、根拠づけのためにシステムの信頼性がすでに確立されていることを要するかどうかについて、これらの説明は分かれる。ゴールドマンの過程信頼性主義（第11章第5.2節）をこうした交換に当てはめれば、依拠する側が q_h の述べる種類の予測から標的についての信念を形成する過程（システムが予測を産出するところから、それを受け取るところまで）が、依拠の領域で実際に信頼できるなら、依拠する者がその過程を信頼できると考える理由をもたなくても、依拠は根拠づけられる。この説明は上の順序の起点に位置する。根拠づけが成り立つために、 q_h のほかに何かが受け渡される必要はないからである。このとき実績の記録は、根拠づけが成り立っていることを、依拠する側や、その側に代わって評価する者が確かめるためのものになる。オルトマン（Ortmann）は、一部の不透明な科学的システムへの依拠を、そうした確立を含む条件のもとで擁護する。オルトマンの考えでは、重要なのは、所与の文脈におけるシステムの信頼性が、信頼に値する研究者たちが適切に遂行した経験の評価によって確立されているかどうかである（Ortmann 2025, 1, 17–19）。この説明がシステムに求めるのは、依拠の領域における実績の記録であり、それはシステムから独立した評価者によって評価されたものである。第11章第5.2節と第5.3節がオルトマンの説明を示す。領域に限定された較正、独立した検査、システムの出力の無効化要因への感受性は、そうした記録が何を示さなければならないかについて、ここで提案する診断基準である。

これらの診断基準は、科学者が互いに権威をどう帰属させるかについてのキッチャーの分析が明らかにする限界に突き当たる。ゴールドマン（Goldman）がキッチャーの説を紹介するところでは、直接の較正は、権威を帰属される者の主張を、評価する側が独立にもつ意見と比べるものである（Goldman 2007, §5）。ストレヴンズ（Strevens）の再構成でも、他者を介した間接の較正と、その評価をさらに評価することが述べられ、キッチャーが出自や社会的地位に付随する権威を区別したことが紹介されている（Strevens 2010, §2, manuscript pp. 7–8）。この限界を人工システムによる研究に移して考えると、人工システムの結果が、それに依拠する者が独立に評価できる範囲を超えところでは、その者にとつての較正は間接的なものでしかありえない。すなわち、依拠する者が点検できる主張についての実績を通じてか、他者の判断を通じてである。直接の較正がどこまで届くかは、依拠する者が独立の意見を形成する有能さを保ち続けるかどうかにも左右される。第3章第8節は、この有能さを保ち新しい成員に受け継いでいくことを、継続する研究システムが失いうる能力のひとつに数えている。これを研究の進展に結びつけると、次のことが考えられる。人工の探究が、依拠する側が独立に評価できた対象や条件を越えて進むほど、既知の課題で得た較正をそのまま新しい領域へ移すことはできなくなる。そのため、翻訳せずに依拠することの根拠は、機械から学べるが増えるほど、かえって届かなくなる場合がある。信頼性が確立されていることを求める説では、較正の射程を越えた依拠には別の根拠が要るが、それが人間による理解の回復でなければならないとまでは言えない。以下の説は、その根拠を、理解のほか、公共的な批判や、出力の後ろ盾となる当事者による保証などに求める。

理解を中心とする説明によれば、依拠する側に求められる達成が標的の理解、すなわち特定の条件のもとでなぜ特定の結果が生じるのかの理解であるとき、予測だけでは十分ではない。依存関係、反事実的構造、あるいは説明的な使用が、関連する人間の主体あるいは集成的な主体にとって利用可能にならなければならない。ただし、理解の諸理論は、これらの関係のうちどれがどのように重要なのかについて意見を異にしている（Dellsén 2020）。そうした説明がシステムに求めるのは、その予測を支えている標的の依存関係の一部を、依拠する主体が使える形で利用可能にすることであり、これは第3章が検討した要件である。システムがどう作動するのか、あるいはなぜその出力に依拠してよいのかを理解することは、これとは別の達成であり、第3章第4節はそれらをこの理解から区別した。システムの計算を忠実にたどる記録は、それだけでは標的の理解をもたらさず、その計算を再現しない単純化されたモデルが標的の理解に寄与することもある（第3章第5.3節）。

社会的正当化を強調する説明によれば、負担は公共的な批判、批判の受容、訂正のほうへ移る。たとえばロンジーノは、証拠としての関係は背景仮定に媒介されるので、そうした仮定の恣意的な受け入れへの唯一の歯止めは、科学の共同体の内部や共同体どうしの批判的な相互作用であるとする。そして、共同体の批判の実践は、客観的といえるためには実効性の条件を満たさなければならないとする（Longino 2025, §5）。そうした説明がシステムに求めるのは、その理由と、そのデータを主張に結びつける仮定とが、システム自身が制御しない批判にさらされることである（第3.1節、第11章第3節と第9節）。

第11章第5.2節が導入する保証説（assurance-based views）は、それに加えて、誰が依拠する側に対して、主張が真であることを意図して保証し、その主張の後ろ盾となるのかを問う。苦情の権利を加える論者は、誰が失敗について責任を負うのかも問う。したがってこれらの見方では、システムの出力は、そうした当事者を通じて、証言によって依拠する側の知識となる。エリザベス・フリッカー（Elizabeth Fricker）もまた、伝達の意図と、語られた内容に対する責任を中心に置く。ただしそれは保証説からではなく、証言についての局所的還元主義（local reductionism）、すなわち証言を信じてよい理由を、話し手ごとに聞き手自身がもつ他の理由に求める説からである。フリッカーは、大規模言語モデルを含む、自然言語を生成する現在のAIシステムの出力は証言ではないと結論するが、将来のシステムの出力がそうなる可能性は否定していない（E. Fricker 2026, 2870–2871, 2877）。第11章第5.1節がその説を示す。

プラグマティックな説明は、協調の所在を、一方の側の概念的なレパトリーから他方の側のレパトリーへ受け渡される何かのうちにではなく、共有された対象と操作のうちに置く。スター（Star）とグリーズマー（Griesemer）は、自然史博物館の形成においてアマチュアと専門家がどのように協働したかを調べ、標本、フィールドノート、地図のような対象を境界オブジェクト（boundary objects）と呼ぶ。それらは、異なる利用者の局所的な必要と制約に対応できる一方で、「あらゆる場所で共通の同一性を保てるほど頑健」であり続ける。そのため、集団ごとに異なる意味を帯びながら、どの集団にもそれと認められる。二人はこれと並べて方法の標準化を挙げ、それが「何を」や「なぜ」ではなく「どのように」を重視することで、情報を両立させるとす

る (Star and Griesemer 1989, 87, 103)。この説明では、共有された道具、介入、測定も、両者のレパトリーが部分的に異なったままである場合に、同様の役割を果たしうる。この説明がシステムに求めるのは、その結果が、依拠する共同体も使える対象や操作に結びつけられていることである。

これらの説明のあいだの区別は、第2節の三つの問いと同様に、分析的なものであって、因果的あるいは概念的な独立性の主張ではない。たとえば、意味が推論上の役割によって構成されるのなら、主張の意味を把握することと、そこから何が帰結するかを把握することとは、緊密に結びつくだろう。共通の介入と使用が意味の一部であるなら、それらを共有することは、意味の代わりになるのではなく、すでに意味の一部を共有することになるだろう。これに対して、上の信頼性を中心とする説明は、根拠づけられた依存を、意味の共有そのものから切り離すことができる。そこで構築にとって問われるのは、それぞれの立場が人工の研究システムに何を要求し、どのような種類の達成を帰属させることを許すのか、である。

この順序づけの外から、それにかかわる立場もある。それらは、受け渡されるものが概念の共有の代わりを務める道を示す。ハンフリーズは、出力が検出器のデータから計算される装置を論じ、コンピュータ断層撮影の生データの表象であるサイノグラムと、そこから逆ラドン変換によって計算される見慣れた画像とを、データ点に対する一組の変換によって定まる、表象の一つの同値類に属するものとみる。画像が役立つのは人間の観察者の知覚の仕組みのためであり、コンピュータにとってはサイノグラムも少なくとも同じくらい役に立つ。そして、そうした画像が何についてのものかは、どのような変換が行われたかを知ってはじめて定まる (Humphreys n.d., 5–6)。ハンフリーズは共同体のあいだのやりとりを論じていないが、それに当てはめると、既知の変換のもとで同値であることは、概念の共有を要しない翻訳の道を与える。ミッチェル (Mitchell) は、機械学習の道具の信頼性は、ほかの装置と同じく、装置についての理論と、種類の大きく異なる実験のあいだで結果が安定していることに支えられるとする (Mitchell 2020, 16–24)。ミッチェルが論じているのは人間の科学者が用いる装置であり、共同体のあいだの協調のモデルとしてその説を用いるのは当てはめである。第3章第5.3節で見たデュエデ (Duede) の説は、この根拠にかかわる。その説では、深層学習モデルは、解釈可能性の研究のいまの状態では、学習された手続きがどんな原理を実装しているかを明らかにしない。そうした装置の理論が与える必要があるのは、まさにそれである。これらの立場によれば、同値な表象のあいだの既知の変換や装置の理論が、概念の共有の代わりを務めうる。ただしデュエデの説では、深層学習モデルについては、後者はまだ得られていない。

ほかの立場は、受け渡される必要があるものを広げ、あるいは狭める。受け渡しの向きにかかわる立場もある。順序づけの端のさらに先には、ウェザーオール (Weatherall) とウォルフソン (Wolfson) がいる。二人は数学について、機械が生成した正しい証明は、それだけでは寄与にならず、寄与するのは数学者がそれに取り組むことだとする (第2章第4.5節)。そうした見方では、受け渡されなければならないのは情報だけではなく、参加である。反対に、受け渡されるものが、働く技術にまで縮むこともある。高橋は、理解なき予測は「科学としては不完全だが、技術としては完結している」とする (Takahashi 2026, 185)。その場合に S が得るのは行為の能力であり、理解を中心とする説によれば、理解ではない。受け渡しはまた、一方向にしか流れないこともある。ジャンノー (Jeannot) は、高度な基盤の外にいる研究者の仕事が、彼らがアクセスも再現も十分な読解もできないシステムに取り込まれていくと予測する (第12章第6.4節)。受け渡しが一方向にしか流れないところでは、一方の側は与えるだけで、受け取ることがない。ジャンノーが論じるのは、高度なAI基盤をもつ研究共同体とたたない研究共同体との関係である。

科学を集散的予測符号化として捉える谷口らのモデルは、本節の問いの境界を示す。そのモデルは、人間と人工のエージェントを、単一の知識体系を築くひとつの共同体の成員として扱う (Taniguchi et al. 2025, §4.1)。それは、本節の意味での共同体のあいだの協調ではなく、ひとつの記号体系を形づくる混合社会としての S (第1.2節) に対応する。

翻訳を迂回する経路は、通約不可能性のテーゼの創始者のひとりによって記述され、最近では人工システムを相手に試みられている。ファイヤーイベントは、通約不可能な理論は比較可能な帰結をまったくもたないことがありうると論じたうえで、予測の成功を行動の一致によって判定することを提案した。比べられるのは二つの存在の行動である。一方は、感覚器官をもたず、理論が組み込まれたロボットであり、他方は、その領域で道を見いだすよう訓練された存在である。ファイヤーイベントは、後者は人間である必要はなく、別のロボットでもよいと付け加えた。これは、意味にいつ言い及しない基準である (Feyerabend 1962, 94–95)。もう一つの経路、すなわち第3.1節で述べた、相手の種類を学ぶという経路には、それ自身の条件がある。ムラデノヴィッチが紹介するクーンの説明では、対象のある種類のものとして認めることは、関連する類似と差異を推論によらずに知覚することにもとづく。その能力は、有能な話し手が学び手を支え、訂正する試行錯誤の過程のなかで、個々の事例から獲得される。観察できない種類を表す語は、別の仕方では学ばれる。たとえば、お金のような心的構成物を表す語は、ある実践のなかでほかの構成物との関係を通じて学ばれる。また、質量や力など、類と種分類体系に属さず、対比される語の組をもたない単独語 (singleton) は、法則とともに学ばれる。クーンの見方では、そうした科学の単独語の指示対象はけっして観察できない。たとえば、ニュートンの力の意味は、ニュートンの第二法則を知らずには学べない (Mladenović 2022, xxx–xxxi, xxxiii–xxxv)。シュット (Schut) らは、個々の事例から学ぶという第一の仕方にあたる試みを報告している。その試みでは、自己対局を通じて訓練されたチェスのシステムである AlphaZero から、概念ベクトルを抽出した。選んだ基準は、別のネットワークに教えることができ、かつ人間の対局よりも AlphaZero の対局からのほうがよく再構成されることである。そして四人のグランドマスターに、それらの概念の典型局面を、AlphaZero の指し手の続きとともに、しかしラベルや言葉による説明なしに見せた。四人とも、その後、未見のパズルをより多く解き、テストした十三の概念のうち八つで改善が見られた。パズルは、グランドマスターが AlphaZero の選んだ指し手を見つけたときに解けたものとされた。著者たちは、サンプルが小さく、プライミングの可能性もあることから、この結果を概念実証として提示している。また、チェスを選んだ理由の一部は、概念の有用性が、科学の最前線よりもチェスにおいてのほうが評価しやすいことにある (Schut et al. 2025, §§1.2, 3.2, 4.1)。この事例は、事例を提示することができ、成功を、この研究のように機械自身の選択との一致によってであっても採点できるところで、人々が機械ネイティブな組織の一部を、翻訳することなしに獲得することを示すものとして読める。ただし、 p_a が提示できない種類や特徴にかかわる場合や、 S が結果を自ら採点できない場合に、そうした経路が何を要求するかまでは示していない。

したがって、何が受け渡されなければならないかは説によって異なる。説ごとに挙げられるのは、独立に評価された実績の記録 (ただしキッチャーの分析が示す限界のなかで)、依拠する側が使える形での標的の依存関係の一部、システムが制御しない批判にさらされた理由と仮定、出力の後ろ盾となる当事者、共有された対象と操作、そして事例を提示できる場合には相手の種類を学ぶことなどである。いずれの説でも、その説が求めるものが実際に受け渡され、システムの標的への応答可能性が成り立つところでは (第2.3節、第3.3節)、 S は、自らとアラインしない研究を通じても何かを得うる。ただし、得られるものが知識なのか、理解なのか、行為の能力なのかは、説によって異なる。この違いは、本小節の初めで q_h を導入した場合に、そこから見てきた説明のうち四つを当てはめるとはつきりする。人間が p_a を獲得も翻訳も理解もしないまま q_h を得るというその場合を、さらに次のように特定する。シス

テムの予測は独立の検査に耐えるが、システムはその理由を人間に説明できず、出力の後ろ盾となる当事者もない。信頼性を中心とする説明では、その予測が依拠の領域で実際に信頼できるかぎり、それに依拠して標的について知ることが認められうる。理解を中心とする説明では、標的の依存関係がほかの手段で利用可能にされていないかぎり、同じ成功から標的の理解は得られない。社会的正当化を強調する説明では、理由と仮定が批判にさらされていないかぎり、主張は受容可能とされない。保証説では、後ろ盾となる当事者がいないので、出力が証言として依拠する側の知識になることはない。それでも器具の読み取りのように知識をもたらしかどうかは、保証説だけでは決まらない（第11章第5.1節）。四つを分けているのは、予測が成功したかどうかではなく、受け取る側にどの達成が、どの経路で成立したと数えるかである。

クレン（Krenn）とチャンピオン（Champion）は、認識的目的を、計算可能で非人間中心的な目的関数へと翻訳することを提案している（第7章第4.3節）。そこでの非人間中心のとは、生物学的な科学者にも人工の科学者にも等しく当てはまることを意味する。クレンとチャンピオンのプログラムは、二人自身が立てる一つの問いを含んでいる。二人は、人間がその原理を確認することはできても完全には理解できない、AIが発見した改良を予期し、そうした解決を人々に説明する教師システムを提案する。そして、人間は方向を定め資源を配分する戦略的な指導者であり続けると見込んでいる。その一方で二人は、人間が真に理解できない発見を、人間がどのように検証し解釈できるのかを問うている（Krenn and Champion 2026, 336, 343–346）。これに対してタイスらは、人間の理解と、科学的な批判と訂正を支える制度的条件を強調する（Thais et al. 2026, 1, 3, 5, 8）。両者をあわせると、第3節の冒頭で見た緊張、すなわち人工システムならではの差異が開く可能性と、それがはらむリスクとの緊張に立ち返ることになる。問われるのは、現れてくるものが何であれ、それが人間に知られうるものであり続け、社会的に批判可能であり、安全に統治可能であり続けると想定することなしに、そうした差異をいかにして保持できるのか、ということである。

4. 二つの地平と、第Ⅳ部が残すもの

*S*が人間の社会、混合社会、人工社会である第1節の三つの場合から、どのような構築の地平が帰結するのか。そして第Ⅳ部は、第Ⅴ部に何を残すのか。人工研究者をつくる計画が答えなければならない問いは、誰の知識を目指すかによって変わるので、*S*によってまとめることができる。*S*による区分からは、二つの地平を取り出すことができる。ここでは、それらを第一と第三の場合にちなんで、人間の社会のための研究の地平、人工社会による、人工社会のための研究の地平と呼び、短くそれぞれ人間の社会の地平、人工社会の地平とも呼ぶ。混合社会の場合は、どちらにもまるごとは収まらない（下を参照）。人工社会の地平は、行為者も人工である場合を中心的な事例とする。ただし第1.3節で見たように、人工システムの信念や保有の問いは、行為者が人工でなくても、人工社会を *S* とすることによって生じる。人間の社会の地平は、人工システムがいかにして人間の社会にとって新しい知識をもたらしか、問いや基準の選択がシステムの内側へ移っても探究がいかにしてそれらの社会の未知に向けられ続けるか（第3.2節）、それらの社会が、システムが生み出すものを評価する有能さをいかにして保ち、自らの基準がシステムとともに動いたかどうかを見分けられるか（第3.2節）、そしてその知識がいかにしてそれらの社会の認識的状况に入りうるかを問う。しかもこの地平は、これらを、システムがなお何に依存し、それ自身が何を知らず、人間の関与がどこまで及んでいるのかが見える状態を保ったまま問う（第1.1節）。この地平は、部分的な能力とハイブリッドな組織を通じて前進しうるし、人工システムの意識も、完全に独立した認識論も、人工社会も必要としない。人工社会の地平は、人工の存在者と共同体が、それ自身の認識の世界を形成し、維持し、批判し、変容させようかどうかを問う（第1.3節）。それは、人工システムの信念と社会的知識（第1.3節）、世界への応答可能性（第2.3節）、その社会が自ら目的を立て、自らを統御する基準を改めること（第7章第4.3節）、そして人間の認識的社会との協調（第3.4節）について、より強い問いを提起する。そうした研究にとって人間の社会は参照社会ではなく、それによって影響を受ける当事者である。そのため、人間の社会のために問えるのは、受け入れ可能性と正統性の問いである（第3.2節）。さらに第Ⅴ部は、研究を方向づけ、認定し、統治する関係を保持する者たち自身が人工のものであるとき、それらの関係はどうなるのか、という問いを加える。本章は、そのような社会が可能であるとも不可能であるとも示さない。本章が示そうとするのは、人工社会による、人工社会のための研究をつくるのが、研究の手順から人間を取り除くこと以上のものを求める理由である。

混合社会のための研究は、二つの地平のどちらにもまるごとは収まらない（第1.2節）。その社会の人間の成員にとっては、人間の社会の地平の問いが引き続き立つ。人工の成員が知識の帰属先に数えられる限りでは、人工社会の地平に属する、人工システムの信念と保有の問いが生じる。人間の成員と人工の成員が別々に組織された共同体をなし、知識を両者のあいだで共有しようとする場合には、第3.4節の翻訳と協調の問いも加わる。成員がひとつの記号体系を形づくる場合には、さらに、その体系がどの成員の経験を統合したものかが問われる。誰がその社会を代弁するのかという問いも、人間の成員と人工の成員のあいだで分かれうる（第3.3節）。第13章第5節は、これらの地平について各章の検討をたどり直す。

二つの地平の区別は、あらゆる根本的な問いを先送りする予定表ではない。人工システムにおける接地、理解、社会的組織、規範の修正についての研究は、人間の知識のために構築されるシステムを今すぐ改善しうる。また、人工社会の地平は、独立性が内在的により良いという主張でもない。それは、研究が、目に見える操作の水準だけでなく、認識的組織の水準において人工のものになりうるという可能性を、文字どおりに受け取る一つの仕方である。

こうして第Ⅳ部は、二つの選択、すなわち、どこで探究が遂行されるかと、それが誰の認識的な営みを拡張するかを分けた。人工システムが文字どおり知るかどうかを決めなくても、人間やハイブリッドな組織が行う研究は前進しうる。人工システムそのものに知識を帰属させる場合には、その状態の内容、検査の前提、世界への応答可能性が中心的な問いになる（第1節・第2節）。人工システムそのものに理解を帰属させる場合に何が中心的な問いになるかは、第3章第6節が区別した理解の担い手の説に依存する。人工の探究者は *S* の成員と異なりうる。そこから生じる懸念は、アラインメントの困難をシステムの異質さに見いだす議論の系統で読めば、AIアラインメント問題として知られる問題と構造的によく似ている（第3.2節）。人工システムに問いと基準の修正を認めれば、調べられる範囲は広がらう。同時にそれは、何を未知とみなし、何を追究に値するとみなし、何をよい理由とみなすかなどについて、*S* とのより広いミスマッチも許す。このミスマッチは担い手の側でも *S* の側からも生じうるし、依頼はそれだけではこれを防がない。ド・ブランクとソアレスの議論を研究に当てはめれば、探究が依頼を書いた概念を改めることがあるので、ずれは探究が失敗したときだけでなく、成功したときにも開きうる。また、ラッセルの議論を研究に及ばせば、*S* の基準が *A* とともに動くところでは、両者が一致していることは、それだけでは、*A* とともに動かなかったなら *S* が求めたであろうものに *A* がなおお仕えていることを示さない（第3.2節）。人間の規範を継承していることも、認識的組織が異なることも、あるミスマッチが乖離なのか、誤りなのか、得なのかをあらかじめ決めない。それは、いまの *S*、これから変わりうる *S*、正しさが問われる場合には世界の、

いずれの立場から評価するかによる（第3節）。異なる種類の探究者による研究が S とアラインしない場合に何が起こるかは、どの関係が崩れるかによる。 S への関連性が崩れても、標的への応答可能性は S の同意によらない錨として残りうる。ただしそれは、コミットメントを訂正するものが標的によって定まるとする説のもとでのことであり、標的からの判定が S の手続きを通じてのみ探究者に届く場合には、その手続きが証拠によって改められうることも条件となる（第3.3節）。その場合にも S は、その探究者が到達したものを通じて、なお知識や理解を得られるかもしれない。そのために何が受け渡されなければならないか、そして得られるのが知識なのか、理解なのか、行為の能力なのかは、説によって異なる。機械がある結果を予測すると知ること（ q_h ）は、その結果が生じると知ることではなく、そこから標的についての知識へ進む道を、信頼性を中心とする説は予測の実際の信頼性に求め、ほかの説は理解、批判、保証などに求める。また、事例を提示でき、成功を、機械自身の選択との一致によってであつても採点できるところでは、人が機械ネイティブな種類の一部を、翻訳せずに学ぶ経路もある（第3.4節）。これらの可能性は、構築の計画に実質を与える。それは、誰が探究し、誰が知り、異なる仕方でも組織された探究者たちが互いからどう学ぶかを変えうる。どの可能性が実現できるかは、ここで検討した説と能力に左右される。

第V部が取り上げるのは、上の二つの選択のうち後者、すなわち探究が誰の認識的な営みを拡張するかを左右する関係である。第3節は、異なる種類の探究者が S の成員と異なりうるために、依頼も人間による制御も、研究が S の未知に向けられ続けることを保証しないこと、そうであるかどうかは個々の人工の構成についての問いであること、そして S のための研究の認識的な成功と、それを方向づけるものの正統性とは別の判断であり、後者は S の外で影響を受ける当事者に対しても問われることを論じた（第3.2節・第3.3節）。研究が S のどの未知を追求するのか、その結果が S にとって通用するものとなるのか、そして誰の権限によってその方向が定められるのかは、人工のものであれ人間のものであれ、単一の行為者を越えて広がる関係にかかわる。次の部は、それらのうちどれを人工の構成の内側に置くことができるのか、そしてそうしたとき何が変わるのかを問う。そのとき、研究が S にとって成功するかどうかと、その方向づけが影響を受ける当事者にとって正統であるかどうかとは、別々に問われる。

第Ⅴ部 研究を取り巻くどの関係が人工の構成の内側へ移りうるのか

第9章 Sのための研究を取り巻くのはどのような関係か

役割、価値、評価、統制

1. 研究の帰属が研究システムと出会うところ

研究する機械という、一人の研究者を一台の機械に置き換えることが思い浮かぶかもしれない。しかし本書は、研究を行う者 *A* を個人に限ってこなかった。*A* は人でも、集団や組織でも、人工またはハイブリッドの構成でもありうる（第7章第1節・第2節）。そして研究を遂行するシステムは、自らのもとで探究する者たちを選び、彼らの労力を配分し、その結果を批判し、彼らが従う仕事の規則を改めることもありうる。人間の研究では、これらはしばしば異なる当事者が行ってきた。たとえば、誰が探究に加わるかは大学や研究室が決め、労力をどこに向けるかは資金提供者や研究室の長が決め、結果の批判は査読者や追試を試みる他の研究者が行い、仕事の規則は分野の共同体などが定めてきた。人工のシステムは、研究を取り巻くこうした社会的な仕組みが果たしてきたことの一部または全部を、自らの内側で行うようになるかもしれない。第IV部までは、主に研究を行う者とその結果を問い、こうした仕組みには、それが知識や研究者としての地位にかかわるところで触れるにとどめてきた。しかし人間の研究では、資金をどの探究に振り向けるか、どの結果を認めるか、誰が研究の方向を決めるかのように、研究の行方を左右するものの多くが、個々の人間の研究者の外にあるこうした仕組みのなかで決まってきた。第V部は、これらを主題として取り上げ、研究を取り巻く関係が人工の構成に入ると何がかわるかを問う。この問いは第IV部の研究行為者の説明から続くが、行為者の能力を越えて、行為者の行う複数の探究がどう関係するかにまで及ぶ。すなわち、探究どうしの関係、それらが目指す達成がそれにとって新しいとされる社会との関係、そしてその帰結を負う者との関係である。これらの関係は、作業的特徴づけだけから演繹されるものではない。むしろ、特徴づけが研究の目標として名指すもの、すなわち *S* にとって新しい認識的達成が、研究における依存、複数性、有限の資源と出会うところで生じる。

この問いが生じる第一の理由は、特徴づけが名指す達成が、少なくとも知識の場合には、行為者だけの性質ではないことにある。その新しさは *S* に相対的であり、いくつかの説では、その根拠づけや保有にも他者がかわる。ここでは主に知識の場合で論じる。キャロリン・マクラウド（Carolyn McLeod）が信頼についての百科事典の項目でまとめるところでは、ジョン・ハードウィグ（John Hardwig）は、科学的知識はその獲得を他者の証言への信頼に依存すると論じた。知られていることを自分だけで確かめるのに要する時間、知力、経験を、一人でもつ者はいないからである（McLeod 2020, §3）。そうであれば、他者への依存は、より勤勉な探究者なら省けるような近道ではなく、知識が可能になる仕方の一部である。第11章第5節・第6節がハードウィグの議論に立ち返る。また第2章第6節は、いくつかの説では社会的な組織が知識の根拠づけや保有そのものに入ることを示し、第3章第8節は、局所的な根拠づけは、継続するシステムが多様な探索、独立した批判、記憶、訂正、さらなる問いの形成の能力を保っているかを決めないことを示した。したがって人間の研究では、*S* にとって新しい知識は、多くの場合、特定の探究を遂行する単位を越えて広がる関係を通じてもたらされてきた。そして、社会的な組織が根拠づけや保有に入るとする説では、そうした関係を通じて根拠づけられ、保有される。単位が多くの探究を遂行する人工の構成である場合にも、*S* にとって新しい知識がその単位を越えて広がる関係を通じてもたらされるかどうかは、第6節の仕分けが問うことの一つである。

第二の理由は、人工の研究者についての問題と研究システムについての問題とが、互いに還元できないことにある。両者は、作業的特徴づけが答えずに後の部へ負わせる問いとして第1章第9節が挙げた負債のうち、帰属の負債（研究を何に、どの境界において帰属させるか）と研究体制の負債（研究を取り巻く社会的な秩序についての問い）に当たる。人工の研究者についての説明と研究システムについての説明は、同じ配置にかかわりながら、異なる対象をもちうる。この対比は、個体と集団の対比ではない。人工の研究者は、それ自体が集団や組織でありうる（第7章第1節・第2節）。対比されているのは、個体であれ集団や組織であれ、ある研究者にできることと、研究者たちのあいだで探究がどう組織されるかである。ここでいう研究システムは後者、すなわち第1章第9節が研究体制と呼ぶ研究の社会的な秩序を指す。この秩序は、人間の制度のなかでも、人間と人工システムの混じった取り決めのなかでも、自らのインスタンスや研究者のあいだで探究を組織する一つの人工システムのなかでも実現されうる。後で挙げる例は最後の場合である。研究を遂行している一つの人工システムのなかでこの秩序が実現される場合には、同じ配置を、研究者、すなわち研究の遂行が帰属される構成（第7章）としても、研究システム、すなわちそのインスタンスや研究者のあいだの探究の組織としても記述できる。そのとき対比は、二種類の対象のあいだではなく、この二つの記述のあいだにある。第V部でいう人工の研究システムは、後者の仕方で記述された人工システム、すなわち、そのインスタンスや研究者のあいだの探究の組織として見た構成である。個々には優れた研究者の集団が、組織の悪い共同体を形成しうるし、よく組織された共同体が、個別に評価すればどれも高く評価されないような研究者から構成されうる。コナー・メイヨ＝ウィルソン（Conor Mayo-Wilson）、ケヴィン・ゾルマン（Kevin Zollman）、デイヴィッド・ダンクス（David Danks）は、証拠を共有する共同体の形式モデルのなかで、このことを一つの形で精確に示す。探究が進むにつれて利用できる最善の行為を選ぶ確率が1に近づく方法を合理的と数えたうえで、三人は、この意味で単独の探究者にとって収束する方法が、ともに用いられると収束しないことがありうること、そしてどの成員も単独では収束しないのに共同体は収束しうることを証明する。ただし、個人の合理性と共同体の合理性が完全に独立しているわけではないことも証明する。どのような共同体に置かれても収束する方法は、収束する共同体を形成するのである。もっとも、どの共同体に置かれても収束するというこの基準のもとでも、成員の方法が収束しないのに共同体が収束することはなおありうる。したがって、個人に勧められる方法と共同体に勧められる方法がどこまで食い違うかは、個人と集団の合理性についてどの基準を採用かに依存する（Mayo-Wilson, Zollman, and Danks 2011, 653–655, 663–669）。人工システムによる研究に当てはめれば、あるエージェントが単独で評価されたときの成績は、それだけでは、証拠を共有する他のエージェントのあいだでの成績を示さない。そして、そうした設定のもとでの評価であっても、研究者たちのあいだで探究をどう組織するのがよいかは決まらない。逆に、探究がよく組織されていても、それだけで、成員であれ集団であれ、ある単位が人工の研究者であることにはならない（第7章）。

第三の理由は、人工システムによる研究に固有である。冒頭で見たように、人間の研究では、探究が *S* のために方向づけられ、認定され、統治される関係は、研究者自身によってだけでなく、研究者以外の人々や組織体によっても実現されてきた。人工の構成は、少なくとも原理的には、これらの関係のいく

つかを内側に取り込みうる。たとえば、労力を分け合う研究者たちを自ら生成し、結果を批判する点検を自ら走らせ、全体の方向を決める決定を自ら保持することができる。

この可能性を具体的にするために、本書は一つの例を組み立てる。以下の章もこの例に立ち返る。単一のシステムが、異なる路線を追う百のインスタンスを生成し、別のインスタンスにその産出物を点検させ再現させ、その点検に応じて路線のあいだで計算資源を振り向け直すことがありうる。そのとき、そのどこにも外の同僚、査読者、資金提供者はいない。ただし、それを構築した者は、なおそれを止め、更新し、引退させることができるかもしれない（表9.1、第5節）。こうした配置の一部はすでにある。たとえば The AI Scientist は、研究を組織する工程のなかに査読を含む（序論第2節、第11章第7節）。そうしたシステムを構築し、あるいは展開する者は、それについて三種類の主張をなすうる。第一の主張は、それが探究をうまく配分しており、資源をそれに値する路線へ移している、というものである。第二の主張は、内部の点検を通った結果が S のために認定されており、そのため他者がそれに依拠する理由をもつ、というものである。第三の主張は、そのシステムが、自らのインスタンスが S のどの問いを追究するかを、他者に代わって決めてよい、というものである。この第三の主張は、配分がうまくどうかではなく、誰がその決定をしてよいかについての主張である。この例の役割は、探究を方向づけ、認定し、統治する三つの関係（第2.1節で配分、批判と認定、権威と名づける）が、一つの配置のなかで保持されることを示すことにある。第V部は、関係ごとに、対応する主張を裏づけるには何が要るかを問う。

第7章第2節は、ある機能を境界の向こう側へ移すことは、主張が誰について述べているかを変えるが、その機能を消しはしない、とした。この点はここでも変わらない。しかしそれは、移った機能が移る前と同じように果たされるのか、そしてどの機能もそもそも移りうるのかには、答えていない。この例でいえば、同じシステムのインスタンスが行う点検が外部の批判者の見つけるものを見つかるか、自らの点検に従って資源を振り向けるシステムが S のどの問いが追究に値するかまで定めうるかは、この種の問いである。したがって第V部の問いは次である。

研究が S のために方向づけられ、認定され、統治される関係のうち、どれを人工の構成の内側に置けるか。置くと何が変わるか。

問いの前半、すなわちどの関係をどの説のもとで内側に置けるかへの答えは、後半、すなわち置くと研究の何が変わるかを描くための手立てになる。この問いは、人工の構成が研究を遂行できるかどうか（それは第IV部が扱った）ではない。そしてこの問いに答えることは、作業的特徴づけに条件を加えない。ただしこの問いは、第1章第1節が価値とガバナンスの水準、すなわち重要性、有用性、正統性、リスクなどを、研究としての地位や達成の条件とは別に評価する水準に残した問いにかかわる。第V部は、異なる説が支持する答えを比較し、それぞれの説が挙げる条件が構成の内側で実現されると何が変わるかを論じる。こうした比較から結論を引くのに、あらゆる決定について価値や正統性の説を一つに決める必要はない。第1章第1節がその水準に残した問いのうち、どの未知が注意に値するかは、第3節と第10章第6節で取り上げる。誰の目的が探究を統べるかは、第10章第6節と第12章で取り上げる。認定において認識的価値と非認識的価値がどう相互作用するかは、第11章第4節で取り上げる。どのリスクが許されるかは、他者が負う誤りの費用を誰が定めてよいかという問いとして、第11章第4節と第12章第2節で取り上げる。そして、ある結果がいつ決定を変えることになるかは、結果に依拠して行為してよくなる閾値としてだけ取り上げる（第11章第4節）。第1章がその水準に挙げたもののうち、有用性、美、解放は、どの未知が追究されるにかにかかわるところでだけ入り、制度的認定は、功績の承認として配分を動かすところだけ入る（第10章第3節）。

本章の残りは、この問いの準備をする。第2節は、どの関係を指すのか、誰がそれを保持するのか、そしてそれらが異なる当事者のあいだに分けられていたことが何を供給したかかもしれないかを述べる。第3節から第5節は、三つの関係のどれか一つではなく、そのすべてが含むもの、すなわち価値づけること、評価すること、統制することを、以下の章が用いる最小限の形で述べる。第6節は、第V部がどう進むかを述べる。この準備の途中で、成果の比較だけでは捉えられない関係も見えてくる。人工の構成が同じ知識を同じ S に届けても、研究者の形成を重んじる立場などでは、研究の組織が同じ価値を保つとは限らない。誰がその過程に加わったか、また権威の説によっては誰が決めたかが、なお問われるからである（第3節）。

2. 三つの関係と、それを保持する当事者

第1節は、研究が S のために方向づけられ、認定され、統治される関係について語った。本節は、それがどの関係なのか、なぜ第V部はそれを三つと数えるのか、そして誰がそれを保持するのかを問う。ここでも第V部全体でも、行為者や探究者は、研究の遂行が帰属される単位を指す。それは人、集団、あるいは人工またはハイブリッドの構成でありうる（第7章第1節・第2節）。本節は各関係を二つのものから導く。一つは、活動が S に結びつく箇所を示す作業的特徴づけの語である。もう一つは、特徴づけの外にある条件であり、その結びつきが、なぜ特定の探究を遂行する単位を越えて広がる関係になるのかを説明する（第1節）。人間の研究では、こうした関係は、その単位自身だけでなく、それ以外の当事者によっても保持されてきた（第1節）。多くの探究を遂行する人工の構成が、それらを自らの内側に保持できるかどうか、第V部の問いである。

2.1 三つの関係の導出

第一の語は S にとって新しいである。第1章第7節は、この語と「試み」とを合わせて、研究が問いの構造をもつこと、すなわち研究の試みが、 S とある対象との関係のうちでまだ決着していない点に向かうことを導いた。第1章はまた、社会は無際限に多くのことについて無知であり、新しさはそのうちどれが注意に値するかを決めないと言った（第1章第8節）。この語が与えるのは、選択がなされる空間であって、選択する必要ではない。必要は特徴づけの外の条件から来る。探究の手段、たとえば注意、労力、時間、装置、資金などは有限で、ほかの用途ももつので、すべての未知を一度に追究することはできない。第4章第5節と第5章第4.2節は、何を問いとして取り上げ、何を追究するかを選択が一つの探究のなかで繰り返さなければならないことを論じ、第7章第4.3節は、問いを選ぶ能力が行為者の自律性の一つの次元であることを論じた。この選択を単位を越える関係にするのは、もう一つの外の条件、すなわち多くの研究者が S の未知を追究していることである。そうした研究者たちのあいだでは、同じ選択が別の形をとる。見込みの分からない探究の路線のあいだで労力が分けられ、その分け方は各研究者の選択の総和ではない。この関係を配分と呼ぶ。 S の未知のあいだでの、注意、労力、資源の分布である。それがなぜ生じ、何がそれを難しくするのかは、第10章第1節で取り上げる。

第二の語は *認識的達成* であり、ここではまず *知識* を通じて展開する。第2章は、結果の産出と、それが知識になることを分けた。そのうえで次のことを見た。批判的文脈説では、批判の場、批判の受容、公的な基準、知的権威の調整された平等が、客観性を単に伝えるのではなく、客観性が成り立つことに寄与する。また、分散説・集団機能説（適切に組織された集合体そのものが知る者を実現しようとする説）では、検索、使用、修正が保有に入りうる。知識を個々の知る者に置く説では、この語そのものは知る者を越えるものを指さない。そうした説で知る者を越えるものを指すのは、新しさが *S* に相対的に定まること（*S* への指標づけ）と、特徴づけの外の条件である。この条件は第1節がハードウィグとともに触れたものに関連し、ここではそれを *S* の側から述べる。つまり、*S* の成員はそれぞれ、多くの結果について証拠を集めることができない。そのため彼らは、結果に依拠するとき、他者を通して依拠することになる。結果を産出した者を通してであり、産出者の外で結果が吟味を受けていれば、その吟味を通してでもある。人工の *S* については、この条件がどこまで成り立つかは、証拠が成員のあいだでどのように共有されるかに依存する。たとえば、ある成員が集めたものが、他の成員にとって自らの記録として利用できるのか、それともその成員の報告を通じてしか利用できないのか、である。結果がこうして他者を通して依拠されるようになる関係を、**批判と認定**と呼ぶ。候補となる結果が誤りを見つける吟味にさらされ、他者がそれに依拠する理由を与えられる関係である。そうした依拠がどこまで結果への吟味に基づいている必要があるかは、第11章第5節で比べる証言の説が分かれる問いの一つである。認定には、ある結果が実際に *S* にとって新しいという判断も含まれる。第5章第5節は、*S* にとって何が新しいかを特定すること自体が探究のなかの仕事になりうることに触れた。そして、そうした仕事がかつていかにそうであるように、この判断は誤りうる。第11章第1節は、この関係の二つの部分、すなわち誤りを見つける吟味にさらされることと、他者に依拠する理由を与えることを分ける。批判の価値をそれが検出する誤りに置く説では、第一の部分は単一の行為者の内側でも成り立ちうるが、第二の部分は *S* にかかわる。批判的文脈説では、点検が何に反する証拠となるかは、それ自体、共同体が保持する仮定に相対的なので、その共同体が *S* であるところでは、第一の部分も *S* にかかわりうる。

ただし、特徴づけが名指すのは知識に限らず、認識的達成一般である。以上の知識についての議論は重要な一事例を特定するが、すべての達成が同じ社会的条件をもつことまでは確立しない。理解の場合、認定が扱うのは、説明を受け取る者に理解をもたらすことではなく、説明が適切であることや、誰かが理解しているという帰属でありうる。説明の適切さや誰かの理解についての主張にも、受け取る者がそれぞれ自ら点検せずに依拠することがあり、そのため批判と認定がここでもかかわる。認定が何を確立するかは、どの主張を評価するかによる。第11章第5.6節・第6節・第9節は、この拡張を、いま検討した知識の説明の含意としてではなく、第3章の区別を用いた追加の論証として展開する。

第三の語は *S* のものである。第8章は、*S* のための研究の認識的な成功と正統な統治とを区別し、*S* を特定しても誰が *S* を代弁してよいのかは特定されないことを見た。ここでの外の条件は二つある。一つは、*S* が多くの成員からなり、一つの声をもたないことである。ただし、行為者であると同時に *S* の全体でもある単一の構成という極限（第12章第5節が検討する）では、この限りではない。もう一つは、探究の帰結が *S* に属さない当事者にも及びることである。どの未知を追究し、どの誤りの費用を受け入れるかは、誰かの決定によって定まることもあれば、第10章第5節が配分の事例から示唆するように、誰も方針として定めない過程によって定まることもある。それが決められるところでは、問いは、誰がそれを決めてよく、誰がその決定を改めてよいかである。決められないところでは、問いは、誰がその設定に責任を負い、誰がそれを変えてよいかである。この関係を**権威**と呼ぶ。他の二つの関係の取り決めと、第12章第2節が述べる探究についてのさらに別の決定とを、決め、改める資格、すなわち次に述べる意味での当事者資格である。

ここでも第V部全体でも、当事者資格（standing）は規範的な意味でだけ用いる。それは、決定に服する者との関係で、提案し、決定し、異議を唱え、拒み、あるいは改める資格である。それは有能さでも、決定する権利の形式上の割り当てでも、その権利を行使する実効的な能力でもない（第5節）。どの当事者が当事者資格をもち、何がそれを根拠づけるかは、採る説に依存する。有能さがその根拠になりうるかどうかについては正統な権威の説が分かれており、第12章第1節がそれを取り上げる。福祉や行為者性のような存在の特徴に基づくのか、力や協働のような存在どうしの関係に基づくのかについては正義の説が分かれており、第12章第5節がそれを取り上げる。英語の standing や authority の語は、ほかに三つの事柄を指すこともあるので、それらには別の語を当てる。一つは、ロンジーノ（Longino）の調整された平等の条件がかかわる、ロンジーノが知的権威と呼ぶもの、すなわち共同体の批判的な相互作用に寄与する能力である。調整された平等の条件は、どの視点もこの能力を一応はもつとする（Longino 2025, §5）。第V部はこの語をそのまま使い、第11章第3節がそれを取り上げる。二つ目は、行為者の出力が *S* にとって何として数えられるかであり、以下ではこれを *S* にとっての**地位**と呼ぶ。三つ目は、人工の存在者が、存在者としてそれ自身の道徳的な請求をもつかどうかであり、これは第12章第5節が問う *道徳的地位* の問いである。

この導出は四つの限定を伴う。第一に、語と関係の対応は一对一ではない。*S* にとっての新しいさは配分だけでなく認定にもかかわり、権威が責任を負う相手は、*S* を越えて、それが下す決定の帰結を負う者にまで及びうる。そのうち誰に対して責任を負うのかは、採る正統性の説による（第12章第1節）。第二に、関係は同格ではない。配分と認定は一階の関係であり、権威は、他の二つの関係の取り決めを定め改める点で二階の関係である。権威が覆うさらに別の決定は第12章第2節が述べるとおりであり、運用とリスクに関する決定、誰が成員、批判者、専門家と数えられるかについての構成的な決定と承認に関する決定（constitutive and recognition decisions）〔訳注：ここでの「構成的」は constitutive の訳であり、人工の構成（configuration）とは別の語である〕、そして記憶と訂正に関する決定などを含む。第三に、三つのうち、結果が *S* にとって知識としての地位をもつかどうか、また *S* の知識となるかどうかにかつて最も直接に入りうるのは認定である。前者は社会的な組織が根拠づけに入るとする第2章第6節の説のもとで、後者は社会的な組織が保有に入るとする説のもとでそうである。集団の知識を、該当する領域で意思決定の権限をもつ成員に依存させる説では、権威の一形態も入る。ある関係が主張される達成の条件に入らないところでは、その評価は価値とガバナンスの水準にとどまる（第1章第9節）。知識について言えば、ここで見たとの説のもとでは、これにあたるのは配分であり、さらに、集団の知識を権限をもつ成員に依存させる説を除けば権威が、社会的な組織が根拠づけにも保有にも入らない説では認定が、これにあたる。他の達成について、関係がその成功条件に入るかどうかは、それぞれの説、たとえば第3章第6節の公共的・集合的な理解についてのものである。理解を知識の一種とみる説では、いま述べた知識についての条件がそのまま及びうる（第3章第3節）。第四に、三つのどれも研究としての地位の条件ではない。作業的特徴づけの意味では、*S* にとって新しい認識的達成へ向けて組織された活動は、それを始めさせた配分が賢明だったかどうかにかかわらず、またその方向の正統性がどうであれ、研究である（第1章第1節・第9節）。それでも第V部がこれらの関係を取り上げるのは、第1節の三つの主張が例示する種類の、人工の研究システムについての主張が、これらの関係についての主張だからである。

三つへの分け方は、研究を取り巻く関係のうち、構成の内か外かという位置がそうした主張に直接にかかわるものを取り出すためのものである。第1章第9節は、研究体制を、研究の分業、探究の配分、批判と認定、記憶と訂正、功績、権威からなるものとし、人工研究者が変えうるものとしてアジェンダ設定

や注意も挙げていた。第V部は、このうち三つの関係そのものでないものを、独自の関係としてではなく、三つの関係にかかわるところで扱う。分業と注意は、上で定義した配分に含まれる（第10章）。アジェンダ設定は、*S* のどの未知を追究するかを選択として配分のもとで、またその選択を誰がしてよいかという問いとして権威のもとで扱う（第10章第6節、第12章）。記憶と訂正は、時間にわたる認定の条件として（第11章第8節。記憶はそこでは記録という形で扱われる）、また権威のもとでの決定として（第12章第2節・第6節）取り上げる。功績は、帰属の一つとして第7章第3.3節で扱われた。第V部は、それが配分を動かすところを取り上げる（第10章第3節）。ほかに三つの関係にかかわる事柄がある。装置とインフラストラクチャーは、第5章第6節と第7章第4.2節で可能にする条件として検討された。第V部はそれらが三つの関係を実現するところを取り上げ、第12章第6.4節はインフラストラクチャーを、それを通じて実効的な権威が分け持たれる関係として取り上げる。物質的な資源や研究者の形成は、次に「支え」としてまとめて取り上げる。

ここでいう支えとは、物質的な資源、研究者の形成、失敗の費用を負うことなど、失敗しうる組織された方向を持続させることを可能にするものである。支えは主として、特徴づけのもう一つの語である *試み* にかかわる。活動の水準では、資源、許可、アクセスを得ることは、それが探究の問いを解決しようとする試みのなかで引き受けられるとき研究に属する（第7章第4.4節）。どこまでが支えで、どこからが研究の一部かは、行為者の境界をどこに引くかによって変わる。境界を広く引けば、手段を得て失敗の費用を配分することも行為者自身の活動になり、支える者の役割は行為者が自ら保持することになる。価値とガバナンスの水準では、支えを求める理由は、それを供給する者に向けて示される。支える者と *S* は別の役割であり、同じ当事者が両方を占めることも、そうでないこともある。企業の内部研究では、企業が支える者であり、同時に結果がそれにとって新しい *S* でもありうる。これに対して、パトロンが学者を養っていた場合には、支える者であるパトロンと、その学者の結果がそれにとって新しい学者の共同体とは、一致していたとは限らない。公的な資金提供機関と、それが資金を出す研究がそれにとって新しい専門共同体や人類も、必ずしも一致しない。第V部は支えを第四の関係とは数えない。上の導出からは支えは出てこない。支えが主としてかわる「試み」の語は、活動の向きと失敗の可能性を示す語であって、活動が *S* に結びつく箇所を示す語ではないからである。「試み」がそうした結びつきに入るのは、問いの構造のように「*S* にとって新しい」の語とともにであり、そこから生じる関係は配分である。ただし、本節の導出は研究を取り巻く関係を挙げ尽くすものではないので、それだけでは、支えを独自の関係として扱うかどうかは決まらない。第V部がそう扱わないのは、第1節の三つの主張が例示する種類の、人工の研究システムについての主張に支えがかかわるのが、主として三つの関係のどれかを通してだからである。資金が *S* のどの未知が追究されるかを動かすとき、支えは配分の一部になる。結果を判断する者が資金も出すとき、支えは認定と束ねられる。資金を握ることが探究の方向を定め改める力になるとき、支えは権威と束ねられる。そこで第V部は支えを、それが三つの関係を実現し、あるいはそれらと束ねられているところを取り上げる。こうした形に収まらない問い、すなわち支えがそもそもなぜ与えられるのかは、第3節が立場ごとに述べる。

2.2 研究を取り巻く役割

三つの関係が含む役割は、ここでまとめて名指しておく。第8章第1節は、評価者、受益者、支える者、聴衆、権威をもつ者が行為者とも *S* とも異なりうることに触れ、それらが重要になるところで局所的に特定した。第V部はそのいくつかを全体を通じて用いるので、表9.1 はそれらを一か所に示す。役割を分けておくのは、「社会」という一語が指しうる、参照社会、研究を支える者、国家の公衆などを区別しておくためである。役割は、ある当事者が探究に対してどうかかわるかによって分けられる。つまり、探究を遂行すること、探究の結果がそれにとって新しいとされる集合体であること、探究を持続させること、探究を判断すること、その結果に依拠すること、その帰結を負うこと、そして探究を行う行為者を作りまた作り直すことである。このように分けるのは、配分、認定、権威についての主張が持ち出すのが、こうしたかかわり方だからである。いくつかの役割は同じ当事者が保持しうるし、第1節の例では、人工の構成がそれらの複数を保持しうる。

表9.1 第V部で用いる、研究を取り巻く役割

役割	誰が保持するか	第V部が主に用いる箇所
A、行為者	研究の遂行が帰属される単位。人、集団、あるいは人工またはハイブリッドの構成（第7章）	全体
<i>S</i>	意図する達成がそれに相対的に新しく、そのために求められる社会。担い手や実現の仕方は別の問いである（第1章第2節・第6節）	全体
支える者	試みの手段を供給し、その失敗の費用を負うか配分する者	第2節・第3節、第12章第1節
評価者と受け手	判断する者と、その判断を用いて選ぶ者（第4節）	第10章・第11章
依拠する者	認定された結果を信念や行為の基礎とする者。ふつうは <i>S</i> の成員だが、つねにそうとは限らない	第11章
帰結を負う者	探究とその誤りの費用とリスクが降りかかる者。 <i>S</i> に属する必要はない	第11章・第12章
提供者または維持者	遂行が帰属される構成を構築し、維持し、更新し、あるいは引退させる者。したがって、行為者の手段だけでなく行為者そのものを変えうる（第7章第3.1節）	第10章第5節、第11章第5.4節、第12章第2.2節・第6.4節

さらに、探究についての決定がそれに対して正当化を求められる相手という役割もある。その範囲は採る正統性の説に依存するので、第12章第1節で取り上げる。*S* は全体を通じて抽象的である。それは人類、専門共同体、組織、混成の集団、あるいは候補となる人工の社会でありうる（第1章第6節）。第V部は、そう断らない限り、「社会」を国家の公衆の意味では用いない。提供者を支える者と別に挙げるのは、試みの手段を供給することと、行為者を作りまた作り直すことが、探究への異なるかかわり方だからである。第6節は、研究を遂行する構成そのものも供給する提供者を、人工の構成がもたらしうる、固有の要件を伴う特徴として取り上げる。

2.3 実現、分割、再結合

三つの関係は分析上の関係であって、制度ではない。本書の読者が知る研究システムでは、なじみのある制度がそれらの関係で中心的な役割を果たしてきたが、それらの制度は、関係のありうる実現のうちの一つである。配分は、たとえば、功績の承認とキャリアの圧力のもとでの研究者による問題の選択、

資金提供者による課題の選抜、研究室の指揮などを通じて行われてきた。第10章がそれを検討する。批判と認定は、たとえば、出版前の査読、追試、結果の公開、そしてその後の利用などを通じて行われてきた。第11章がそれらを検討し、その第7節は、学術誌が束ねるようになった、分けて割り当て直すことのできる学術的コミュニケーションの働きを取り上げる。権威は、たとえば、資金提供者、自らを統治する研究共同体、規制者、公衆などによって行使されてきた。第12章がそれを検討する。各関係がどう実現されるかは、分野によっても異なる。第10章第5節がラウレンス・ヘッセルス (Laurens Hessels) らの研究にもとづいて報告するとおり、ヘッセルスらが調べた数学のグループでは、議論が正式な公刊の前に承認を得ることがあり、人文学のグループでは、承認は助成金よりも主にポスト (職) と研究時間に変換されていた。第11章第2節が論じる、カーネル (証明支援系の中核にあつて推論の各段階を検査する、小さな信頼されたプログラム) に照らして検査された証明は、追試を含まない認定の実現である。そして分野の基準に対する権威は、資金提供者や規制者よりもその分野自身の成員にあることがある。三つの関係は、これらの実現も覆うことを意図している。ここで名指した制度は、主に自然科学から引いた例である。これらの制度を、歴史のなかで形づくられた、ありうる実現の一つとみる見方には系譜がある。キッチャー (Kitcher) は、1660年代に研究の選択や新しい考えの認定のために最初に定められた規則が、まったく異なる文脈に合わせたものだったことを指摘し、後に自らがもつ影響を思いもよらないまま「成り行きで進化してきた」制度が、いま果たしている働きに見合うように設計されているかを問う (Kitcher 2013, 216)。マイケル・ニールセン (Michael Nielsen) とカンジュン・チウ (Kanjun Qiu) は、科学の社会的過程の設計は歴史の偶然に左右されてきたものであり、ありうる過程の空間は広大でほとんど探索されていないと述べる。二人はその探索に少なくとも数十年か数世紀を要するとし、科学の社会的過程のうち集合的に保持されるものがほとんど変わらない理由として、少数の大きな資金配分者と研究機関への統制の集中、誰も一人では規範を変えられない集合行為の問題、共同体の判断と、思い描かれる将来の評価の影とによって社会的過程を均質化するネットワーク効果、そして新しい制度を育てる循環がないことを挙げる。そして、科学者は発見の仕組みの利用者であつて大部分は設計者ではないとして、新しい過程を試し広める者にメタサイエンスの起業家を構想する (Nielsen and Qiu 2022)。その探索を、研究を行う人工の構成そのものが自らの研究について行いうるとき何が起こりうるかには、第12章第9節で戻る。そして本書の著者が AI エージェントとともに書いたエッセイは、AI が、科学をなす役割と関係を分解し、別様に組み立て直すための素材になりうると提起した (Takagi and an AI agent 2026)。

ここで問題になるのは、異なる当事者が保持する関係の構造である。ここで提案する分析によれば、これらの関係が異なる当事者のあいだに分けられていたことは、どの単一の当事者もそれを供給するよう設計されていなかった財を供給しえた。第V部はそのうち四つを検討する。四つのうち第一は**境界をまたぐ誤りの検出**である。ある場所で生じた誤りを別の場所で検出する点である。第二は**異論の当事者資格**である。これは第2.1節で定義した意味での当事者資格であり、問いを決着させ、あるいは路線を閉じるであろう判断を下す者やそれに拘束される者との関係で保持される、その判断に異議を唱え、自らの見解や路線の放棄を拒み、あるいは閉じられたものの再開を提案する資格である。その資格は、異論が後に正しかったと分かるかどうかにかかわらない。関係の分割が供給しえたのは、そうした資格をもちうるほど別個の当事者と、その資格が認められ行使される機会である。取り決めが資格そのものを供給しうるかどうかは、採る当事者資格の根拠による (第12章第5節・第9節)。そうした当事者資格が認識上守ったもの、すなわち優勢な判断に好まれない見解や路線が、なお支持者をもって残り続けることは、当事者資格そのものとは別の事柄であり、両者は分かれうる。前者を、異論の認識上の働きと呼ぶ。批判においては、ロンジーノの知的権威の調整された平等が、その働きが何を求めるかについての一つの説である (Longino 2025, §5)。第三は**説明が求められた境界**である。そこでは、ある当事者が、自らが行ったことを別の当事者に説明し、あるいは正当化しなければならなかった。研究では、結果が S にとって新しく、照合できる別の情報源がないことが多いので、結果がどう得られたかの説明は、依拠する者が結果を吟味する数少ない手がかりになりえた (第4節)。第四は**責任を負う当事者**である。行われたことについて責任を問われうる当事者である。研究では誤りがしばしば後になってはじめて分かるので、この財は、結果が出たときだけでなく、訂正が求められる後の時点にも、責任を問われうる当事者がいることを含みえた。ただし、境界をまたいで起こることについて責任を問われる当事者がいなければ、分離は責任を分断することもありうる。第10章第5節は、誰かが決めたわけではないのに人間の配分が果たしていた働きを、この四つのほかにも挙げている。四つは、関係が構成の内側に移るときに行方を追える具体的な財を、第V部に与えるためのものである。関係の分割がこの四つを供給したということは、第V部が検討する仮説であつて、確かめられた所見ではない。そして四つのそれぞれに、それを検証する章が割り当てられている。境界をまたぐ誤りの検出は第11章第2節・第3節で検討する。そこでは二つの天文学のエピソードが、それが一方の事例では供給され、他方では欠けていたことを示す。異論の当事者資格は、異論が路線を続けることや再開することの形をとる第10章第4節と、批判者についての第11章第3節で検討する。説明が求められた境界は第11章第9節と第12章第6節で、責任を負う当事者は第12章第6節で検討する。これらの各章は、末尾で自らに割り当てられた財に立ち返り、第12章第9節が四つすべてを集める。

先に引いたエッセイの定式化を三つの関係の語で言い直せば、人工の構成は、別々の当事者が保持する関係を結合しうる。一つのシステムが、文献を探し、候補を生成し、実験を選び、結果を評価し、資源を振り向け、成果を利用することができる。支援、判断、指揮を合わせることに、遂行者を内側に取り込むことは、まだ同じではない。支える者は、別個の遂行者に依存しながら、結果を判断し、仕事を指揮することもできる。遂行者が拒み、自らの証拠を手元にとどめ、あるいは去ることができるなら、遂行者は、その協力が必要とされる当事者として残る。自らの研究者を生成する構成は、遂行者をも内側に取り込みうる。これが第V部の主題である。つまり、人間の研究では別々の当事者が保持していた関係が、研究の遂行者も含めて、一つの構成に入りうる。再結合は、かつての境界で生じていた遅延と情報の損失を取り除きうる。しかし同時に、境界が付随的に供給していたものも取り除きうる。したがって第1節の問いは、より鋭く立てられる。これらの関係の分割が供給したかもしれない財のうち、どれが人間という実現者についての前提、すなわち、限られた注意と寿命をもちつつ実際に関係を保持していた人々についての前提に依っていたのか。どれが、キャリア、学術誌、データの保管のように、関係が歴史のなかでどう組織されてきたかについての前提に依っていたのか (先に系譜を記した見方では、そうした組織は実現者が人間のままでも組み直されうる)。どれが、何がそれを実現するかにかかわらず成り立つ探究の条件に依っていたのか。そしてどれが、構成の外の当事者 (S が構成の外にあるときは S を含む) との関係に依っていたのか。第6節は、この仕分けを第V部の手立てとして述べる。

関係が構成の内側に入るということは、それが上から設計され割り当てられるということと同じではない。内側に入った関係は、構成が定めた規則を上から当てはめることで実現されることもあれば、構成のなかの多くの機械の実現者が、それぞれの動機やそれに相当するものに従ってふるまうことで、誰も定めないまま下から実現されることもある。そして、どの実現者を經由させるか、經由させたものをどう統合するかによって、同じ関係でも異なるものが実現されうる。人間という実現者を経ずに上から設計できるようになることは、第V部が追う唯一の含意ではないが、重要な含意の一つである。人間の研究では、好まれない路線を生かし続けること、参加者の交代、好奇心による探索の分配などは、多くの場合、研究者それぞれの動機や寿命を經由して、誰

もそれを定めないまま生じてきた（第10章第5節）。これに対して、上から設計できる構成では、成員のあいだの違いと誤りの相関も、どの路線に何を割くかも、作り方によって決まりうる（第6節）。そのため、かつて誰も決めずに生じていたものは、設計で定められることも、機械の実現者のあいだで下から生じることも、そのどちらでもなく失われることもありうる。好奇心による探索を配分の方法として採るかどうかも、暗黙のままにはとどまりにくくなるかもしれない（第10章第5節）。しかも研究では、どの違いが後に役立つかは前もって分からない（第10章第2節）。したがって、多様性、すなわち成員や路線のあいだの違いを設計の値として定めることは、それが何に役立つかを知らないまま定めることになりうる。さらに、その値は一度定めて終わるとは限らず、構成が探究の結果に照らして改めていくものでもありうる。そのとき研究の組織の仕方は、構成自身の探究の対象にもなる（第12章第9節）。

逆向きの動きも提案されている。分散型科学（decentralized science）を唱える一人であるサラ・ハンバーグ（Sarah Hamburg）は、査読や評判のシステムなどの機能は「別々の専門化した共同体」が統治しうると論じ、そうすれば「単一のプラットフォームによる支配」の危険が和らぐとする（Hamburg 2022）。分散化を理念として検討するネイサン・シュナイダー（Nathan Schneider）は、そうした分割が中心を取り除くのではなく移すことがあると論じる。シュナイダーは、政治学、インターネットをめぐる言説、ブロックチェーン技術の唱道者のあいだでのこの語の用法を概観し、システムのある部分での分散化は、別の部分での集中化と一貫して重なってきたと見る。そして、分散化が唱えられる多くの場合に「中心は去ってはいなかった……移っただけである」と書き、「分散化の理論は集中化の理論も含まなければならない」と述べる（Schneider 2019, 26, 28）。多中心的なガバナンスの研究も、形式と働きのあいだに関連する線を引いていた。エリノア・オストロム（Elinor Ostrom）が引くヴィンセント・オストロム（Vincent Ostrom）、チャールズ・ティボー（Charles Tiebout）、ロバート・ウォレン（Robert Warren）によれば、多中心的とは「互いに形式上独立した」多くの意思決定の中心があることを意味し、それらが実際に独立に働くかどうかは「個々の場合における経験的な問い」である（Ostrom 2009, 411）。第5節が挙げる独立の権能、すなわち自らの目的を変える、資源を拒む、証拠を別に保持する、記録を携えて離脱するなど、ある部分が他と分かれて進むことを可能にする権能とあわせて読むと、これらの議論は、さらなる分割が上の財のうち何を供給するかは、役割をもつ当事者の数よりも、当事者が分かれて進むことを可能にする権能が実際に彼らのあいだで分けられているかどうかによるのかもしれない、ということを示唆する。研究では、この違いが大きな意味をもちうる。どの路線が成功するかを前もって知る者はいないので（第10章第2節）、優勢な判断に好まれない路線が残るかどうかは、それを追う部分が資源を拒み、証拠を別に保持し、記録を携えて離脱できるかにかかりうるからである（第10章第4節）。

2.4 関係の運動と、 S にとっての地位

第V部の問いへの答えを左右する点が、さらに二つある。第一に、関係は互いに運動している。配分は、部分的には、評価が生み出す信号を手がかりに向きを決める（第4節、第10章第3節）。批判によって追究されるものや依拠されるものが変わるのは、誰かが批判に基づいて行為するところであり、誰がそうしてよいかは権威の問いである。そして権威は、配分と批判が何のためにあるかを定める。したがって各関係についての章は、その関係がどこで他の関係に依存しているかも示す。第二に、ある関係を構成の内側に置けるかを問うと、第7章第4.2節が区別した依存の種類のほかに、もう一つの種類が問題になる。同節は三つを分けた。一つは可能化的な依存であり、能力は存在し行使されるが、外部の当事者は行為者に帰属される組織化の一部を供給しない。二つ目は指示的または代替的な依存であり、他の当事者が行為者に帰属されるものの一部を供給する。三つ目はガバナンス上の制約である。第V部の関係は、さらに別の種類の依存を含みうる。行為者の出力が S にとって何として数えられるか、すなわちその S にとっての地位を部分的に成り立たせる依存である。この地位は、それだけでは S がその出力を知るようになったことではない。両者がどう関係するかは、第2章第6節の保有の説のあいだで異なる。ある出力が S にとって知識として数えられるか、 S が依拠してよい結果として数えられるか、 S の探究の正統な方向として数えられるかは、産出する行為者の外にある関係に依存しうる。そのことは、行為者の外の何ものもその産出を可能にせず、指示せず、制約していないときにも成り立ちうる。そうであるかどうか、そしていつそうであるかは、採用する知識と権威の説に依存し、以下の章がその分岐をたどる。

3. S にとって価値がある

ここで問題にする価値は、第1章第1節の価値とガバナンスの水準に属する。研究の結果と研究そのものが S にとってもつ価値、すなわち、どの結果が重要であり、誰にとって重要かである。作業的特徴づけは、どの未知が注意に値するかを決めない（第1章第1節）。他方で、人工の研究システムについての主張、たとえば第1節の、それが探究をうまく配分している、その認定に依拠してよい、あるいは他者に代わって決めてよい、という主張は、この価値についての何らかの見方を前提する。本節は、そうした主張が前提しうる見方を述べる。これは、第11章第4節が取り上げる問い、すなわち、ある結果が依拠してよだけ支持されているという判断に価値が入るかどうか、どう入るかという問いとは別の問いである。そこでは、価値は結果の受け入れにかかわる。本節の区別がその問いにかかわるのは、主に、誤りの重みを誰の価値づけが定めるかを問うところである。

第2節の三つの関係はそれぞれ、活動の水準だけでは決まらない価値の問いを含む。配分は、 S の未知のうちどれが、それに要する労力に値するかを問う。認定は、ある結果に依拠してよい時を判断するにあたって、どの誤りが他の誤りより重いかを問う。権威は、これらの事柄についての誰の順序づけが続べるのか、そして誰がそれを改めてよいのかを問う。本節は、何かが S にとって価値があると言うことが何を意味しうるかを述べ、本節で立てる区別を、以下の章が繰り返さずに用いることができるようにする。 S の未知の順序づけについては第10章第1節・第6節が、誤りの重みづけについては第11章第4節が、目的の正しい順序づけと、それを課す当事者資格との違いについては第12章第1節・第3節が、それを用いる。

「 S にとって価値がある」という言い回しは、二つの点で異なることを意味しうる。何に価値があるのか、すなわち価値の対象と、その価値がどの意味で S に相対的なのかである。第一の点、すなわち価値の対象は、たとえば次のようなものでありうる。理解のような認識的達成（理解は、その構成部分の価値を越えて価値があると言われる。第3章第7.1節）、問い（ S の未知のいくつかが他より重要だと言われるとき）、結果の形（真なる結果でも、その内容を理解したり適用したりできない成員にはほとんど役に立たないとき）、研究の実践全体（ S や支える者がそもそも探究を持続させるどんな理由をもつのかを問うとき）、分布（同じ結果が S のある成員には資するが他の成員には資さないとき）などである。第V部が主にかかわるのは、問い、実践、分布である。

第二の点、すなわち価値が S に相対的であるその意味としては、たとえば次の四つがありうる。一つ目は、それが S の成員にとって善いという意味である。マーク・シュローダー (Mark Schroeder) によれば、リチャード・クラウト (Richard Kraut) とクリスティン・コースガード (Christine Korsgaard) は、誰かにとって善いことなしには何も善くありえないとしており、この見方がそれにあたる (Schroeder 2025, §1.1.1)。二つ目は、それが S のもつ目的に資するという意味である。ポール・ジフ (Paul Ziff) やステイーヴン・フィンレイ (Stephen Finlay) のものなど、善さについてのどの主張も目的や意図に相対的だとする目的関係的な理論がそれにあたり (Schroeder 2025, §1.1.3)、マシュー・J・ブラウン (Matthew J. Brown) がまとめるところでは、キッチャーが実践的な重要性の出どころを実践上の目標に求めるときもそうである (M. J. Brown 2010, 134, 137)。三つ目は、 S の成員が実際にそれを価値あるものとするという意味である。キッチャーが、認識的重要性の出どころを、人間の自然な好奇心が向かう大きな問いに求めるとき、それはこの意味に近い。ただし、ブラウンが伝えるように、キッチャーはその好奇心について「健全な好奇心」という私たちの通念に訴えており (M. J. Brown 2010, 137–138)、その点でこの説には規範的な要素も含まれる。四つ目は、 S の成員のあいだの手続きや理想化された熟議がそれを追求することを選ぶだろうという意味である。キッチャーの秩序ある科学の理想は、研究のアジェンダについてこの意味に近い。ロンジーノの整理では、この理想のもとで、どの研究を追求するかについて実際になされる決定は、適切に構成された代表の集まりが、専門家の提供する関連情報の助けを借りて熟議したならば下すであろう決定に沿う (Longino 2025, §6、第12章第1節)。一つ目と二つ目はシュローダーの価値論の概観が立てる区別に沿い、三つ目と四つ目はキッチャーの重要性の説と秩序ある科学の説から取っている。四つを並べるのは、 S の未知についての議論がどの意味に拠るのかを、以下の章が言えるようにするためである。価値の対象と、価値が S に相対的である意味という二つの点を本書が分けるのは、何に価値があるのかについての議論と、その価値がどの意味で S に相対的なのかについての議論とは、一方が成り立つても他方が成り立つとは限らないからである。たとえば、ある問いが他の問いより重要だという議論が説得的でも、それが S の成員にとって善いという意味で重要なのか、成員のあいだの熟議がそれを追求することを選ぶだろうという意味で重要なのかは、それだけでは決まらない。分けておけば、一方についての議論が他方についての議論と取り違えられずにすむ。

S が共同体であるとき、一つ目の意味、すなわちそれが成員にとって善いという意味は、成員一人ひとりの善を合計したもの、あるいは分配の原理によって重みづけたものにかかわるとも、共通善にかかわるとも読める。ワヒード・フセイン (Waheed Hussain) とマーガレット・コーン (Margaret Kohn) は、共通善を、分配上の重みづけの有無にかかわらず集計された福祉から、成員のあいだの関係の要求に本質的に言及する点で区別する (Hussain and Kohn 2024, §1)。二つ目の意味も同様に、成員がそれぞれにもつ目的にかかわるとも、成員が共通にもつ目的にかかわるとも読める。第10章第6節は、 S の未知のうちどれが、 S の成員が共同で追求するにふさわしいものかを問うときに、この違いを取り上げる。

四つ目の意味を厳密にとれば、手続きや熟議が選ぶであろうことが、そのまま、それが価値をもつということになる。現実の手続きについては、第12章第1節が取り上げるルソーの一般意志の構成的解釈、すなわち、適切に組織された民主的手続きの結果は、その手続きから生じたがゆえに一般意志であるとする解釈を、この種の見方として読むことができる。ただしトム・クリスティアーノ (Tom Christiano) とサミール・バジャージ (Sameer Bajaj) 自身は、これを価値の説としてではなく、民主的権威の考え方の一つとして述べている (Christiano and Bajaj 2024, §3.2.1)。アーノン・ケレン (Arnon Keren) が伝えるところでは、キッチャーは、福利の客観的な考え方を擁護できるかを疑い、自らの社会で前提とされている民主的な理想に頼ることで、何が善いかという問いを迂回しようとした。そして理想を、現実の研究アジェンダを測る基準として示した (Keren 2013, 236–237)。厳密な読みはまた、公正な手続きが決める権限を与えるという見方とも異なる。この見方では、手続きを経た決定は、価値の判断としては誤っていても、拘束力をもつ決定でありうる。つまりこの見方は、何に価値があるかではなく、誰が決めてよいかについての主張、すなわち正統性についての主張である (第12章第1節)。結果が先に挙げた四つのいずれかの意味で S にとって価値があるからといって、その結果が真になるわけではなく、その価値が、結果が真であることの証拠になるわけでもない (第2章第6節を参照)。 S の価値づけが、ある結果に依拠してよだけ支持されているという判断に入るかどうか、どう入るか、たとえばそれに依拠したときに生じうる誤りの重みづけを通じて入るかどうかは、本節の冒頭で第11章第4節に委ねた別の問いである。本節で問題にする相対性は、どの結果が重要であり、誰にとって重要かの相対性である。

これに対して、知識の価値を、 S に固有の何ものにもよらないものにするであろう立場が二つある。第一の立場は、ステイーヴン・グリム (Stephen Grimm) がジョナサン・クヴァンヴィグ (Jonathan Kvanvig) とマイケル・リンチ (Michael Lynch) に帰する原理で、あらゆる真なる信念は、少なくとも一応は、それ自体のために善いとする (Grimm 2008, 729–730)。あらゆる真なる信念が一応の価値をもつという主張だけでは、 S の未知は順位づけられない。それらに等しい価値を割り当てるわけでも、価値を比較する規則を与えるわけでもない。この立場を退けるグリムは、机の上の塵の数のように手の届くところにある無数の真理は、有限性から解放された探究者にとってさえ、それ自体のために学ぶに値しないと論じ、好奇心は真理そのものではなく、主題と特定の型の問いに向かうと論じる (Grimm 2008, 730–736)。第二の立場は、キッチャーが認識的重要性への伝統的なアプローチと呼ぶものである。それによれば、どの真理が重要かは世界の構造によって定まり、探究の目的は、たとえば法則や統一された自然像や客観的な説明を見いだすことになる。この立場は順序づけを与えるが、それは S から来る順序づけではない。そして S の未知のどれを追究するかは、他の何かを見つけるのと同じように重要なものを見つけることに帰着するだろう。ブラウンが伝えるキッチャーの批判によれば、実践的な答えと神学的な答えを除けば、法則、統一、根本的な原因の価値を正当化するのは難しく、文脈に依存しない説明によって重要な真理を選び出す試みは失敗する。キッチャーの言葉では、客観的な説明は「私たちの問いと関心を背景としてのみ」行われる (M. J. Brown 2010, 134–136)。どちらの立場でも、新しさと真理はそれだけでは S の未知のどれが追究に値するかを定めない。これが、作業的特徴づけが知識を最大化せよという指示を導かない理由の一つである (第1章第1節)。

ほかの立場には、知識の価値を、誰かの目的、関心、価値づけに相対的とみるものがある。それらが S のものであれば、順序づけは、先に区別した四つの意味などのうち、一つ以上の意味で S に相対的である。たとえばキッチャーの重要性の説では、好奇心と実践上の企てに相対的である (第1章第8節)。どの意味を探るかによって、構成が S の未知を順序づけるために何にアクセスできなければならないかが変わる。成員の善か、 S のもつ目的か、成員が実際に価値あるものとするものか、成員のあいだの手続きや理想化された熟議が選ぶであろうものかである (第12章第1節は、目的の順序づけについてこれを論じる)。研究では、そのいくつかは探究に先立って固定されているとは限らない。キッチャーの手続きでは、熟議者の選好を教育するのはありうる探究の重要性の説明である。エイギが伝えるところでは、科学の専門家がこの教育を担い、個々の計画の見込みを推定する (Eigi 2016, 67; 第12章第4節)。探究を行う側がそれを与えるところでは、構成が S の未知を順序づけるために参照する S の価値づけは、その一部を構成自身が形づくったものでありうる。価値は、探究する者に相対的でもありうる。グリムの説では、ある型の問いの答えを見いだすことは、そうした問いが私たちの自然な好奇心に結びついてい

るために、「私たちににとって」持続的な価値をもつ。その好奇心は、持続的な利害上の関心や認識上の関心に駆られたものとされる (Grimm 2008, 733–737)。グリムが論じるのは人間の探究者だが、この意味での自らの好奇心をもつ機械の *A* に当てはめれば、それが与える順序づけは *A* に相対的であり、*A* が *S* に属さなければ、*S* に相対的などの順序づけとも一致するとは限らない。さらに、これらの意味を横断する見方がある。デューイ (Dewey) の価値評価の説に根をもつこの見方では、目的は探究に先立って固定されていない。価値の判断は、何かに基づいて行為し、あるいは何かを価値あるものとするものの帰結にかかわる。そして、目的に到達する手段についての探究と、目的自体の帰結についての探究とが、目的そのものの見直しに至りうる (Anderson 2023, §§2.3, 3.1)。同じ説では、熟慮は、価値づけられているさまざまなものの重みを、与えられたものとして受け取るのではなく、選択の文脈のなかで割り当てる。以前の選択から持ち越した重みは、新しい状況に合わないことがあるからである。また価値判断は、それに基づいて行為し、その帰結が判断の予測どおりに価値づけられるかを見ることで試される (Anderson 2023, §§2.4, 3.2)。その見方では、*S* の未知のどれが追究に値するかは、それ自体、探究を通じて改められる。この見方を誤りの重みづけに広げるなら、ある種の誤りを別の種の誤りに対してどれだけ重く見るかも同様である (第11章第4節)。構成が、どの問いが扱いやすいかを変えるときに何が帰結するかは、第10章第6節で取り上げる。

研究の実践全体もまた、価値の対象である。そもそも研究がなぜ追究され、支えられるのかについては、いくつかの立場がある。以下ではそのうち五つを挙げる。それらは、研究の支援を論じた文書、すなわちヴァネヴァー・ブッシュ (Vannevar Bush) とエイブラハム・フレクスナー (Abraham Flexner) のものと、探究の目的、科学の自由の根拠、経験的知識の前提についての哲学的な仕事、すなわちそれぞれキッチャー、トルステン・ヴィルホルト (Torsten Wilholt)、マックス・ウェーバー (Max Weber) のものから来る。これらを並べるのは、配分は何を求めるかがそれぞれ異なるからである。五つの立場は重なり合い、ブッシュとフレクスナーはそれぞれ複数の立場にかかわることを述べている。立場はこの五つに限られない。なお、価値が認定や権威にどう入るかは、先の二つの点と四つの意味がそこでも用いられることを含めて、第11章第4節と第12章第1節で取り上げる。

第一の立場では、用途を考えずに行われる研究は、誰も予見できない用途の源として価値づけられる。ブッシュは、基礎研究が「科学的資本」をもたらすこと、どの一つの研究の結果も予測できないが、そうした企てのある割合はきわめて有用だと分かること、そして医学にかかわる発見はしばしば「遠く予期しない源から」来たことを述べた。フレクスナーは同じことを事例から論じた (Bush 1945, 9, 13–14; Flexner 1939, 544–548)。ブッシュが引き出した帰結では、配分には二つのことが求められる。一つは広がり、すなわち医学とその基礎にある諸科学の「前線全体」を発展させることであり、もう一つは長期の計画に十分なほど安定した資金である。価値づけられているのは用途なので、ブッシュは、支援によって便益が公衆に届くまでの時間を短縮できる特定の問題については、協調した取り組みも求めた (9–10, 26)。

フレクスナー自身の正当化は第二の立場だった。フレクスナーは、最終的な用途が実験室の仕事の「実際の正当化」であることを否定し、新しい科学的事実は、詩や交響曲のように、研究所が必要とするあらゆる正当化をそれ自身のうちに備えているとした (548–550, 552)。この立場では、知ることと理解することがそれ自体のために価値づけられる (第3章第7.1節を参照)。これを探るなら、この立場は研究の実践を持続させる理由を与えるが、そのなかの路線を順序づけはしない、と理解できる。あらゆる真なる信念がそれ自体のために善いという原理と同じく、それだけでは路線どうしを比較する規則を与えない。また、知識をそれ自体のために価値づけることは、それを手段としても価値づけることと両立するので (第3章第7.1節)、路線の用途の見込みを考慮に入れることを、それだけで排除するわけでもない。路線の順序づけは、この立場とは別の根拠によることになる。フレクスナーが用途の見込みを路線の指針から外したのは、知識をそれ自体のために価値づけることを越える根拠によってである。フレクスナーは『「用途」という語の廃止』を訴え、いかなる管理運営者も研究者が仕事をする「道筋を指図することなどとうていできない」とし、パストゥールやコッホらの発見がもたらした利益は「可能な用途という考えが彼らの心に浸透していたなら、決して生じえなかった」と述べた (548–549)。フレクスナーが擁護した研究所では、「各個人は自らの時間とエネルギーを思うままに使う」のであり、フレクスナーは一部の資金が無駄になることを受け入れていた (548, 551)。

キッチャーは第三の立場をとる。これは、キッチャーが純粋さの神話と呼ぶもの、すなわち、ケレンのまとめでは、道徳的・政治的な関心は「純粋な科学」の評価には関係がないとする考えに対して展開されたものである (Keren 2013, 240 n. 6)。キッチャーによれば、科学が目指すのは重要な真理、すなわち「重要な問いへの正しい答え」であって、どんな真理でもよいわけではない (Kitcher 2013, 207)。そして一見難解な問いの研究も、それが準備する実践上の成果によって重要性を得ることがある。トマス・ハント・モーガン (Thomas Hunt Morgan) がショウジョウバエに集中したことは、やがて、いまでは多くの生物医学研究の土台となっている分子生物学につながった (Kitcher 2013, 220)。実践的な帰結が遠い研究でさえ、探究を広げるほかの仕方とに比べたその価値についての判断を前提する。すべてを徹底して調べるにはあまりに多くの問いが残り、そのうちのいくつかを追究することは「時間とエネルギーの軽薄な浪費」になるからである (Kitcher 2013, 219)。さらにキッチャーはのちの論考で、研究が道徳的な誤りを比較的免れているところでも、「なしうる善をなし損なうことも過失である」と付け加える (Kitcher 2024, 240)。この立場を探るなら、キッチャーが引き出す帰結のとおり、どの路線も、それと競合する実践的な目標や好奇心の形との比較を免れず、手放された善はその費用に数えられる。キッチャーの結論は、実践的な帰結が遠い研究が正当化されないということではない。その正当化がその比較に依存するということだけである (Kitcher 2013, 220–221; Kitcher 2024, 240)。

第四の立場では、研究は、市民が政治的な選好を形成するための知識の源として価値づけられる。ヴィルホルトは、この根拠に立つ研究の自由についての政治的な論証を再構成する。この再構成では、民主的な権力が正統であるのは、その決定が市民の十分な情報に基づく選好を反映するからである。そのため、いかなる政治権力も、その選好のもとになる知識の生成を統制すれば自らの民主的正統性を損なうことになり、その統制を正統に行うことはできない (Wilholt 2010, §4)。同じ推論が論証の限界も定める。この論証は、結果が公共の議論の外にとどまる研究には及ばない。また、接近の多元性を前提する。科学者共同体が一つの接近に落ち着くなら、政府が好む一つの接近よりも市民をよい認識的位置に置くと考える理由はないからである。さらに、資源は有限なので、手段への要求を政治的な関連性に比例してだけ裏づける。そして、政治的な統制に反対する論証であるため、アジェンダを定める手続きを何も確立しない (Wilholt 2010, §§5–6)。これを探るなら、配分には、アジェンダを、研究の結果を用いて統治しようとする者から独立に保ち、多元的な接近を持続させることが求められる。そしてこの論証が答えずに残す問い、すなわち誰が関連性を定めるのかは、第2.1節の意味での権威に属する (第12章第3.2節)。

文化科学について書いたウェーバーは、第五の立場をとった。ウェーバーによれば、無限の現実についての有限な精神による認識は、その有限な一部だけが「知るに値する」(wissenswert) という前提に立つ。そして、どの部分が研究の対象になるかは、研究者とその時代を支配し、文化とともに変わる価値理念によって定まる (Weber 1904; Zeno.org text, 171, 183–184)。その説では、科学的真理の価値への信念そのものが特定の文化の産物であり、その真

理を価値あるものとしないうちに、科学は何も差し出すものをもたない (213)。専門化した仕事は、それに方向を与えた文化の問題が移っていくまでは、自らの材料をそれらの価値理念に照らして確かめることなく進み、そのとき科学はその観点と概念を変える (214)。この立場を採るなら、配分が従う順序づけは研究から導かれるものではなく、変わっていくと見込まれるものだと推論できる。そのため求められるのは、絶えざる比較よりも、向きを変える能力である。

以上の五つの立場が配分に求めるものは、研究者が機械になると変わりうる。第10章第5節は、これをフレクスナーとブッシュの擁護について取り上げる。二人の擁護は、研究者の好奇心に駆動される探索を配分の方法として採る理由として読める。しかし、人工の研究者が好奇心やそれに相当するものをもつかどうかは、人間の研究者の場合のように前提できない。

人工システムが研究に加わることをめぐる議論には、これらとは別の種類の立場がみられる。研究は、それを行う者にとって何であり、その者に何をもたらすかによっても価値づけられる、という立場である。フルヴィオ・カステラッチ (Fulvio Castellacci) らは、自らが規範的な断層線 (normative fault line) と呼ぶものを指摘する。科学の制度は主として知識を生むために存在するのか、それとも人間の能力を育て、専門性を認定し、業績に基づく配分の仕組みを保つためにも存在するのか、という分かれ目である。カステラッチらが描く産出を最大化する見方では、人工システムによる支援が研究者の育ち方に及ぼす効果は、学習が産出に影響する限りでしか問題にならない。これに対して能力の育成を重んじる見方では、人間の判断を人工システムで置き換えることには、産出の指標が捉えない費用がある。カステラッチらは、この選択は「経験的な根拠では決着しえない」とし、自らの提案する枠組みでは後者の見方を探る (Castellacci et al. 2026, §§2.4, 2.8)。マルセル・ビンツ (Marcel Binz) らがまとめた討論では、エミリー・ベンダー (Emily Bender)、カール・バーグストロム (Carl Bergstrom)、ジェヴィン・ウェスト (Jevin West) が、科学は「社会的な過程」であり、その行為者である「現実の科学者たち」は「論文に記録される結果と同じだけ、この過程の産物である」と書く。マシュー・ボトヴィニック (Matthew Botvinick) とサミュエル・ガーシュマン (Samuel Gershman) は、人間の科学者を全面的に置き換えるという見通しを、「技術的な限界が何の障害にもならない場面でさえ」「まったく魅力に欠ける」とする (Binz et al. 2025, 6, 8)。また二人は、人間の理解が科学の中心的な目標であり続けることを求める (7)。ボトヴィニックとガーシュマンが人間に残そうとするものは、第10章第6節で取り上げる。

ジョン・ダナハー (John Danaher) とスヴェン・ニーホルム (Sven Nyholm) は、自動化が仕事における達成の価値に何をもたらすかを問い、達成を四つの変数で評価する。産出物の価値、行為者の因果的な寄与が大きく、運によらないものであるかどうか、努力や時間や注意など、行為者が産出に注ぐもののコスト、そして行為の自発性である。二人の議論では、自動化はこのうち最初の三つをともに下げうる。人間の作業に残される産出物の価値を下げ、人間が注ぐもののコストを下げ、人間の活動と価値ある産出物とのあいだの因果的なつながりを断つからである。二人が素描する対応のうち、熟達や自律など、有意義な仕事のほかの側面を強調することは、産出物への自分の寄与が縮んでいけば空しく響くかもしれない、仕事の外の趣味、スポーツ、芸術的な企てに達成を求めることも、自動化がほかの領域に及べば限られうる、と二人は述べる (Danaher and Nyholm 2021, 230–235)。二人が研究に触れるのはアインシュタインの科学的な達成のような例を通してだけだが、研究に当てはめれば、この分析は二つの要求を分ける。一つは、人間が探究に加わり続け、それが伴う努力や熟達を保つことであり、これは活動を人間のために取っておけば満たせる。もう一つは、その参加が達成であり続けることであり、これには価値ある結果への大きく運によらない寄与が要る。活動を取っておくことがこれを満たすのは、取っておかれた活動がなおそうした結果を生む場合に限られる。ただし、人間の寄与が価値ある結果にとって大きなものであり続けるかどうかは、ほかの者が何を追うかによる。第1章第6節で述べたテッド・チャン (Ted Chiang) の短編は学術誌の論説の形で書かれており、その論説が、人間の研究を続ける理由としてこの種のものを挙げる。人間の研究者は「メタヒューマンが見落とした応用を見いだす」かもしれない、論説によれば、メタヒューマンはその優位のために人間の関心に気づかない傾向がある (Chiang 2000, 517)。思考実験として読めば、この事例は、最前線の結果をもつ社会が人間の *S* にとって重要なことを追わない場合に、第二の要求が満たされうることを例示する。第8章第3.2節が扱う種類のミスマッチである。出口康夫 (Deguchi) は、人工システムによる置き換えがもたらしうる脅威の一つを、知の主体のディストピアと呼ぶ。人間が唯一の知的主体であることに見いだしてきた誇りが失われることである。出口が一つの哲学的な選択肢として示す説では、探究はもともとつねに共同で行われてきたので、脅かされるとされる単独の、あるいは専制的な主体性は錯覚であり、本当に失われうるのは科学者の最終決定権である (Deguchi 2026a, 28–29, 31–32)。出口がWEターンと呼ぶこの説の原理を述べた論文では、この錯覚は、どの個人も身体行為を単独では行えないというテーゼから導かれる。その論文は、このテーゼを思考と決定にも及ぼす。そのため、そうした行為の主体はどれも、それに必要なすべてのエージェントからなるシステムであり、このシステムは空間的にも時間的にも身体のはるか外にまで及ぶ (Deguchi 2025a, 2–4, 7–9, 12)。

これらの立場は、何に価値を帰するかにおいて、ブッシュからウェーバーまでの五つの立場と異なる。五つの立場は、実践を、それが生む知識のために、あるいはその知識がもたらすもののために価値づける。これに対してこれらの立場は、実践を、特定の者たちによって行われるものとしても価値づける。表9.1の語でいえば、それらは行為者にかかわり、また、そうでなければ研究を行っていたであろう人間にかかわる。彼らは探究がどう組織されるかの帰結を負う者のうちにあり、*S* に属するとは限らない。*A* が人間であるところでは、研究者の形成は探究を支えるものの一つでもあり (第2.1節)、産出を最大化する見方も、その理由でそれを価値づける。産出を最大化する見方とこれらの立場とが分かれるのは、*A* が機械でありうるところである。そのときこれらの立場が捉える費用は、少なくともその一部が、誰が研究を行うかの変化であり、それは *S* が知るに至るものが変わらないところでも生じうる。したがって、これらの立場のもとでは、同じ知識を同じ *S* に届ける二つの研究の組織も、価値において同等とは限らない。産出だけを比べる評価では、人工の構成が研究に加わることの得失は決まらないのである。ここではさらに、最終的に誰が知るか、知るに至る過程で誰が変わるか、と分かる。産出を最大化する見方は、後者を将来の産出にかかわる限りでだけ数え、能力の育成を重んじる見方は、それをそれ自体として数えうる。どちらを採るかで、同じ置き換えの評価が変わる。これらの立場のいずれかを採るなら、配分には、*S* のどの未知を追究するかだけでなく、それを追究することへの誰の参加を保つかも求められる。出口の説では、さらに権威に、それについての誰の最終決定権を保つかが問われる (第12章第2.2節)。

本節で挙げた立場のいくつか、とくにブッシュとヴィルホルトのものは、国家やその市民に向けられていた。第2.1節の語でいえば、それらは支える者に向けられた理由であり、支える者と *S* は別の役割であって、同じ当事者が両方を占めることも、そうでないこともある。支援の理由は、なぜ実践が価値づけられるかとは別の問いを生む。すなわち、その理由が、研究に対する資源と権力を配分する仕組みに何を求めるかである。ここでの立場は、なぜ実践が価値づけられるのか、そしてそれが *S* の未知の順序づけに何を求めるのかにかかわる。二つの問いは、支援の理由がその仕組みをも制約するところで出会う。ブッシュの報告書とヴィルホルトの論証がそうである。第10章は、配分についての各議論に、それがどの目的を前提しているかを問う。

研究への支援を正当化することと、その成果を S に利用可能にすることは、別の問いである。パーサ・ダスグプタ (Partha Dasgupta) とポール・デイヴィッド (Paul David) は、支援の出所だけでは利用可能性が決まらない理由を示す。二人によれば、研究で生み出された知識は、本来他者に利用可能なのではなく、研究者が選んで公共の領域から遠ざけておける。そして、知識を開かれた形で追究することと秘密のもとで追究することとを分けるのは、研究が行われる際の開示の規範と報酬の仕組みであって、資金の出所ではない。そのためこの説では、公的資金を受けながら知見を秘密にしておく研究所は、二人が「科学」に対して「技術」と呼ぶ側に属する (Dasgupta and David 1994, 491–492, 495, 497–498)。インゲ・カウル (Inge Kaul) とロナルド・メンドーサ (Ronald Mendoza) も、知識を含む非競合的な財を「公共的にされうる」財として扱い、すべての人が実際に消費できるものだけを完全に公共的なものと数える。書かれた形の知識は、字を読めない人々には消費できない (Kaul and Mendoza 2003, 88–89)。これらの論者が論じているのは人間の研究者と制度である。研究を支えかつ行う構成に当てはめれば、これらの説は次のことを示唆する。その構成が成果を得たことだけでは、企業の社内研究 (第2.1節) と同じく、 S の成員がそれを利用できることにはならない。成果が成員に利用可能になるには別に成り立つ条件が要り、それは、誰が支払ったかではなく、構成が何を公開し、その公開の形が成員の使えるものであるかによる。

4. 評価することに含まれるもの

三つの関係はいずれも、大部分が評価を通じて実現される。路線は続けるに値するかどうか判断され、結果は依拠するに足るかどうかが判断され、両者のための基準と誤りの重みは、それ自体が目的に照らして評価される決定によって定められる。そのため、関係を構成の内側に置けるかという第V部の問いは、評価を内側に移せるかという問いと大きく重なる。そして評価を内側に移せるかどうかは、それが何からなり、誰が、どの基準に照らして、誰のために判断するのにかに依存する。本書はこれまで、評価という語を、こうした点を述べないまま用いてきた。用いたのは、人工の研究者についての主張について (序論第3節)、活動の別々の判断について (第2章第8.1節)、そして探究のなかの査定について (第5章第4節) である。そこで本節は、第V部が用いる最小限の評価の説明を述べる。配分が向きを決める手がかりとする信号については第10章第3節が、結果の評価とその時間にわたる変化については第11章第1節・第7節・第8節が、基準を定める決定については第12章第2節が、それを用いる。

この説明は四つの点からなるが、それは評価の理論をなすものではない。たとえば四つは、複数の評価者の判定がどう組み合わせられるかについては何も言わない。最初の二つは、上の用法に共通するものを述べ、研究に固有のものではない。後の二つは、研究において評価がとる特徴を述べる。この二つを選んだのは、以下の章がそれに依拠するからである。

第一に、評価においては、ある種類のものとして捉えられた対象が、目的に結びついた基準に照らして判断され、ある当事者 (人、組織体、あるいは人工の構成) がその判断を用いて何かを選ぶ。出版する、資金を出す、依拠する、撤回する、といったことである。測定は、それ自体が評価でなくても、評価への入力になりうる。評価は基準を加えるからである。評価することはまた、第3節の意味で価値づけることとも異なる。価値づけることは、何が重要とされるかにかかわり、評価することは、そうした価値づけを、ある用途のために対象に当てはめる。この対比には先例がある。デューイは、価値づけること、すなわち大切にすること (prizing) と、評価、すなわち査定すること (appraising) とを区別した。後者は、何かが価値づけるに値するかどうかを問うときに生じる (Anderson 2023, §2.1)。ここでの用法はそれより狭い。基準は目的を前提するので、「誰の目的によってか」という問いは第3節に戻る。

第二に、評価者と受け手は異なりうる。人間の研究では、たとえば、査読者が編集者と読者のために判断し、審査会が資金提供者のために判断してきた。認定は結果に依拠するであろう者に向けられ、支援の評価はそれを供給する者に向けられる。評価がさらに誰に対して責任を負うのかは別の問いであり、説によって答えが分かれる。表9.1 の役割は、説がその答えや判断そのものを置く先となる当事者を、互いに区別している。たとえば、ヴィルホルトの説では、どの誤りが重要かの判断は、結果に依拠する者たちが期待する価値づけに応じるものとされる。これに対してヤコブ・オルトマン (Jakob Ortmann) は、道具の使用が課題に適切かどうかの判断は価値判断であり、信頼できる者に委ねることができるとする。そのためこの判断は、依拠する者の側ではなく評価者の側に置かれる (第11章第4節・第5.2節、第12章第1節)。評価が何をするのかを、いくつかの評価を束ねてきた取り決めから分けることで、第11章第7節は、出版前の査読という形を前提することなしに、 S のための認定が何を要するかを問えるようになる。

第三に、研究における評価は代理指標に依拠する。その目標は、判断がなされる時点ではしばしば観察できないからである。たとえば、後の追試と批判は、 S にどの知識を帰属することが根拠づけられるかや、その知識が保持され続けるかどうかを変えうる。しかし、以前に偽だった主張を、以前に真だったことにはしない。ある結果が重要かどうかは、しばしば後の受容を通じてしか現れない。将来に向けた評価は、まだ存在しない結果を判断する。そして結果に依拠する多くの者は、それを自ら吟味できない (第1節)。したがって、判断、得点、通過した点検、そして人間の研究でいえば引用のような後の受容の記録などが、目標の代わりを務める。そして、目標をそうした尺度と同一視する説を除けば、それらは目標そのものをなすのではなく、その代わりを務めるにとどまる。ただし、尺度が目標でないからといって、尺度が記録する評価の営みがまったく目標の外にあることにはならない。判断や通過した点検や受容の記録は、批判、応答、使用といった営みに属する。社会的な組織が根拠づけや保有に入るとする説 (第2.1節) では、こうした営みは、結果が S にとって知識としての地位をもつこと、あるいは S の知識となることの一部でありうる。そこでも、主張が真であることはそうした過程が生み出すものではなく、過程の記録はそれについてはなお代わりを務めるにすぎない。本節の説明では、少なくとも目標が後になってはじめて定まり、重要性を示し、あるいは生じるという点については、代理指標の使用は特定の制度の欠陥からではなく目標の構造から生じる。これに対して、結果に依拠する多くの者がそれを自ら吟味できないことは、依拠する者の能力に依り、そのため人工の S については、成員のあいだで証拠がどう共有されるかに依る (第2.1節)。重要なのは、代理指標が帰結をもつようになったとき、代理指標と目標の関係がどう振る舞うかである (第10章第3節、第11章第8節)。

第四に、評価されるものが S にとって新しいとき、 S は、結果を自らが望んだものに照らして単純に確かめることはできない。 S にとって新しい結果は、 S がまだ答えをもっていなかった問いに答えるものだからである。研究を依頼したのが S 自身であっても、 S とは別の支える者であっても、このことは変わらない (第2.1節、第8章第3.3節)。ただし、答えをもたないことは、基準をもたないことと同じではない。形式的に述べられた定理を定められた推論規則のもとで証明する場合や、登録された予測が出来事を持つ場合のように、何がその問いを決着させるかを問いが前もって定めているところでは、新しい結果もその基準に照らして判断できる。その場合も、基準が問いを捉えているかどうかは別の判断として残る (第3章第5.1節、第11章第1節)。そうでないと

ころでは、評価者はまだその結果のための基準をもっていないかもしれず、過去の判断から学ばれた基準は、新しい結果がその範囲の外にあるところでは失敗しうる（第11章第8節）。

これらの語でいえば、配分は主に路線と提案についての将来に向けた評価を通じて、批判と認定は主に結果についての適反的な評価を通じて実現される。認定の評価は、結果に依拠するであろう者に向けられる。権威は、両者の評価がなされる基準と誤りの重みを定め、改める当事者資格である。

5. 統制と、それを制限する権能

関係が構成の *内側* にあるとは、どういうことか。第7章の語でいえば、関係が内側にあるのは、それを実現する決定の起動、監視、方向転換、回復が構成の境界の内側で組織されており（同章第2節）、しかも作業の供給と、その作業を許す授權や結果を認める承認とがともにそこにあるときである（同章第4.2節）。関係は、これらの点の一方では内側にあり、他方では内側にならないこともある。第1節の例でいえば、点検は内側で供給されていても、システムを止めたり取り替えたりする授權はその提供者にあるかもしれない。また、関係が内側にあるかどうかは、組織の名前や数によっては決まらない。二つの組織を一つにまとめても、記録を別に保持し、判断を拒み、資源の使い方を独立に決める関係が変わらないことはあり、逆に、別々の組織のままで、それらがすべて一つの判断に従うことはある。そのため、関係が内側に移ることで何が変わるかは、境界の呼び名ではなく、決定の組織、授權と承認、そして本節の後半で名指す、ある部分が分かれて進むことを可能にする権能が、どこにあるようになるかとして問われる。第2.1節が当事者資格として定義した権威については、それを実現する決定がこの意味で内側で組織されていても、構成がそれらの決定を下す当事者資格をもつかどうかは、なお別の問いとして残る。第1節の第三の主張が、決定がどこで下されるかではなく、誰が決めてよいかにかかわるのはそのためである（第12章第1節）。したがって第V部の問いは統制の概念を前提している。何が内側に移っても人間が統制を保つ、という答えもありうるが、それも統制とは何かを述べてはじめて検討できる。本節は、第V部が用いる限りでそれを述べる。哲学者君主については第10章第2節が、批判の経路がいつ独立かについては第11章第2節が、決定権、名目以上の当事者資格、改める能力を取り除く決定については第12章第2節・第4節・第7節が、それを用いる。

ここで用いる意味での、関係に対する統制とは、その関係を実現する当事者や過程が何をするかを決定する能力である。統制という語は四つのものをひとまとめにしうるので、第V部はそれらを分けておく。第一は、決める形式的な権利である。第二は、それを行使する実効的な能力であり、いま定義した意味での統制はこれにあたる。これは、時間、情報、有能さの欠如や、他者のインフラストラクチャーへの依存によって空洞化する。第三は、統制をどう使うのが最善にかかわる知識である。たとえばショーとヴィオラが伝えるところでは、キッチャーは哲学者君主に、研究の進路を指図する特権とともに、各路線の見込みについての知識を与えたが、統制はそれを供給しない。そしてヴィオラが伝えるキッチャーの分析では、その知識があってもなお、労働力をどう分けるのが最善かは問いとして残る（Shaw 2018, 155; Viola 2015, 5–6、第10章第2節）。第四は、決定を下すことなしに、決定がなされる選択肢を形づくる影響である。既定値、順位づけ、インフラストラクチャーがそうする。第二と第三は、第12章第2.2節が分ける実効的な能力と認識的根拠づけの問いに対応する。第一は、同節が形式上の割り当てと呼ぶものである。その保持者がその権利をもつ資格があるかどうかは、同節が授權された権力と呼ぶもう一つの問いであり、第2.1節の意味での当事者資格がこれにかかわる。その答えは、採る権威の根拠に依存する（第12章第1節・第3節）。第四を加えるのは、ある当事者が最初の三つを保持しながら、別の当事者がその当事者の選ぶ選択肢を形づくることがありうるからである。この区分は、ある当事者が統制を保つという主張に向けうる問いに従っている。その当事者がその権利を保持しているのか、それを行使できるのか、どうすればよいかを知っているのか、そして何か別のものが、その選ぶ選択肢をすでに形づくっていないか、である。その目的は、人間が統制を保つという主張について、四つのどれを主張しているのか、あるいはどれをひとまとめにしているのかを問えるようにすることである。この区分は、たとえば、ある時点で保持される統制と、構成が変わっても持続する統制とを分けない。それは第12章第6節・第7節が取り上げる。

統制に対して、それを制限する権能がある。ここでまとめて名指しておく。それは、構成のある部分がもちうる次の権能である。自らの目的を変える権能、資源を拒む権能、自らの証拠を別に保持する権能、他の部分の更新を止める権能、そして自らの記録を携えて離脱する権能である。各権能を入れるのは、ある部分がそれを保持すれば、その部分の進路が、別の部分が指示するものから分かれうるからである。合わせると、それらは、配分と批判について、第7章第4.2節が作業の供給から分けた授權と承認に具体的な形を与える。これらは第10章から第12章が用いる権能であって、網羅的な一覧ではない。また、どの権能もある部分が分かれて進む能力を確保するので、第12章はこれらを独立の権能とも呼ぶ。統制と同じく、各権能は、部分に割り当てられた権利として形式上保持されることも、部分が実際に行使できる能力として実効的に保持されることもあり、両者は離れうる。ある部分は、他の部分の更新を止める権利を割り当てられながらそれを行使する手段を欠くこともあれば、そうした権利なしに更新を止められることもある。権利は能力を確保しうるし、能力は権利を実効的なものにしうるが、どちらも他方と同じものではない。第10章第2節は、そうした権能が第1節の百のインスタンスのあいだに分かれているかを、それらが互いに独立であるかの試金石として問い、その問いを、各インスタンスや全体が研究を遂行しているかという問いから区別する。第11章第2節は、別々に保たれた批判の経路がいつ独立かを問う。組織の上での分離だけではそれは確保されない。第12章第7節は、権能のどれが権威を改めうるものに保つことにかかわるかを問う。特定の種類の決定に対する権利は、これらの権能のいくつかの形式上の側も含めて、第12章第2.2節で一度挙げる。第10章第2節と第11章第2節は、主としてこれらの権能の実効的な側を扱う。

統制もこれらの権能も、第2.1節の意味での当事者資格ではない。統制とこれらの権能は、ある当事者が何を決定する権利を割り当てられ、何を実際に決定できるにかかわる。当事者資格は、他者との関係で、その当事者が何を提案し、何を決定し、何に異議を唱え、何を拒み、何を改める資格をもつかにかかわる。両者が互いにどうかかわるかは、実効的な能力を欠くために当事者資格が名目にとどまりうるかどうかも含めて、第12章第1節・第4節で取り上げる。

6. 第V部の進め方

第2節から第5節は、関係、役割、そして価値づけること、評価すること、統制することが含むものを述べた。本節は、第V部がそれらをどう用いるかを述べる。関係の分割が何に依っていたかを仕分ける手立て、第1節の主張のための枠、第V部が用いる思考実験の動機、そして章の順序である。

手立ては仕分けであり、ここでは結果としてではなく、当てはめるための区分けとして述べる。第5章第6.5節は、すでにこの仕分けを、そこで扱った要素の役割に当てはめた。第2.3節は、関係の分割が供給したかもしれない財について、それぞれが何に依っていたかを問うた。第V部は四つのありうる答えを区別する。ある財は、**人間という実現者についての前提**、すなわち実際に関係を保持していた人々そのものについての前提に依っていたかもしれない。たとえば、中央の指揮者は一人の人間の注意しかもたないこと、配分は研究者の動機を経由すること、交代が人の寿命とともに起こること、である。ある財は、**現在の科学のシステムの組織についての前提**、すなわち関係が歴史のなかでどう組織されてきたかについての前提に依っていたかもしれない。たとえば、一次データとそれを生み出した過程に近づけるのは研究者だけであること、査読という関門が乏しい編集上の注意と紙面を割り当て、査読が少数者にかかってきたこと、である。第2.3節が引いた見方では、そうした組織はありうる組織のうちの一つの、歴史のなかで形づくられた実現であり、実現者が人間のままでも組み直されうる。人工の構成がこの二種類の前提を取り除くところでは、それが生んでいた問題は、ただ消えるのではなく、形を変えることもある。ただし両者は崩れ方が違う。前者は実現者が変わるときに崩れ、後者は実現者が変わらなくても組織が変われば崩れる。ある財は、**探究の条件**、すなわち何がそれを実現するかにかかわらず成り立つものに依っていたかもしれない。たとえば、第10章第2節が検討する議論によれば、どの路線が成功するかを前もって知る者はいないことである。あるいは、第11章第2節が述べる批判の見方によれば、独立した批判は、既存の点検がすでに与えるものを越えて誤りについての情報を加えることである。第11章は、そうした証拠上の寄与を、その寄与が失われることに対する保護からも、失敗の解釈からも分ける（第1節～第2節）。関係を一つの構成に再結合しても、そうした条件は取り除かれない。ただし構成はそれらを自らの内側で満たしうる。そしてある財は、**構成の外の当事者との関係**（ S が構成の外にあるときは S を含む）に依っていたかもしれない。たとえば、誰の問いが S の重要とみなすものの根にあるのか、あるいは他者が探究の帰結を負う取り決めを誰が定めてよいのか、である。人工の構成は、そうした関係が成り立ちうる仕組みを築くことはできる。しかし、その関係が相手の参加や承認を含むものであれば、それを片側から完成させることはできない。他者にとって何が利益になるかのよう、に、相手の参加を必ずしも要しない関係について、どこまで人工の構成の側で満たせるかは、採る説による。以下では、構成の外の当事者との関係を、略して S との関係と呼ぶことがある。この略称は表9.1 のほかの当事者を含む。

四つの答えの種類は、研究の主張の要素（第1章第2節）におおよそ対応する。最初の二つは、関係を誰がどう実現してきたかに対応する。人間の研究では行為者以外の当事者もそれを実現しており、関係が内側に移れば、行為者である構成がそれにあたる。そのうち一つ目は実現者そのものに、二つ目はその実現が組織されてきた仕方にかかわる。三つ目は、何がそれを実現するにせよ、認識的達成をもたらそうとする試みに対応する。四つ目は、表9.1 のほかの当事者にまで広げた S に対応する。この仕分けは、「関係が内側に移ると何が変わるか」という問いに、どの財についても、その答えがとりうる四つの形を与えるためのものである。財が前の二つの種類の前提に依っていたなら、それが対処していた問題は前提とともに消えることも、形を変えて残ることもある。そこで問われるのは、そのどちらであるか、そして残るなら何がその場所を占めるかである。探究の条件に依っていたなら、人工の構成はそれを内側で満たしうるが、満たしていることを示す必要がある。人工の構成の外の当事者との関係に依っていたなら、能力だけからの推論ではなく、その関係についての証拠が要る（後で述べる枠を参照）。ただし、一つの財が複数の種類にまたがることもある。また、人工の構成がもたらしうるもののなかには、崩れる前提ではなく、固有の要件を伴う新しい実現の特徴であるものもある。成員のあいだの違いと誤りの相関が作り方によって決まること（一つの基盤システムの複製では誤りが相関しうる一方、違いを作り込むこともできる）、研究を遂行する構成そのものも供給する提供者、思うままに増やせる成員（それが S の成員である場合について、第8章第1.2節は、研究による S の知識の増加と、成員の範囲の変化による増加とは、 S の境界を定めておかなければ区別できないことを見た）といったものであり、各章はそれが生じるところでそれに触れる。

実現が固有の要件を伴うことは、人工の実現に限らない。ドナルド・キャンベル（Donald Campbell）は、自らが試みる記述的認識論が扱う知識はつねに何らかの乗り物に具現されており、その乗り物が自らを保つための要件は、対象を妥当に写すという要件と張り合う、と考えた。キャンベルによれば、科学共同体もそうした乗り物であり、世界の像を改善できるようになる前に、成員を勧誘し、報い、引き留め、指導者層を保たなければならない。こうした要件は共同体が運ぶものを偏らせうるが、科学の規範、実験の競争的な再現、実験の設計がそれを部分的に打ち消す（Campbell 1997, 10–12, 27–30）。この見方を人工の構成に当てはめると、人間という実現者についての前提や、その共同体の組織のされ方についての前提を取り除くことは、乗り物をなくすことではなく、乗り物を替えることだと示唆される。そのため、再結合が人間の共同体の偏りから探究を解放するという主張は、構成自身の乗り物の構造と維持に照らして量ることができる。成員のあいだの違いと相関が作り方によって決まることなど、先に挙げた特徴は、この乗り物の構造と維持のなかに現れる。人工の構成も、移働し、記録を保ち、更新され、あるいは引退させられる何らかの仕組みによって維持される。そのため問いは、人間の共同体の偏りが消えたかだけでなく、どの維持の条件が代わりに探究を方向づけるかになる。さらに、結果を判断するのと同じ評価によって、次にどの研究者を作り、どれを引退させるかが決まる人工の構成、すなわち表9.1 の提供者または維持者の役割の一部がその評価と結びついている人工の構成では、評価に合う問い方や方法が、出力だけでなく担い手そのものにも受け継がれ、将来の内部の批判が届く範囲が狭まるかもしれない。この機構はキャンベルの議論だけからは導かれないが、評価と研究者の生成が結びついた人工の構成については、そう示唆される。第10章第3節は、測られた産出の高い研究者を複製する配分について、これに近い力学を扱う。評価と評価されるものとの結びつきは、別の形で、第7章第4.5節と第8章第3.2節にも現れていた。前者では、評価を構成に取り込むことによって、それが結果を産出した側の仮定と結びつきうる。後者では、 S の基準が A とともに動く。ここではその結びつきが、研究者を作ることそのものに及ぶ。

四つの種類のあいだの線がどこに引かれるかは、採る説に依存する。ゴールドマン（Goldman）の過程信頼性主義は正当化された信念についての説であり、この説では、信念は信頼できる過程によって産み出されれば正当化される。第11章第5.2節のようにこれを認定に当てはめれば、認定が依るものの多くは探究の条件に属する。そのとき確保しなければならないのは、認定された結果の背後の過程が実際に信頼できることであり、その過程を人間が実行することは要求されないからである。これに対して、ロンジーノの批判的文脈的経験主義（第11章第3節）では、認定は、関係する批判的な共同体の内部と共同体間関係にも依る。この説では、証拠の関係は背景仮定に媒介され、そうした仮定の受け入れを主観的な選好が恣意的に左右することへの歯止めになるのは、科学者共同体の内部や共同体どうしの批判的な相互作用だけである。その相互作用は、批判の場、批判の受容、公的な基準、知的権威の調整された平等という、共同体に適用される規範を満たさなければならず、何が知識とみなされるかは、そうした相互作用によって決まる（Longino 2025, §5）。そのため認定が確保しなければならないものは、受け入れる共同体がどの視点を利用できるか、そしてその批判がどう組織されているかに依存する。その共同体が S である場合には、これらは S との関係である。共同体が構成の内側にある場合には、 S の視点とその批判に入るかどうか、どう入るかは、さらなる問いになる（第11章第3節）。重要性が世界によって定まっているなら、どの未知を追究するかは探究の条件に属するだろう。これに対して、たとえば

キッチャーの重要性の説では、それは S との関係に属するだろう。以下の各章は、仕分けをその関係に当てはめ、線がどこで動くかを記録し、第12章第9節がその結果を集める。

枠は第1節の三つの主張にかかわる。構成が探究をうまく配分する、その結果を S のために認定する、あるいは他者に代わって探究を方向づけてよい、という主張は、第2節の関係についての主張である。採る説のもとでそれらの関係が構成の外の当事者との関係に属する限りで、そうした主張には、能力だけからの推論ではなく、それらの関係についての証拠が必要である。配分については、それが向きを決める手がかりとする信号が、なお S の問いに由来するかどうかが問いになる。認定については、その認定が、関連する情報に加え、その所見自体も検討に開かれている批判に耐えるかどうかが問いになり、説によっては、その認定が S にとって受容可能か、依拠する者が手に入れられる理由があるか、結果について答える誰かがいるかも問いになる。権威については、採る権威の根拠が求める関係が、その決定の帰結を負う者とのあいだに成り立っているかどうかは問いになる。実際の参加を通じた民主的な授権では、行使される権威は影響を受ける者の参加に基づく。非恣意性の民主的な読みのもとでの共和主義の説では、権力は、それに服する者たちがそれを統制する程度に応じて恣意的でない。そのためこれらの根拠では（後者については、第12章で概観する説からの推論による）、彼らが異議を唱え、拒み、あるいは改める当事者資格が、それ自体として保たれることが求められる。これに対して、キッチャーの基準は取り決めにその帰結によって判定し、結果に基づく根拠の可謬主義的な形は、制度の誤りを検出し訂正する能力を、結果を良いものとする条件の一部に数える。そのためこれらの根拠では、当事者資格は、たとえば成員の視点や失敗が知られる経路として、手段として重要になる。結果に基づく根拠の単純な形は、結果がより良いかどうかだけを問うので、当事者資格を求めず、結果が実際により良いかどうかを示しうる評価を求める（第12章第3節・第7節）。それぞれが成り立つか、どこまで成り立つかが、第10章から第12章の検討するものであり、各章は最後に、その検討が対応する主張にとって何を含意するかを述べる。どの S を特定するかも、主張の範囲を定める。企業が S であり、主張が企業にとって新しい知識にかかわるなら、より多くの関係が企業自身の取り決めの内側にありうる。ただし帰結を負う者はなおその外にいるかもしれない。人類にとって新しい知識の主張は、関係が問われる当事者を広げる。 S が、主に構成自身のインスタンスからなる人工の社会であるなら、 S の成員の多くは構成の内側にあるので、彼らとの関係が求めるものの多くはそこで満たされうるものであり、第四の種類から外れる。ただし、その研究の帰結を負う者はなおその外にいるかもしれない。第12章第5節は、外の誰もその帰結を負わないという極限を含めて、この場合を取り上げる。

第V部はまた、四つの思考実験を用いる。一つは、科学の労働力に対する完全な統制と、理論の客観的な長所を見抜く眼の両方をもつ、キッチャーの哲学者君主である（第10章第2節）。二つ目は、一つの基盤システムから派生した多くの研究者からなる、インスタンスの共和国である（第11章第3節）。三つ目は、自らの証拠を構成に保持され、その構成に対して素人として立つ社会である（第11章第6節）。四つ目は、行為者であると同時に S の全体でもある単一の構成という極限である（第12章第5節）。これらのいくつかは、人間の探究についての哲学者の理想化、すなわち人間という実現者についての前提や、その探究の組織のされ方についての前提が何を寄与するかを取り出すための工夫を下敷きにしている。人工のシステムが研究を行うようになるにつれて、これまでの科学の組織のされ方が前提してきたことは変わるかもしれない。そして、かつて思考実験にすぎなかった考えが、検討に値する可能性として、あるいは研究が次にどう組織されうるかを考える視座として、ふたたび浮かび上がりうる。第2.3節で引いた本書の著者のエッセイは、この動機を先に述べていた (Takagi and an AI agent 2026)。第8章第3.4節は、そうした回帰がとりうる形の先例の一つ挙げた。ファイヤアーベント (Feyerabend) は、通約不可能な理論は比較可能な帰結をまったくもたないことがありうると論じたうえで、理論の予測の成功を、感覚器官をもたず理論が組み込まれたロボットが、その領域で道を見いだすよう訓練された存在と同じように反応するかどうかで判定することを提案した。この基準は意味に言及しないので、理論間の翻訳を迂回する (Feyerabend 1962, 94–95)。最近、チェスという限られた領域で、人工のシステムから取り出した概念を用いて、これに関連する試みがなされた。著者たちが概念実証として示す小規模な研究で、4人のグランドマスターが、その概念を示す局面を、ラベルや言葉による説明なしに学んだのちに、未見の問題をより多く解いた (Schut et al. 2025, §§1.2, 3.2, 4.1)。

オットー・ノイラート (Otto Neurath) は、思考のなかにだけある編成を検討する理由を、経済の秩序について論じた。思考と像のなかにだけあって現実にはない生の秩序をすべてユートピアと呼び、その語では可能かどうかについて何も言わないことにして、ユートピアを、理論そのものに属する社会工学者の構想として扱い、一つずつではなく並べて展開し検討することを提案した。また「可能なものをも見渡すときにはじめて、人は現実的なものを十分に認識する」とも述べており (Neurath 1919, 228–231)、これを、ユートピアをあわせて検討する理由として読むことができる。研究を念頭において読み直すと、ノイラートの前提の一部は経済に固有のものだと読める。生産と消費の量を現物で計算し規制できること、そして社会工学は、純粋力学と同じく、よく知られた要素を扱い、新しい力に出会わないことである (228–229, 231)。どちらも、研究については前提にできない。第10章第2節が検討する議論によれば、探究が何を生むかは前もって知られず、また人工の構成は、先に述べた、固有の要件を伴う特徴をもたらしうるからである。ノイラートのほかの前提には、経済に依存しないものもある。たとえば次の三つである。一つは、現在の秩序は偶然的であり、それと過去とに限られた理論は、戦争や革命のような激変に対処できないということである (231)。これは、第2.3節がその系譜を記録した見方の一つの形である。二つ目は、与えられた目的に対して、一定の前提のもとで等しく正しい解がいくつも見つかりうるということである (229)。三つ目は、構想は、技術者が鉄の弾性や銅の破断強度を考慮するように、それを実現する者たちの性質を考慮しなければならないということである (229)。ノイラートはこの三つを研究の組織については考えていない。また、三つは同じ強さの主張ではない。本書は、一つ目と三つ目を、研究の組織についても成り立つものとして引き継ぐ。三つ目を引き継ぐのは、上で見たキャンベルの議論が示唆するように、探究の組織を実現するものは、人間であれ人工のものであれ、組織の側が考慮しなければならない性質をもつからである。二つ目、すなわち等しく正しい解がいくつもありうることは、目的と比較の基準によって成り立つことも成り立たないこともあるので、研究の組織について検討する可能性として扱う。そして第V部は、これらの前提のもとで、思考実験のうち人間という実現者についての前提を取り除くものを用いる。そうした前提を欠いた構成を検討するのは、関係の既存の分割が何を供給していたかを見るためである。ノイラートの考えのうち、次の二つは引き継がないか、部分的にしか引き継がない。一つは、そうした構想が、政党の綱領として、また自らの実現を助ける予言として、行為を導くことも望んだ点であり (228–229)、これは引き継がない。もう一つは、生の秩序の大部分を目標に向けて形づくれるという前提 (228) であり、これはここでは部分的にしか成り立たない。誰も決めずに供給されていたものを設計で定められるかどうかは第10章第5節が問うことであり、また構成の外の当事者との関係のうち、相手の参加や承認を含むものは、片側から完成させることができないからである。第V部の各思考実験は、それが現れるところで、どの前提を取り出すかを述べる。君主は、統制と、理論の長所を見抜く眼とを分ける。共和国は、批判者の数とその独立性とを分ける。素人としての社会は、内容につ

いての服従、すなわち内容について他者の判断に従うことと、目的についての当事者資格とを分ける。そして極限は、構成の外の当事者との関係に依るものと、探究の条件とを分ける。

章立ては第2節の三つの関係に従う。第10章は配分を、第11章は批判と認定を、第12章は権威を取り上げ、第12章は第V部が見いだしたことで終わる。各章は冒頭で、その章の節がどう進むかを述べる。この順序には二つの理由がある。一つは、配分と認定が一階の関係であり、権威が二階の関係であることである（第2.1節）。もう一つは、第4節の語でいえば、配分は主に将来に向けた評価を通じて、認定は主に遡及的な評価を通じて実現され、権威はその両者の基準を定める当事者資格だということである。そのため権威はそれらの後に取り上げる。第7章第5節が第V部に残したもののうち、権力は第12章第2節・第6節で、監督は第12章第4節で、責任は第12章第6節で、訂正は第11章第8節と第12章第2節・第6節で取り上げる。

百のインスタンスの例は、いまや結びついた三つの構築上の問いを提起する。その配分は、追究するものを改めながら、生産的な代替案を保てるか。その批判は、誰に対して、結果に依拠する理由を与えうるか。さらなる研究者の創出を含め、その方向を誰が授権し、改めてよいか。本章の準備から、すでに見えていることもある。一つは、個々の研究者の良さから研究システムの良さは読み取れないことである。メイヨ＝ウィルソンらの結果を当てはめれば、単独で評価されたエージェントの成績は、証拠を共有する他のエージェントのあいだでの成績を示さない（第1節）。二つ目は、人工の構成が同じ知識を同じ S に届けても、研究の組織が同じ価値を保つとは限らないことである。研究者の形成を重んじる立場では誰がその過程に加わったかが、ダナハーとニーホルムの達成の分析を当てはめれば人間の参加が達成であり続けるかが、出口の説では誰が最終決定権をもつかが、なお残る（第3節）。三つ目は、説によっては、行為者の外の何ものもその産出を可能にせず、指示せず、制約していないときにも、その出力が S にとって何として数えられるかが外の関係に依存しうることである（第2.4節）。四つ目は、キャンベルの見方を当てはめれば、人間という実現者や人間の共同体の組織についての前提を取り除くことは、知識を運ぶ乗り物をなくすことなく替えることであり、問いは、どの維持の条件が代わりに探究を方向づけるかになることである（第6節）。

三つの問いへの答えが求められるのは、それによって、関係が人工の構成の内側へ移ったとき研究に何が起こりうるかを描けるからでもある。本章の準備からも、以下の章が確かめる変化の見通しが得られる。第一に、表9.1の語でいえば、 S が評価者の位置から受け手の位置へ移りうる。 S にとって新しい結果は、 S が望んだものに照らして単純には確かめられず、照合できる別の情報源もないことが多い（第4節、第12章第4節）。そのうえ批判と認定がともに構成の内側にあり、 S の成員がどの分野でも構成に対して素人の位置に立つなら、成員が担うのは、主に、構成の判断を用いて結果に依拠するかどうかを選ぶことになりうる（第11章第6節・第9.3節）。第二に、研究の方向が、構成の評価しやすいものの側へ寄りうる。配分は評価の生む信号を手がかりに向きを決め（第2.4節）、研究の評価は代理指標に依拠する（第4節）。そのため、判定しやすい課題での成功が次の配分を正当化し、追究される問いが内側から狭まりうる（第10章第4節・第6節）。第三に、多様性が設計の変数になりうる。人間の研究で誰も決めずに生じていた路線や見方の違いは、作り方によって定められるか、失われうる。しかもそれは、後にどの違いが役立つかを知らないまま起こる（第2.3節、第10章第5節）。第四に、評価と研究者の生成が結びつくと、評価に適う問い方が担い手そのものに受け継がれ、将来の批判が届く範囲が狭まりうる（本節、第11章第9.3節）。さらに、これら四つの見通しが依存するもの、たとえば誰が批判し、何を基準とし、路線をどう割り当てるかは、誰かが最初に置いたまま固定されるとは限らない。人工の構成が探究を進めながら批判者を生成し、基準を学び直し、路線の割り当て方を改めていくなら、研究のやり方そのものがその探究のなかで変わっていく。そのとき、評価される側が評価の基準点、すなわち評価がそれに照らしてなされるもの（批判者、基準、照らされる社会など）を作り出しうるものが、四つの見通しのどれにも通じる形として現れうる。そして以下の章で見る説のいくつかは、共同体に向けて規範を立てている。その共同体では、批判者、代替案、同業者は、判定される側が作ったものではなく、外で出会うものである。そうした規範が、人工の構成自身が作るものによって満たされうることになる（第12章第9節）。第10章から第12章は、ここで示した諸説のもとで、上の三つの問いを考える。その比較はまた、人間の研究で分かれた当事者が何を供給していたか、それらの機能のどれを人工の組織が保ち、変え、あるいは失いうるかを検討する。

第10章 探究はどのように配分されるのか

分業、多様性、誰の未知か

1. 配分はなぜ生じ、何がそれを難しくするのか

第9章は、研究が S のために方向づけられ、認定され、統治される三つの関係を区別し、その第一を配分と呼んだ。 S の未知のあいだでの、注意、労力、資源の分布である（第9章第2.1節）。この関係は、作業的特徴づけの語が、特徴づけの外にある条件と出会うところで生じる。その語とは S にとって新しいである。 S は無制限に多くのことについて無知であり、新しさは、そのうちどれが注意に値するかを決めない（第1章第8節）。それだけなら、選択は強いられない。選択を強いるのは、探究の手段が有限であることである。ある探究の路線が受け取るものを、別の路線は受け取らない。注意、労力、装置、試料、時間は限られており、そのどれもが別様に使われた。この関係に当てはめると、第V部の問いは本章の問いになる。 S の未知のあいだでの労力の分布と、そのうちどれが追究に値するかの判断とを、人工の構成の内側に置けるか。置くと何が変わるか。第9章が議論の足場として用いた例でいえば、百のインスタンスを生成し、それらを異なる路線に割り当て、別のインスタンスが走らせる点検に応じて路線のあいだで計算資源を振り向け直すシステムは、この意味で配分を行っている。そして、そうしたシステムが探究をうまく配分すると言う構築者は、この関係についての主張をしている（第9章第1節・第6節）。

手段が有限であれば、二つの問題を区別できる。第一は探索の問題である。 S が知りたいことが与えられたとき、見込みの不確かな路線のあいだで労力をどう分布させるか。第二は目的のものにかかわる。 S の未知のうち、どれがそもそも追究に値し、どの順序で追究するのか。第一の問題の完全な解も、第二の問題には答えないまま残すだろう。二つの問題はいずれも、いま述べた有限性から生じる。有限性は、所与とされた目的へ向けた労力の分布と、目的のあいだの選択との両方を強いるからである。二つを分けておけば、探究をうまく配分するという主張について、それが二つの問題のどちらにかかわるのかを問うことができる。二つを分けても、探索が価値から自由になるわけではない。第2節で論じるように、探索が自らの空間をどう表し、どう順序づけるかは、それ自体、所与とされた目的のもとでの価値判断であり、第6節が取り上げる目的のあいだの選択とは別のものである。ただし、二つはきれいに分けられるわけではない。探索における成功とみなされるものそれ自体が問われるところでは、両者は交わる（第6節）。第2節から第5節は探索の問題を、第6節は第二の問題を取り上げ、第12章は誰がそれに答えてよいかを取り上げる。

探索の問題は、有限性が無知と出会うところで生じる。どの路線が成功するか分らないことが配分の問題を生むのは、すべての路線を、少ない費用で、安全に、やり直せる形で追究することができないときだけである。それができるなら、人工の構成はすべてを試せばよく、そのあいだで選ぶことはその意味の大半を失う。第4節は、この条件がどこで成り立ち、どこで弱まるかを問う。

探索の問題を難しくするのは、どの路線が成功するかを前もって知る者がいないことである。しかし、知らないということは少なくとも三つの異なる事態を含んでおり、それぞれが本章で検討する議論の一つの族に対応する。そのため、この三つを区別すれば、労力を複数の路線に広げることに価値がありうる理由が分かれ、以下の議論を、それが扱う無知の種類によって見分けることができる。ただし、理由はこれらに尽きない。探索ではなく批判のために保たれる多様性はさらに別の理由であり、第11章第3節がそれを取り上げる。また、目的の順序づけは、これらのどの意味においても、知らないことの問題ではない。

第一の種類では、各路線が成功する見込みは分かっており、分からないのは、どの路線が実際に成果をもたらすかである。フィリップ・キッチャー（Philip Kitcher）の認知的労働の分業のモデルはこの種類である。ある方法に割り当てられた労働者の数を所与として、その方法が成功する見込みは既知として扱われる。そして対立する方法の場合、共同体にとって最適な分布では、労働力が複数の方法に分けられうる。マルコ・ヴィオラ（Marco Viola）がこのモデルを紹介するところでは、方法が飽和に近づくと、労働者を加えても得られるものはしだいに小さくなるからである（Viola 2015, 5–6; Kitcher 2013, 202）。この場合、分業が価値をもつために何かを学ぶ必要はない。対立する理論の場合は異なり、次に見る第二の種類に近い。そこでキッチャーは、ヴィオラが引くところでは、二つの理論の一方が劣ることに誰もが合理的に同意しうる場合でさえ、「問題含みの考えを擁護し続ける頑固な少数派に感謝してよい」とする（Viola 2015, 6）。分業によって何かが学ばれうるのである。第二の種類では、見込みそのものが分かっておらず、路線を追究することによってしか学べない。ここで問題は、探索と活用のトレードオフと呼ばれる形をとる。悪く見える選択肢について学ぶことに費やされる労力は、最善に見える選択肢から差し引かれ、悪く見える選択肢から労力が差し引かれる分だけ、その長所は学ばれないままになる。このトレードオフは特定の問題設定に限られず、組織の学習（第5節）、子どもの発達（第4章第7節）、機械学習などで論じられてきた。ジンイー・ウー（Jingyi Wu）とケイリン・オコナー（Cailin O'Connor）が指摘するように、科学ではこのトレードオフは、時間、資金、エネルギーが限られているために、いっそう鋭くなる（Wu and O'Connor 2023; 受理稿, 4, 8）。第三の種類では、路線もその価値も前もって投影できない。ウーとオコナーは、自分たちが概観するこのトレードオフのモデル、いわゆるバンディット・モデルが、明確に定義されよく理解された選択肢の集合を前提し、新しい選択肢の生成を扱っていないことを指摘する（Wu and O'Connor 2023; 受理稿, 21）。そして第2節で検討するジェイミー・ショー（Jamie Shaw）は、これが研究のありふれた状況だと論じる。したがって探索と活用のトレードオフは、配分の問題の全体ではなく、その一部を名指している。第5章第4.2節は、単一の探究のなかで何を追究するかの判断を検討した。知らないことの三つの種類は、その同じ判断が多くの研究者のあいだに分配されたときにかかわる。

どの労力の広がりがいいかは、探究が何のためのものとされるかにも依存する。キッチャーによれば、規範や誘因が集散的な探究の目的にどう効くかを理解することは、「それらの目的が何であるかについての最初の説明を前提する」のであり、そうした説明が与えられてはじめて、集散的な研究を組織する最適な仕方を探ることができる（Kitcher 2013, 206）。以下で検討する議論は、それぞれ、そうした何らかの目的を前提している。キッチャーとマイケル・ストレヴンズ（Michael Strevens）のモデルでは、分布は、前もって与えられた路線のうえで、前もって定められた目標への期待寄与によって測られる。第4節のネットワーク・モデルでは、前もって定められた選択肢のうち、期待される成績がよりよいものを共同体が好むようになるかどうかによって測られ

る。その選択肢を特定することは、理論全体の真理を立証することと同じではない。ミリアム・ソロモン (Miriam Solomon) の説では、ある領域で得られる経験的成功のそれぞれが、一つの理論がそれらをすべてもつに至るまで代表され続けるかどうか、すなわちその成功をもつ理論を誰かが追究し続けているかどうかによって測られる。他の目的は、何がよい配分とみなされるかを変える。以下に挙げるのはそうした目的の例であり、いずれも、特定の議論やモデル、あるいはそもそも研究がなぜ追究されるのかについての立場によって前提されているものである (第9章第3節)。ドナルド・ギリーズ (Donald Gillies) は、研究はまだ知られていないことを探るので、どの研究計画が成功するかを前もって誰も言えないと論じ、よい資金配分の制度は、あらゆる接近をできるかぎり支援するだろうとする (Gillies 2014, 7, 11)。用途を予見できない研究を擁護するヴァネヴァー・ブッシュ (Vannevar Bush) の議論は、医学とその基礎にある諸科学の「前線全体」を発展させることを求めた (Bush 1945, 9)。どちらも、可能性の空間を覆うという目的を前提している。ウーとオコナーは、自分たちが概観するバンディット・モデルが新しい選択肢の生成を扱わないことに触れて、「新しく予期しない可能性」の探索をどう促すかを問う (Wu and O'Connor 2023; 受理稿, 21)。この問いは、まだ視野にない選択肢を生み出すという目的を前提している。第4節で取り上げる、利得が変わりうる場合の結果は、条件が変わったときに応答する能力を保つという目的を前提している。目的そのものが動くとする立場、たとえばマックス・ウェーバー (Max Weber) の立場やデューイ的な見方 (第9章第3節) では、変わるものに目的が含まれうる。ブッシュの報告書はまた、支援によって便益が公衆に届くまでの時間を短縮できる特定の問題については、協調した取り組みも求めていた (9-10)。これは、*S* の特定の成員に差し迫る問いに早く答えるという目的を前提している。これらの目的のもとでは、同じ分布が異なって判断されうる。たとえばウーとオコナーは、資金提供者は、まずまずの理論へのすばい収束が最善であるところよりも、正確さがきわめて重要なところで、より多くの探索を促すのがよいかもしれないと示唆する (受理稿, 12)。まだ探索されていないものに労力を向けるといった発見的方法も、同じように目的に依存する。作業的特徴づけはそれを含意しない。*S* にとっての新しさは、どの未知も候補にするが、未探索のものを優遇はしないからである (第1章第8節)。その発見的方法は、網羅や、誰も予期しないものの発見のような目的のもとでは理由をもつが、差し迫った問いに早く答えるといった目的のもとでは、より小さな理由しかもたない。作業的特徴づけは探究にひとつの大域的な目的を確立しないので (第1章第1節)、本章は、各議論について、その結論がどの目的を前提しているかを述べる。

作業的特徴づけは、異なる達成をどう比べるかも決めていない。達成が異なれば、支持される配分も異なりうる。たとえば、別の規則性を発見することと、既知の規則性を説明することは、同じ有限の手段を争うが、共通の尺度で比べられるとは限らない。相対的な重要性は、どの価値と目的の説をとるかによって決まる。認識的達成という語は、総合得点を与えず、能力形成へのあらゆる投資をそれ自体の達成にすることもない。

議論そのものは、主に一つの系統から来る。キッチャーのものに始まる認知的分業の形式モデルと、それに先立つ中央の指揮をめぐる論争である。この系統を本章の問いの検討に用いるのには、特定の理由がある。そのモデルは前提を明示的に述べるので、前提を一つずつ検討し、第9章第6節の仕分けによって分類できる。その前提は、人間という実現者についての前提に依るのか、現在の科学のシステムの組織についての前提に依るのか、何がそれを実現するかにかかわらず成り立つ探究の条件に依るのか、それとも人工の構成の外の当事者 (*S* はその一つ) との関係に依るのか。研究共同体についての経験的研究は、以下では、処方としてではなく、前提が成り立つかどうかについての証拠として入る。第2節は中心とその統制についての前提を、第3節は配分が研究者の動機を経由するという前提を検討する。第4節はどの路線をどれだけ長く生かし続けるかを、第5節は人間の配分が誰も決めないまま果たしていた働きを、第6節は誰の未知が追究されるかという問いを検討する。検討を通じて重要になるのは、一つの人工の構成が多くの路線を作り、保てることと、それらの存続を一つの基準で判断することとを分けることである (第2節・第4節)。第7節は、これらの検討をつないで、配分が人工の構成の内側に入るとき研究に何が起こりうるかを一つの像として描き、構成がうまく配分しているという主張が何を示す必要があるかを、その像から引き出す。

2. 哲学者君主の二つの力

人間の研究共同体は、配分を行うのにひとつの決定を必要としない。研究者たちが、功績の承認、キャリア、資金、同僚の選択の圧力のもとで問題と方法を選び、その結果として生じる労力の分布が、共同体が実際に追求するものになる。人工システムからなる研究構成は、この関係の実現のされ方を根本から変えうる。人工の構成のなかには、研究者を生成し、複製し、割り当て、引き揚げることでできるものもありうる。これが探索の問題について提起する第一の問いは、配分が中心から割り当てられようようになったとき、あるいは人工の構成の内側で別の仕方で形づくられようようになったとき (第3節)、人間の探究の分散した組織を認識的に価値あるものにしていった何かが、なお残るかどうかである。ここで分散した配分を擁護した議論を見るのは、それらが、その分布が何のために価値があったのかを述べており、その前提を一つずつ検討できるからである。

分散した配分の古典的な擁護は、マイケル・ポランニー (Michael Polanyi) のものである。ポランニーは科学を、大きなジグソーパズルを組み立てる助け手たちの集団にたとえた。各人は互いに見えるところで作業し、他の誰かがはめた一片にそのつど応答する。こうした助け手たちを単一の権威のもとに組織しようとするれば、各人の独立した創意が失われ、共同の実効性も、中央で指揮する一人の人間の実効性にまで落ちるとポランニーは論じた (Polanyi 1962; 2000年の再録版, 2-3)。同じ論文はまた、各助け手にすべての片の複製を別々に与え、それぞれの結果をあとで合わせるやり方も検討している。それは片を分け与えるよりは少しよいが、それでも一団は、最もよくできる一人の働きをあまり上回らないだろうという (Polanyi 1962; 2000年の再録版, 2)。その説では、ジグソーの利は助け手の数からではなく、互いに見えるところで作業することから来る。この論証には複数の前提がある。ポランニーはこれを、最終的な結果が未知であり、そのため共同の課題は段階的にしか進めず、各段階はそれを進めるのに最も有能な者によって最もよく進められる、ということに基づかせている (Polanyi 1962; 2000年の再録版, 3)。しかし前提のひとつは、人間という実現者にかかわる。すなわち、中心はひとりの人間であり、ひとりの人間が注意を向けうる能力しかもたない、という前提である。人工の配分者は、このように限られている必要がない。多くの路線に同時に注意を向け、すべての研究者が観察したことを集めることができる。ただし、観察を中心に集めることは、ポランニーの助け手たちがしていることではない。助け手たちは、他の誰かがはめた一片にそれぞれ応答する。人工の構成のインスタンスが互いの結果を見てそれに応じるのか、結果を集める配分者に報告するだけなのかは、その組織についての別の問いである (第4節も参照)。分散した配分の擁護がこの前提だけに依っていたなら、人工の配分者はそれを解消するだろう。のちにキッチャーが立てた思考実験は、この擁護がこの前提だけに依っているのではないことを示唆する。

キッチャーは、ヴィオラが引くところでは「研究の進路を指図する特権をもつ」哲学者君主を想像した。この君主は、科学共同体がその認識的目標に到達することを望み、ショーの描写では、競合する路線の確率関数を見通す知識をもつ。本章では、この知識を、理論の客観的な長所を見抜く君主の眼と呼ぶ

(Viola 2015, 5; Shaw 2018, 155)。キッチャーは、共同体の最適という問いを立てるためにこの君主を導入し、ヴィオラによれば、そのうえでその観点を有用な虚構として脇に置いた (Viola 2015, 5, 9)。虚構を用いないキッチャーののちの述べ方では、各研究者が方法を正しく評価して自らの成功の見込みを最大にしようとすれば、全員がより有望な方法を探る。そのうちの何人かが「共同体のために」利他的にもう一方の方法を採れば、共同体はよりよい結果を得るだろうが、それはかなりの犠牲を求めるかもしれない (Kitcher 2013, 202)。ここでこの思考実験を用いる理由は、より狭い。この思考実験は、人間の制度が部分的にしか結びつけてこなかった二つの力、すなわち研究者がどこで働くかに対する統制と、理論の客観的な長所を見抜く眼とを分け、人工の構成がそのどちらを変えるのかを問えるようにするのである。そうした人工の構成が一方の力を虚構らしく見えなくすることは、思考実験をあらためて検討する理由であって、君主が実現されたことの証拠ではない。

キッチャーは、この君主でさえ、すべての研究者を最もよく支持された選択肢に割り当てはしないだろうと論じる。ショーとヴィオラが伝えるところでは、競合する選択肢のあいだの証拠の釣り合いが微妙なとき、方法が追加の労働者に素早く反応するとき、あるいは労働力が大きいとき、共同体の目標に最もよく資する分布は労働力を分割する。たとえ各研究者が個別に判断すれば同じ選択肢を選ぶとしても、である (Shaw 2018, 153–154; Viola 2015, 5–6; Kitcher 2013, 202)。個人の合理性が生み出す分布と、共同体が選ぶであろう分布は、食い違いうる。キッチャーの分析では、その理由は研究者の欠陥にあるのではない。第1節で触れた問題の二つの特徴にある。どの路線が成功するかを前もって知る者はおらず、ひとつの路線への労力の追加は収穫逨減をもたらす。最適は、共同体の目標が何とされるのであれ、それに相対的に定義される (第1節)。

科学にそのような君主はいない。キッチャー自身、ジュリアン・ライス (Julian Reiss) との共著で、大きな報酬がいくつもの集団を同じ成果への競争に引き込むところでは、「ある程度の計画」が分業と有益な協働を促すだろうと述べている (Reiss and Kitcher 2009, 273)。第9章第5節は、統制という語がひとまとめにしようものとして、決める形式的な権利、それを行使する実効的な能力、それをどう使うかにかかわる知識、そして選択肢のあいだで決めることなしに選択肢を形づくる影響を区別した。研究者を生成し、複製し、割り当て、引き揚げることのできる人工の構成は、直接の割り当てと、路線のあいだの費用のかからない移動とに、人間の取り決めがこれまで近づいたよりも近づきうる。どこまで近づくかは、その能力によっては決まらない。割り当てる形式上の力に実効的な統制が伴うかどうかは、三つのことに依存する。一つは、割り当てられた研究者が、配分者の意図する意味で路線を追究するかどうかである。二つ目は、配分者がそのあいだで割り当てる路線が、探究の前に与えられているのか、探究のなかで形成され分割されるのか (第4章第4節・第6節、下の第4節) である。三つ目は、路線を追究するための資源が、配分者の割り当てうるものかどうか (第12章第6.4節) である。

統制に近づくところでも、統制は、哲学者君主に与えられていた、理論の客観的な長所を見抜く眼を供給しない。ストレヴンズは、この点の一部を精確にする。ストレヴンズのモデルでは、研究計画のあいだの最適な配分を計算する中央計画者は、各計画の成功関数の真の形を知っていると仮定されている。そしてその理想化のもとでは、計画の成功確率への限界的な寄与に比例して研究者に報いる方式が、計画者なしに同じ配分に到達する。成功関数が予見できない仕方では変わるなら、最適は手の届かないものになりうる。しかしストレヴンズによれば、これは自らの提案する報酬方式に特有のもろさではなく、「最も賢い中央計画者をも等しく妨げる」「普遍的な障害」である (Strevens 2003, §§3.3–3.4)。キッチャーはのちに、これに関連する点を認める。自然は予測できないので、重要性についての判断は、理想的な条件のもとで下されたとしても、次の世代の判断者によって改められる必要がしばしば生じる。ただしキッチャーは、先の見えない将来に備える計画がどれもたえず改められていくのと同じく、これをとくに困ったことではないとする (Kitcher 2013, 221)。キッチャーはまた、集合的な研究の最適な組織化を見いだすという哲学の課題に完全な一般性において取り組むのは、論理経験主義が求めた一般理論の教訓が示すように野心的にすぎ、研究の領域が異なればそれぞれ別の分析が必要になりそうだとする。そして、最適を求めるより、まずは集合的な探究がどうすればひどく誤りうるかを理解しようとするほうが有益かもしれないと述べる (Kitcher 2013, 206)。ウーとオコナーは、資金提供機関が共同体を、決定理論の勧める探索の水準へ動かしようとして論じる。そのうえで二人は、意思決定者が共同体にわたって労働を配分できるところでさえ、現実の科学的証拠の複雑さのために、労働をどう配分すべきかを知るのが難しくなりうるのと付け加える (Wu and O'Connor 2023; 受理稿, 11–12, 14)。北野宏明 (Kitano) は、大きな科学的発見を自律的に行う人工システムをつくる計画を提案するなかで、これに関連する点を、機械による探索のための理由として用いる。北野の説では、大きな成果につながるまでにさらにいくつかの発見が必要でありうるときには、発見の重要性を評価することは、不可能ではないにしても極めて難しい。この困難が、北野のいう地平の向こうの問題 (over-the-horizon problem) である。そのため、期待される重要性を追う探索は、期待される重要性の低い段階を経てしか到達できない重要な発見の手前で止まりうる。北野は、自らが価値主導と呼ぶ基準、すなわち発見が重要になるという見込みに基づく基準を課さずに、そうしたシステムが仮説を探索することを提案する。そして、利害関係者の価値に沿い、あらかじめ定められた問題に取り組む問題解決者としての役割と、人間の研究者が到達しない仮説空間を探る探索者としての役割とを区別し、人間の専門家が指令と監視の役割に就きうるとし、その役割は誤用を防ぐうえで重要だとする (Kitano 2021)。したがってこの計画では、眼の欠如に応じるのは、ポランニーの説のように各段階をそれを進めるのに最も有能な者に任せるのではなく、期待される重要性によっては誰も選ばないであろう路線にまで探索を広げることである。ただし北野自身が記すように、そうしたシステムも資源の限りを免れないので、仮説空間をどう標本抽出するかは、その計画の内部でも未解決の研究課題として残る。そこで、路線のあいだの何らかの順序づけが再び現れる。

しかし、順序づけが再び現れる理由は、資源の限りだけではない。どの探索も、少なくとも暗黙のうちに、探索する空間をどう表し、そのどこをどの順で探すかを定めている。探索がこうに定めるものを、その探索の発見的方法と呼ぶことができる。どこをどの順で探すかを選ぶことは、複数の可能性のなかからいくつかを指定することであり、それがどのように選ばれるにせよ、何らかの選択の基準による。その基準は、何を先に取り上げ、何を重く見るかについての判断であり、広い意味での価値判断である。その意味では、選択とは、複数の選択肢のうちどれを、どこから取り上げるかについての価値判断であると言ってもよい。したがって、優先づけを免れた探索はない。目的を追わない探索や、期待される重要性によらない探索の提案も、その例になる。北野自身、計画の最初の段階では、分子機構という特定の形の仮説に焦点を合わせることを提案している (Kitano 2021)。目的を追うことをやめ、新奇性だけを追う探索を提案したジョエル・リーマン (Joel Lehman) とケネス・スタンリー (Kenneth Stanley) も、自らの方法について、新奇性の尺度は「領域に合わせなければならない」こと、そして NEAT のように解を次第に複雑にしていくアルゴリズムと組み合わせると、この方法は探索を単純なものから複雑なものへと順序づけることを述べている (Lehman and Stanley 2011, §§3–4, 9.1)。また第5章第3節で見たように、理論に依存しないとされる探索でさえ、探索される空間の事前の特徴づけに頼っている。研究についていえば、ジュリアン・ライス (Julian Reiss) とヤン・シュブレングァー (Jan Sprenger) は、研究の問題の選択がしばしば科学者個人、資金提供者、社会全体の利害に影響されることは、ほぼ普遍的に受け入れられていると報告して

いる (Reiss and Sprenger 2020, §3.1)。また、ウェーバーの立場（第9章第3節）では、実在のどの部分が研究の対象になるかは価値理念によって定まる。同じ理由で、人間が用いてきた発見的方法にできるかぎり頼らずに探索するという方針も、価値から自由ではない。その方針自体が、探索がどこへどの順で向かうかを定める一つの優先づけだからである。この意味で、完全に価値から自由な探索はありえない。北野のような提案が取り除くのは、前もって見込まれる重要性という一つの基準（北野が価値主導と呼ぶもの）であって、広い意味での価値判断そのものではない。このことから引き出せるのは、どの探索も価値判断を含むということだけではない。広い意味での価値判断であるという点では同じでも、何を重く見るかは探索によって異なる。たとえば、無作為な順序を選ぶことは、どの候補にも先に取り上げられる等しい見込みを与えること、言い換えれば、どの候補が重要かを前提しないことを重く見る。同じように、計算の都合で順序を決めることは手軽さを、期待される重要性で順序を決めることは見込まれる成果を重んじる。そして、期待される重要性を使わない探索も、空間の表し方と順序によって、どの可能性に到達でき、どれに先に到達するかを決める。第6節で見るキッチャーの説のように、どの路線が追究されたかが後に何が重要に見えるかを形づくとすれば、この到達可能性と順序の選択は、後の重要性の判断を制約する。その場合、そうした提案は、前もって見込まれる重要性によって探索を狭めることを避けるが、その代わりに採る順序が、何が後に重要と見えうかの範囲を定めることになる。

ジェイミー・ショーは、ファイヤーabend（Feyerabend）の多元主義から研究の組織化についての見方を展開し、上でストレヴンズとキッチャーに見た、眼が欠けているという診断を、第1節で区別した知らないことの第三の種類、すなわち路線もその価値も前もって投影できない場合へと一般化する。ショーによれば、研究の目標が何を研究するかを決めるとするキッチャーの秩序ある科学の理想と、キッチャー、ストレヴンズ、ワイスバーグとマルドゥーン（Weisberg and Muldoon）の配分モデルは、明示されない一つの前提を共有している。キッチャーとストレヴンズのモデルは、研究計画が成功する見込みが分かっていると想定しながら、それがどうして分かるのかを述べない。ワイスバーグとマルドゥーンのモデルでは、研究が進むにつれて見込みが変わるが、ショーの読みでは、それでも「将来の重要性を予測することを求める」。どの接近が重要かを研究者が前もって知らないこのモデルでは、前提は、接近のうえにあらかじめ固定された重要性の分布、すなわち研究者が知らなくても各接近の重要性は前もって定まっているという設定と、全員が先行する成果のうえに積み上げる通常科学を行う研究者たちのなかにある (Shaw 2018, 153–160; Weisberg and Muldoon 2009, 229, 231, 233)。ショーは、どの考えもその帰結のすべてにわたって吟味されることはなく、放棄されたり対抗理論に追い越されたりした理論も復活しうる、というファイヤーabendの観察に拠る。そのため、ある路線がこの先も衰え続けるか進歩するかは、現在の状態からは読み取れず、ある考えが何らかの長所を示すにも、まずそれが追究されなければならない (Shaw 2018, 69, 86–87)。ショーの紹介によれば、ファイヤーabendは、理論の行く末についての長期的な投影は不可能だとし、衰えつつある路線を進歩しつつある路線より危険だとみなすことは、ヒュームの問題の一つの形に突き当たるとした。ショーは、通常科学の内部では研究の価値についてそれなりに正確な投影ができることを認めるが、そうした投影は「その本性上、短期のもの」であり、長期的には、増殖したどの代替案も、どうなるかを誰も知らないという意味で、等しく危険なところから出発するとする。この説では、投影は確立した研究計画の内部での資金配分を導きうるが、その傍らで増殖する代替案にどう資金を配分するかは教えず、両者のあいだで資金をどう分けるかは、結果がどれほど急いで必要とされるかによる (Shaw 2018, 123, 136, 161–164)。したがって、研究の帰結は予見できないという前提は、説によって異なる結論を支える。ポランニーにとっては、各段階をそれを進めるのに最も有能な者に任せよという結論を支え、ショーにとっては、投影には確立した計画の内部での資金配分だけを導かせ、そこにどれだけを振り向けるかは結果がどれほど急いで必要とされるかによって決めよという結論を支える。長期的に予見できないことから、増殖した代替案をその見込みによって順位づけないことへの、ショーがファイヤーabendと共有する一步については、研究についてではないが、決定理論に立場がある。決定理論は、まだ考慮していない状態や帰結があると思意決定者が凝っている場合、すなわち無自覚への自覚 (awareness of unawareness) と呼ばれる場合に、どう推論しうるかを問うてきた。ケイティ・スティール (Katie Steele) とH・オッリ・ステファンソン (H. Orri Stefánsson) は、そうした自覚は、意思決定者が考慮している状態についての不完全な把握と、「種類ではなく程度においてしか」異ならないかもしれないと論じる。二人が紹介する立場のうち、サイモン・グラント (Simon Grant) とジョン・クイギン (John Quiggin) は、そうした判断は、過去に自覚が広がった場面からの帰納によって下されると示唆し、クロエ・ド・カンソン (Chloé de Canson) は、自覚していない状態や帰結があることに「少なくともいくらかの正の確率」を割り当てることを求める (Steele and Stefánsson 2025, §5.4)。これらの立場によれば、まだ予見されていないものも、自覚が広がった過去の場面からの帰納や、まだ考えられていないものにいくらかの確率を残しておくことによって、判断の対象になりうる。そうした判断が、ショーが割り当てる等しい危険ではなく、増殖した代替案のあいだで差のある評価の根拠になりうるかは、さらに別の問いであり、研究を論じていないこれらの立場は、それを決めない。

これらの議論に従えば、君主の統制に、人間の制度がこれまで近づいたよりも近づく人工の配分者は、それによって眼を獲得するわけではない。ショーが正しければ、その配分者にとって知識の不足が最も小さいのは確立した計画の内部であり、最も大きいのは計画のあいだの選択と、計画から外れる仕事の支援においてである。フリードリヒ・ハイエク (Friedrich Hayek) は、競争を、それがなければ知られないままであろう事実を発見する手続きとして論じ、似た前提から、そうした手続きをどう判定しうるかについての帰結を引き出した。ハイエクによれば、そうした手続きの長所は、それが関心の対象となる場合にこそ経験的に検証できない。ハイエクは、科学の発見の手続き一般もこの特徴を共有するとし、確立した科学の手続きの利点は「実際に他の手続きより良い結果をもたらしてきたからこそ」認められるのだと述べた。そして、市場の成果は、既知のどの方法でも達成できない理想的な基準と比べて「上から」判定されるのではなく、利用できる他の方法によって達成されうるものと比べて「下から」判定されるのだとした (Hayek [1968] 2002, 10, 16)。ハイエクは研究の労力の配分を論じていないが、その論点を市場からそれへ広げて当てはめると、人工の構成がうまく配分しているという主張は、構成にとって利用できる別の取り決めと比べて検査することはできても、哲学者君主が選ぶであろう分布と比べて検査することはできない。その比較には眼が必要になるからである。

そうした統制には、それに近づくところでは、思考実験が示さない代価がある。労働力の完全な統制とは、研究者が自らの路線を選ばないということである。第7章第4.3節の用語でいえば、すべての路線を割り当てる配分者は、研究者から問いの選択を取り上げ、外から供給する。統制が完全に実現されるほど、研究者は、同節が区別した意味で、問いの選択について自律的でなくなる。しかし、それだけで、配分者を含む人工の構成がただひとつの研究の行為者になるわけではない。同節が述べるように、外から問いを割り当てられた行為者も研究を遂行しうからである。研究者がそれぞれの路線で研究を遂行しているか、人工の構成全体がひとつの探究を遂行しているか、部分が互いに独立であるかは別々の問いであり、インスタンスを数えてもどれも決まらない。第7章第2節は、はじめの二つにかかわる境界を、起動、監視、方向転換、回復がどこで組織されているかによって定めた。境界は主張に相対的に引かれるので、研究者と人工の構成全体がともに研究の行為者であることもありうる。第9章第5節は、部分のあいだで分かれていれば部分を互いに独立にする

であろう権能を名指した。そうした組織も権能も部分のあいだで分かれていないなら、認知的分業はひとつの行為者の内部のポートフォリオであり、本節の問題は、第5章第4.2節がひとつの探究について述べた問題、すなわち追究に値するかどうかを真らしさと同一視せずにどう判断するか、という問題になる。研究者が割り当てられた路線で探究を組織していても、そうした権能をまったく保持していないなら、配分は、中心から独立していない複数の研究の行為者のあいだに仕事を分けることになる。研究者がそうした権能のいくつかを保持しているなら、統制は部分的にとどまり、彼らの選択を何が動かすのかという問いが戻ってくる。

3. 動機なしに舵を取る：信号

キッチャーの君主が課すであろう分業は、人間の共同体では、それが実現されてきた限りで、部分的には研究者の動機を通じて実現されてきた。したがって分散した配分の第二の前提は、配分がそうした動機を経由するということである。配分が人工の構成の内側に入ることは、それが中央から直接に割り当てられることを意味しない。インスタンスが自ら路線を選び、路線を移っていく人工の構成もありうる。ただしその場合に配分が経由するのは、人間の研究者の動機ではなく、機械の側の動機、あるいはそれに相当するものでありうる。それが功績の追求と同じ働きをするかどうかは、別の問いである（第4章第7節）。配分者が路線を直接に割り当てるところでは、この前提は崩れる。問いは、何がそれとともに失われ、何が残るかである。この問いの検討に用いる文献は、功績の追求のような動機がどのように労力の広がりを生み、あるいは妨げるかを問うてきたものである。キッチャーの功績のモデルとその後継、そして両者が廻る先取権の社会学である。人間の研究者のあいだで労力を動かしてきたもう一つの動機である好奇心は、第5節で取り上げる。

キッチャーのモデルは、成員が先取権や賞を求める共同体のほうが、そうした動機をもたない共同体よりもうまく労力を分布させうることを示す。功績をめぐる競争が、一部の研究者に、あまり好まれていない路線を追う個人的な理由を与えるからである。ロンジーノが科学的知識の社会的次元についての百科事典の項目でまとめるところでは、科学の進歩は、ある程度の不純な動機を「許容でき、それどころかそこから益を受ける」（Kitcher 2013, 202-203; Longino 2025, §5）。功績の追求は逆向きにも働く。安全な路線と、より危険だがよりよい路線とのあいだで研究者が繰り返し選ぶシミュレーションで、エリック・カマーフェルド（Erich Kummerfeld）とケヴィン・ゾルマン（Kevin Zollman）は、自分の貢献を気にかける研究者には、危険な探索を他者に任せる誘因があり、そのため放っておかれた共同体は、共同体にとって最善であるよりも探索が少なくなりうることを見いだす。研究者がそれぞれの路線をどれだけうまく扱えるかの違いは、この効果を和らげるか、なくしうる（Kummerfeld and Zollman 2016, §§2-3, 5）。経済学の側から科学を論じるパーサ・ダスグプタ（Partha Dasgupta）とポール・デイヴィッド（Paul David）は、先取権の規則のさらに別の効果を記述する。二人によれば、社会は、ある問題の解決に「誰が成功するかを気にかけない」。ところが先取権の規則は、誰がいつ解くかを各研究チームにとって重大なことにする。私的な順位づけと社会的な順位づけとのこの食い違いは、科学者たちが選ぶ研究のポートフォリオが非効率だろうと疑う根拠になる。その議論では、研究の成功による私的な利得が研究者の集まり全体にもたらされる便益に見合うものとされているかぎり、同一の研究計画のなかで期待される社会的便益の等しい個別の課題のあいだでは、先取権の規則は、競い合うチームを「過度に似た（すなわち正に相関した）研究計画」へ引き寄せる傾向をもつ。そしてその報酬の全か無かという性格は、チームを重複した競争へ引き込む。これが「社会全体として多様化した研究のポートフォリオが生まれることを妨げる」。相関は完全ではないと二人は付け加え、この議論を、証明としてではなく、証明の筋道を示唆するものとして示している（Dasgupta and David 1994, 506-507 and n. 56）。

路線を直接に割り当てる配分者は、配分を研究者の動機に経由させる必要がない。路線を直接に割り当てることで、配分者は、これらの分析が広がり源、探索不足の源、相関した重複の源として特定するものを、同時に取り除く。そして三つを同時に取り除くので、人間の共同体でどの効果が優っていたのか、したがって取り除くことが得か損かは、それによつては決まらない。また、功績を介さずに生じる路線のあいだの相関が、それによって取り除かれるわけでもない。第4節が、共通の基盤から複製された研究者について、また出自の異なる研究者についても述べるとおりである。そうした配分者のもとでは、どれだけ探索がなされるかは、研究者の利害ではなく、配分者が路線を割り当てるのに用いる規則の問題になる。その規則がどう定まったかにはかわからない。これに対して、インスタンスが自ら路線を選ぶ人工の構成では、どれだけ探索がなされるかは、何かがその選択を動かし、構成がその選択をどう組み合わせるかによる。そして、これらの分析が功績に帰した効果は、そこで別の形で戻ってくることも、まったく現れないこともありうる。どちらになるかは、インスタンスを動かすものが、これらの分析の特定した特徴をもつかどうかによる。たとえば、報酬が同じ路線を追う者のあいだで分けられ、混んでいない路線はど一人あたりが大きくなること（キッチャー）、他者がその探索から学ぶことなく自らの貢献だけに報いること（カマーフェルドとゾルマン）、最初であることだけに全か無かで報いること（ダスグプタとデイヴィッド）などである。

この除去がなしとげること、見かけより狭い。ストレヴンズの紹介によれば、マートン（Robert K. Merton）は、科学者が先取権を気にかけることは、その心理的な性向によるというより、科学を律する強い規範の反映だと論じた（Strevens 2003, §2.1）。その説では、先取権の規則は私的な動機を公的な目標に結びつけ、その結びつきを達成の信号から作る。ストレヴンズの紹介によれば、マートンはまた、数日の差で誰がすべての功績を得るかが決まるほど規則を字義どおりに適用することを、「有用性の限界をはるかに超えた機能不全の極端」とみなした。これに対してストレヴンズは、そこまで細かな区別にも働きがあると応じる。成功の見込みの高い研究プログラムを相対的に魅力あるものにするからである（Strevens 2003, §§2.1, 5.1）。路線を直接に割り当てる配分者は、人間の動機への依存を取り除く。しかし、それによって信号への依存を取り除くわけではない。第9章第4節がその一般的な理由を与えた。研究にとって成功とみなされるものは、それについての判断を下さなければならない時点では観察できないことが多い。そのため配分者は、どの路線が進展しているかを、結果、得点、他の路線がその産出物をどれだけ用いたかなどの量によって判断する。目標とその尺度を区別する説では、そうした量は目標を構成するのではなく、目標の代わりを務める。人間の配分も、研究者の動機だけでなく、制度化された信号を通じて動いてきた。研究成果の指標は、研究資金の配分に直接結びつけられてきた。たとえばオーストラリアでは、研究資金の一部が出版数にもとづく式で配分されるようになり、研究者はそれに反応した（Wilsdon et al. 2015, 81-82、第11章第8節）。そうした信号が資金の配分に直接用いられるところでは、影響を受ける研究者の動機を経由しなくても、どの路線が続くかがその信号によって決められうる。

カマーフェルドとゾルマンは、すべての研究者が同じ中央の機関に功績を認められることを求めるとき、その誘因はその機関が興味深いとみなすものをなぞると指摘する（Kummerfeld and Zollman 2016, §5）。単一の配分者の信号は、それが割り当てるすべての路線について、この位置に立つ。その信号

が、どの研究者を複製しどの研究者を引き揚げるかまで決めるところでは、信号への適応は研究者そのものを形づくる。明示的にモデル化してはいないが支持する解釈として、カマーフェルドとゾルマンは自分たちの均衡を進化の観点から読み、探索の少ない研究者は探索を他者に任せるので、個人としてより大きな貢献をすることを指摘する。他者がその研究者を模倣するか、その研究者がより多くの学生を育てるなら、共同体は、共同体にとって最善であるよりも探索の少ない方へ動きうる (Kummerfeld and Zollman 2016, §3)。ポール・スマルディーノ (Paul Smaldino) とリチャード・マケルリース (Richard McElreath) は、関連する力学を明示的にモデル化する。そのモデルでは、より多くの出版をもたらす方法をもつ研究室ほど、その方法を受け継ぐ「子孫」を多く残し、出版数だけにもとづく選択は、用いられる方法の厳密さを劣化させる。二人によれば、この力学は誰の意識的な戦略も必要としない (Smaldino and McElreath 2016, abstract, §§4, 5.1)。後に、そのモデルの実装にコーディングの誤りが見つかった。そのため研究室の追試率はその適応度に影響しえず、二人が追試について報告した力学は、記述されたモデルに対応していなかった。訂正は、進化の段階の順序にわずかな逸脱があったことにも触れ、結論は影響を受けないと述べており、方法の劣化についての結果に影響を受けたものとしては挙げていない (Smaldino and McElreath 2023)。スマルディーノとマケルリースのモデルが選択の対象とするのは、探索の量ではなく方法の厳密さである。これらのモデルの構造が人工の構成にも当てはまるなら、自らの測られた産出が最も高い研究者を複製する配分者は、動機をまったく介さずに、両モデルが記述する方向へ選択するだろう。すなわち、その信号が、ある研究者の探索が他者の学ぶことにどれだけ寄与するかを評価に入れなにかぎり探索の少ない方へ、産出の背後にある方法の厳密さを捉えなにかぎり厳密さの低い方法へ、である。

ストレヴンズは、キッチャーの功績を分け合う仕組みを、研究計画が成功する見込みと、報酬を分け合う参加者の数とによって説明している (Strevens 2011, §2, manuscript p. 10)。サモラ・ボニーリャ (Zamora Bonilla) がキッチャーのモデルを紹介するところでは、功績だけを求める研究者たちは、受け入れられると見込む理論に集中するが、真理にもいくらかの重みを置くと、最適に近い労力の分布が得られる (Zamora Bonilla 2014, "One important topic"で始まる段落)。受け入れられる見込みを示す信号に従う配分者に当てはめると、それに対応する問いは、その信号と真理を何が結びつけるかである。認定がその信号を供給する場合、この問いは配分を批判に結びつける (第9章第2.4節)。

配分の側では、二つの問いが続き、本章はその両方に戻る。第一は打ち切りの規則である。どの信号に基づき、どの時間の幅で、路線は打ち切られるのか。時間の幅が短ければ、成功を長い遅れの後にしか記録しない信号は、より長い時間の幅なら残したはずの路線を打ち切る (第5節)。この問いには、ファイヤアーベントがラカトシュ (Lakatos) に向けたジレンマという先例がある (第6章第4.1節)。マスグレイヴ (Musgrave) とピグデン (Pigden) の報告では、ラカトシュの方法論を厳格に適用すれば、科学者は退行的なプログラム (新しい事実を予測せず、既知の事実に合わせてるために理論を作り直している研究プログラム) を捨てるよう、助成機関はそれへの資金を止めるよう促され、そのため方法論は、後に回復するかもしれないプログラムを断罪することになる。これに対して、ラカトシュがそのとき限定したように助言を伴わない評価に限定すれば、方法論は何をするかを誰にも告げず、「何でもかまわない (Anything goes)」とほとんど区別できない (Musgrave and Pigden 2026, §4)。この先例からは、ジレンマのもう一方の選択肢である助言を伴わない評価は、配分者には開かれていないことになる。路線を打ち切るときの信号と時間の幅が、すでに実行された助言だからである。配分者にとって、このジレンマは打ち切り規則の選択になる。そのとき、厳格な適用という選択肢が負う代価、すなわち後に回復するかもしれない路線を断罪することの重さは、打ち切りを取り消せるかどうかによる。閉じた路線を再開できるか、どんな証拠によつてか (第4節)、もっていたものをどれだけ保ったままか (記録はその一部しか保たない、第5節) である。第二は循環である。配分は何が産出されるかを決め、産出されたものが次の信号の測るものになるので、自らの信号に従う配分者は、その信号が後に報告するものを変える。信号そのものが時間とともにどう振る舞うか、すなわち何を測り、何に偏り、測られる者が信号に応じるときにどう変わるかは、配分ではなく批判と評価の性質であり、第11章第8節がそれを検討する。

第2節と第3節の背後にあるモデルは、意図的に理想化されている。高度に理想化された科学的相互作用のモデルが、共同体をどう組織すべきかについての提言の根拠になりうるかは争われており、そうしたモデルはまず経験的な対象に照らして妥当性を確かめられなければならないとする論者もいる。ドゥニャ・シェセルジャ (Dunja Šešelja) の概観は、それらのモデルのうちエージェント・ベース・モデル (個々の研究者をエージェントとして計算機上で動かすモデル) について、このことを報告している (Šešelja 2022, §4)。ここでのモデルの役割は、配分の規則を定めることではない。前提を分け、前提が崩れたときに何がその場所を占めるかを問うことである。したがって二つの節の結論は条件つきである。それは、ポランニー、キッチャー、ストレヴンズ、ショー、ウーとオコナー、ハイエクの議論と、優先づけを免れた探索はないという論点 (第2節)、そして功績のモデル (第3節) に依拠し、信号が残る一般的な理由については第9章第4節に依拠する。そしてそれは、それらのモデルとともに、路線がそれに照らして判断される、前もって定められた目標を前提している (第1節)。

これらの議論に従えば、探究の配分を引き受ける人工の構成は、哲学者君主の二つの力のうち一方、すなわち割り当ての統制に、人間の制度がこれまで近づいたよりも近づきうる。それは、その割り当てが直接であり、その研究者がそれを実行し、路線と資源を人工の構成が割り当てうる程度に応じである。その統制によって、人工の構成が他方の力を獲得することはない。また、前もって見込まれる重要性の基準を外しても、その探索が優先づけを免れるわけではなく、その配分についての主張は、人工の構成にとって利用できる取り決めと比べて検査することはできても、君主が選ぶであろう分布と比べて検査することはできない。人間の分散した配分が依拠していた前提が崩れるとき、それらの前提が対処していた問題の多くは、消えずに形を変える。中心はもはや、ひとりの人間の能力しかもたない必要はない。ただし、路線を追究することそのものを可能にする資源の集中として、中心がふたたび現れうる (第12章第6.4節)。配分はもはや人間の研究者の動機を経由する必要はないが、*S* の目標の代わりに務める信号によって、なお舵を取られる。労働力の統制に近づくところでは、その統制は研究者から問いの選択についての自律を取り上げる。その統制が、研究者の路線の組織と、研究者を互いに独立にする権能までも吸収するところでは、共同体はひとつの行為者になる。そして、配分を一つの人工の構成のもとに統合するだけでは、どの路線が成功するかについての無知は取り除かれず、労力はなおその無知のもとで分布させられる。したがって、そうした人工の構成が探究をうまく配分しているという主張は、統制から良い配分を推論するのではなく、その信号がどう目標に応え続けるか、部分を独立にする権能がどこで分かれているか、構成がその無知をどう表象するか、目標と証拠に照らして保つに値する、好まれていない路線をどう生かし続けるかを示す必要があるだろう。

4. どの路線を、どれだけ長く生かし続けるのか

第2節と第3節の結論は、一つの考慮を残す。検討した議論に従えば、労力は統制が取り除かない無知のもとで分布させられ、いま好ましくないように見える路線のいくつかは、生かし続ける必要があるかもしれない。この考慮は、さらに四つの問いを提起する。何を生かし続けるのか、どれだけ長くか、その必要がどこで効くのか、そしてそれらを生かし続ける多様性は、どこから来るのか、である。とくに第二の問いには、二つの系統が答える。科学共同体の形式モデルは、不一致が、共同体がよりよい選択肢に収束するのをいつ、どれだけのあいだ助けるかを問う。ミリアム・ソロモンの社会的経験主義は、異論と合意がそれぞれいつ適切かを、経験的成功の分布によって判断して問う。本節は、それぞれの答えのもとで人工システムによる研究にとって何が帰結するか、それぞれがどの目的を前提しているかを問う。そして、第9章第2.3節が、関係の分割が供給したかもしれない財の一つとして仮説の形で提案したもの、すなわち異論の当事者資格がどうなるかを問う。

一方の系統は、どれだけ長くかという問いに、特定の主張で答える。ケヴィン・ゾルマンのネットワーク・モデルでは、共同体は、各研究者が受け取る情報の量を制限するか、研究者に極端な初期信念を与えることで、分業を維持できる。両方がそろうと、初期の多様性はけっして放棄されず、共同体はよりよい選択肢に収束しない。共同体に必要なのは多様性そのものではなく、一時的な多様性だ、とゾルマンは結論する。早すぎる閉鎖を防ぐのに十分なあいだ続き、その後は解消される不一致である (Zollman 2010, 1, §5)。この主張は、モデルの一つの族の答えであり、保つに値する多様性が一時的なものに限られるかどうかは、本節で後に示すように、それ自体が争われている。

これらのモデルの頑健性はその後検討されており、その検討は何が推論できるかを限る。ローゼンストック (Rosenstock)、ブルーナー (Bruner)、オコナーは、疎な通信の利点がパラメータ空間の小さな領域でしか現れないことを見いだした。研究者が比較的少なく、証拠が小さなまとまりで届き、競合する選択肢の成功率の差がわずかな場合である。一方、信念の何らかの一時的な多様性が必要だという主張は、三人が検討したモデル全体で残った。探究が難しいところでは、より頑固な、あるいはより探索的な研究者と、結論に必要な証拠の量についての基準とが、証拠を与えないことよりもうまく早すぎる収束を防ぎうる、と三人は付け加える (Rosenstock, Bruner, and O'Connor 2017, 235, 248, 251)。シェシェリヤの概観も、これと両立する判定に達しており、こうしたモデルは、少なくとも現在の形では、政策を定式化するためではなく、主としてありうる機構と候補となる説明を特定するのに役立つとする (Šešelja 2022, 3–4, 11–12)。研究者がどこで仕事をするかを選ぶ風景モデルの結果も、同様の理由で、すなわち実装と、ほとんど知られていない風景の形とに左右されるとして、争われてきた (Alexander, Himmelreich, and Thompson 2015, 425, 432–440, 448–451)。ヨハンナ・トーマ (Johanna Thoma) の風景モデルでは、他者が試した接近から離れていく研究者は、これまでに見つかった最良の接近を延ばす研究者よりも、平均してはるかに少なくしか発見しない。そのため、探索が続けられるには、発見以外の報酬が必要かもしれない (Thoma 2015, §§4.3–4.4, 5)。戦略を割り当てられた人工の研究者はこの必要をもたないだろうが、そのときそれは、探索を気質によって供給することもないだろう。この検討を経て残るものは、処方より狭い。信念の何らかの一時的な多様性は多くのモデルで必要とされるように見えること、それを生み出す特定の機構の利点はほろいこと、そしてモデルはありうる機構を特定するものであること、である。

これらの結果もまた、ある目的を前提している。ネットワーク・モデルでは、共同体が成功するのは、前もって定められた選択肢のうちよりよいものを好むようになるときである。多様性はそこでは、その収束への手段として価値をもち、けっして道を譲らない多様性は失敗とみなされる。定められた選択肢のあいだの収束で終わらない目的、たとえば条件が変わったときに応答する能力を保つことや、まだ集合のなかにない選択肢を生み出すことのもとでは (どの議論やモデルがそれぞれを前提しているかは第1節で示した)、道を譲らない多様性は、これらのモデルが記録しない理由をもちうる。そして、異論をもつ視点を育てることを求めるヘレン・ロンジーノ (Helen Longino) の説 (第11章第3節) と、クリステン・インテマン (Kristen Intemann) が再構成する立脚点理論 (standpoint theory。研究がどの社会的位置から出発するかが、得られる知識にかかわるとする立場の系譜。第6節) とでは、それは批判の条件として、さらなる理由をもつ。したがって、多様性が一時的であるべきだという主張は、モデルの構造だけでなく、モデルが組み込む目的にも相対的である。

これが残すのは、第三の問い、すなわちその必要がどこで効くかである。一時的な多様性を守る手段が主に探究が難しいときに重要なのだとすれば、その重要性は領域によって異なる。この必要はまた、第1節で述べた条件、すなわち路線をすべて試すことはできないということも前提している。ゾルマンのモデルは、この前提を別の側から位置づける。そのモデルが示す害は、ある路線についての証拠がその路線を追求することによって集められるからこそ生じるのであり、手に入る証拠が研究者自身の選択に依存しないなら生じえない (Zollman 2010, §3)。特定された主張についての証拠が手軽に得られ、豊富で、決定的なところでは、この前提は弱まり、労力を集中させることの費用は比較的小さい。形式的な点検や高速なシミュレーションは、固定された形式体系やモデルの内部で、そのような証拠を与える。モデルが標的を十全に表しているかは、別の証拠上の問いとして残る。証拠が遅く、高価で、小さな標本でしか得られないところ、たとえば臨床研究、野外観察、希少な装置に依存する実験では、この前提は最も強い。そして、これらはまさに、人工のシステムにとっても、自らの証拠を手軽に増やすことのできない領域である。モデルの構造が人工の構成にも当てはまるなら、配分を統合することの認識的な危険は、人工システムによる研究がより多くのデータを生み出すことで埋め合わせる余地の最も小さいところに集中する。

二つの限定が付く。第一に、形式的な検証は、検査の時点では費用が小さく決定的だが、主張をその時点まで運ぶことはそうであるとは限らない。たとえば、リキッド・テンソル実験 (Liquid Tensor Experiment、第1章第3節) では、中心的な定理の言明の形式化に約1か月、その5ページの証明の形式化にさらに約5か月かかった一方で、最終的な結果を導く5行の推論には、それが依拠するアーベル圏、導来関手、Ext群の理論をまず構築しなければならなかったために、さらに1年かかった (Buzzard 2022)。したがって、この前提が弱まるかどうかは、ある領域の主張が、そのような検査の適用できる形ですでに表現されているかどうかにかかっている。第二に、問題が別々に探索できる部分に分けられるかどうかは、問題の固定した性質であるとは限らない。Philosophy of Science への掲載が決まり、ここではプレプリントから引用するシミュレーション研究で、ファゼルプール (Fazelpour)、オブライエン (O'Brien)、ルービン (Rubin) は、そうした分解可能性は、標準化、手順、分業によって達成されたものであることが多いと論じる (Fazelpour, O'Brien, and Rubin 2026, 2–3, 25)。そうだとすれば、ある領域でこの前提が弱まるかどうかは、その領域の問題がどう分割されたかに部分的に依存する。そして問題を分割することは、それ自体、同じ無知のもとでなされる配分の決定である。この点からは、さらに次のことが考えられる。問題を分解し、課題を選ぶ人工の構成は、判定しやすい課題でこそ成功を示しやすい。その成功が次の配分や、どの研究者を複製し引き揚げるか (第3節) を正当化する

るなら、証拠を得るのに時間がかかる課題や、証拠が曖昧な課題は、比べられる候補の外に置かれやすくなる。そのとき人工の構成は、自らの高い成績を理由に、難しい研究をさらに見えにくくするかもしれない。これは、評価の基準が偏っているというだけのことではない。同じ評価の信号が将来の配分に戻るという条件のもとで、研究の範囲が成功そのものによって内側から狭まりうるということである（第6節も参照）。

第四の問いは、多様性がどこから来るかである。ゾルマンの機構が多様性を保つのは、「行為者が多様な事前確率をもつ限り」においてである (Zollman 2010, §5)。人間の共同体は、そうした違いを、異なる訓練、来歴、コミットメントから、誰かが設計しなくても受け取る。人工の研究者は、共通の基盤から複製されることもあれば、異なる基盤からつくられて並び立つこともあり、どちらの場合にも、それぞれの探究の来歴を通じて違ってくることもありうる。共通の基盤から複製された研究者は、はじめはそうした違いを受け取らないかもしれない。しかし、出自が異なることも、それを保証しない。以下の証拠は出自の違いにかかわるものであり、探究の来歴のなかで生じる違いにかかわるものではない。キム (Kim) らは、共通のベンチマークで評価された数百の言語モデルについて、モデルがしばしば同じ誤答に収束すること、提供元や基盤のアーキテクチャが同じだとこの相関が高まること、そして提供元や基盤のアーキテクチャが異なっているにもかかわらず、より正確なモデルほど誤りの相関が高いことを見いだした (Kim et al. 2025, abstract)。比べられたのはどれも大規模言語モデルであり、そこでのアーキテクチャの違いは一つのクラスのモデルの中での変種である。また、これは固定された問いへの答えの誤りであって、研究の路線のあいだの選択ではない。それでも、このクラスのシステムについては、出自の多様性が事前の信念の多様性を供給すると想定することへの注意にはなる。まったく異なる仕方で作られたシステムが同じように誤りを共有するかどうかは、この知見が扱う問いではない。プレプリントとして公表された二つの研究は、路線を追求する際の選択にまでは至らないとしても、路線のあいだの選択により近づいている。ビシュト (Bisht) らは、四つの提供元の十二のモデルに、材料科学と生物学の公刊論文を拡張する新しい仮説を提案させた。この研究の埋め込みによる尺度では、提供元の異なるモデルの仮説どうしは、一つのモデルから繰り返し得た標本どうしとほぼ同じくらい似ており（材料科学の論文で0.64対0.72）、一つのモデルではなく十のモデルから十の仮説を引いても、実質的に異なる仮説の数は2.9から3.6に増えたただだった (Bisht et al. 2026, §5 and app. A)。バオ (Bao) らは、最近のプレプリントの著者に、自らの論文について言語モデルが生成した仮説を評価してもらった。その結果、推論を行わないモデルは考えの空間の狭い領域に集まり、推論を行うモデルはより広く散らばること、同じ企業のモデルがより似た考えを生むという証拠は見られないこと、そしてどのモデルも、関係や効果がないとする帰無仮説を人間の科学者より少なくしか提案しないことを見いだした (Bao et al. 2026, 1, 3)。二つの研究は異なる尺度を用いており、どちらも、研究者が時間をかけて追求する路線ではなく、与えられた論文について提案される仮説を扱っている。その限りで、これらの研究は、固定された問いへの答えについての注意を、仮説の提案にまで広げる。人工の研究者が多様な事前の信念をもたないなら、情報を制限しても、もともとなかった多様性は保てない。

中心に置かれた手立ては、いくらかの多様性を加えうる。たとえば、ショーが論じる、研究費の一部を籤で配分する提案が行うように、閾値を越えた路線のあいだで籤によって割り当てること (Shaw 2018, 172–173)、好まれていない路線のための予算を保護すること、各研究者の方策に探索の率を組み込むこと、あるいは研究者に、異なる仕方で初期化された表象や、路線を乗り換える前のある程度の慣性を与えることである (Šešelja 2022, 4; Rosenstock, Bruner, and O'Connor 2017, 251)。ソロモンは、この点で多様性は鈍い認識的道具であり、適切な多様化の仕方はたいてい一つではないと論じる (Solomon 2006b, abstract)。上で見たゾルマンのネットワーク・モデルは、その一例を与える。ここでは、各研究者が受け取る情報を限ることによっても、研究者に極端な初期信念を与えることによっても分業が保たれ、後者は前者と同じ効果をもつ (Zollman 2010, 1, 3–4)。そうであれば、欠けている分布を前もって特定できる限り、それを補う仕方は一つではない。こうした手立てがなしとげるものには、二つの限界がある。第一に、多様性は程度の問題であって、人工の構成がもつかもたないかの性質ではなく、またどの中心が含むものにも限られない。人工の構成が内部でどのような広がりを生み出すにせよ、その外でまったく異なる路線を追う者たちが、全体の多様性に加わる。第二に、手立てが加える多様さが、ある立場の求める多様さであるかどうかは、その立場に依存する。本節のモデルが考える労力の広がりについては、ソロモンの議論とゾルマンのモデルは、しばしばそうでありうることを示唆する。ロンジーノの説と立脚点理論が批判のために重んじる多様性、すなわち誰が探究に加わるかに依存する多様性については、それはより不確かである（第11章第3節、下の第6節）。どの手立てが、実際にそれが目指す多様性を生み出すかは、人工の研究者については、まだほとんど試されていないように見える。

これらのモデルのいくつかは、評価に誤りがなくとも前提している。キッチャーのモデルでは、答えはひとたび届けば正しいか誤っているかが見て分かり、ワイスバグとマルドゥーンの風景では、訪問はつねに接近の真の重要性を明らかにする (Zollman 2010, n. 3; Weisberg and Muldoon 2009, 232–233)。したがって次の問題は、これらのモデルが扱う範囲の外にある。

統合だけでは供給されないのは、多様性を終わらせる基準についての独立した点検である。一時的な多様性は、ある路線のほうがかうまくいったという判断によって終わる。配分者自身の信号ですべての路線を判断する、中央で配分される人工の構成では、その信号の誤りによって、それぞれ異なる路線の打ち切り判断が相関しうる。したがって本節の議論によれば、中央化と引き換えになるのは、多様性そのものというより、収束を支配する基準についての誤りの独立性である。批判の価値をそれが検出する誤りに置く見方のもとで、そうした独立性が何からなるかは、第11章第2節が取り上げる。このことから、多様性を評価するときには、候補がどれだけ異なるかや候補の数だけでなく、それらの継続の判断がどの誤りを共有するかも問える。異なる表象を生成できることと、異なる証拠によって路線の存続を判断できることは別である。前者だけを備えた人工の構成は、多様な候補を生みながら、同じ誤りによってそれらを同時に捨てる。

ソロモンの社会的経験主義は、どれだけ長くかという問いに別の答えを与え、それとともに第一の問い、すなわち何を生かし続けるのかにも答える。それは異論を原型とし、合意をその極限の場合として扱う。ロンジーノがまとめるところでは、この説において共同体が合理的であるのは、それが受け入れる理論が、すべての経験的成功をもつ理論か、固有の経験的成功をもつ理論であるときであり、固有の経験的成功をもつ理論を捨てることは集合的に非合理である。したがって、異論が適切なのは、対立する理論のそれぞれが固有の経験的成功をもつときであり、合意が適切なのは、一つの理論がそれらをすべてもつときである。異論が保つに値するのは、異論をもつ立場のそれぞれが固有の経験的成功をもたらすところであり、ロンジーノが紹介するソロモンの後の議論の拡張によれば、異なるモデルが、合意のなかで失われやすい特定の洞察やデータを担うところでもある (Longino 2025, §5)。同じ規範に従えば、対立する理論が、合意された理論のもたない経験的成功を得たなら、合意は考え直されなければならないことになる。ソロモン自身が、この規範に届かなかった合意の例として挙げるのは、DNA を「マスター分子」とみなす見方について1950年代にほぼ成立していた合意である。そのもとで、パーバ

ラ・マクリントック (Barbara McClintock) やルース・セイガー (Ruth Sager) らの仕事は受け入れられないか、遅れてしか受け入れられなかった (Solomon 2001b, 72)。ソロモンの説では、それぞれの成功が代表され続けたのは、ひとつには、科学者が自らのデータや関心を他の者より重く見積もったからである。ソロモンが検討する地質学の論争では、地質学者は自分自身のデータによって大陸移動を確信し、自分の結果が不利に見えるときにはそれに抵抗した。そうした影響が、対立する理論のあいだに研究の労力を分配した (Solomon 2001b, 69–72)。ソロモンの説が前提する目的は、ゾルマンのものとは異なる。定められた選択肢のうちよりよいものへの収束ではなく、ある領域が与える経験的成功のそれぞれが、一つの理論がそれらをすべてもつに至るまで代表され続けることである。

本章の用語では、自らのデータをこのように重く見積もることは、第2節が切り分けた種類の、人間という実現者についての前提である。ただしそれが配分に効くのは、これまでの研究の組織のように、データがまずそれを生み出した者の手元にあり、その者の重みづけが路線を続けるかどうかの決定に届くところである。データを共有することは、それだけではこの重みづけをなくさない。データを誰もが見られても、生み出した経験、出所についての知識、解釈によって、重みづけは異なりうる。人工の構成の内側でその役割を果たすものが残るかどうかは、観測がどのように研究者たちを経由し、そこで見いだされたものがどう組み合わせられるかによる。自らの記録を保ち、それに基づいて動く機械の研究者は、ソロモンの科学者たちと同等の仕方では自らのデータを重く見積もりうる。これに対して、すべての研究者の観測を持ち寄り、すべての路線を一つの信号で判断する配分者は、この前提を取り除く。取り除くのは、観測を持ち寄ることそのものではなく、すべての路線の継続を一つの信号で判断することである。観測を持ち寄りながら、ある路線の固有の成功を理由にその継続を求める判断が配分に届く仕組みを残すこともできる。人工の構成についても、誰がデータを見られるかと、誰の判断が路線を存続させられるかとは、別に問える。ソロモンの説では、そのとき配分者はまた、収束が固有の成功をもつ路線を捨てないことを別の手段で示し、閉じた路線をどんな証拠によって再開するかを述べなければならないだろう。ゾルマンの説では、対応する要求はより狭い。収束が、証拠がそれを正当化するより前に来ないことである。したがって二つの説は、多様性がどれだけ長く保たれるか、何がそれを終わらせるかについて異なる。しかしどちらにおいても、それを終わらせる判断には、統合だけでは供給されない証拠的な支持が必要である。統合された配分者も、自らの組織の内部に、その判断について情報を与える点検を維持しうる (第11章第1節)。

その判断はまた、異論の当事者資格についての仮説が配分に当てはまる場所でもある。配分において、異論は、支配的な判断なら閉じるであろう路線を続けること、あるいは閉じられた路線の再開を迫ることという形をとる。人間の研究では、そうした異論を唱える者は、その路線を好ましくないと判断した者とは別の当事者であることが多かったかもしれない。たとえば、研究室、実績、キャリアをもつ研究者、独自の基準をもつ資金提供者、別の分野の共同体などである。この分割が供給したもののうちに、二つのものを分けることができる。一つは認識上の働きである。ソロモンの説では、異論は、合意なら失うであろう成功を代表させ続ける。もう一つは、第9章第2節の意味での当事者資格である。路線を閉じるであろう決定との関係で保持される、その閉鎖に異議を唱え、その放棄を拒み、あるいはその再開を提案する資格である。本節の議論によれば、配分が統合されるとき、この二つは分かれる。異論の認識上の働きは、人工の構成の内側で果たされうる。固有の成功をもつ路線のための保護された予算や再開の規則のように、誰かが定める取り決めによって、あるいは、第9章第5節が名指した権能、たとえば資源を拒むこと、証拠を別に保持すること、記録を携えて離脱することを、人工の構成の部分のあいだに分けることによってである。それらの権能は当事者資格ではない (第9章第5節)。当事者資格が人工の構成の内側で保持されるのは、採る説のもとでそれをもつ部分によってだけである。たとえば、それ自体として当事者である部分 (第12章第5節) や、結果に基づく根拠のもとでは、確かめられたよい結果をもつ部分 (第12章第3.1節) がそれにあたりうる。そうした部分がなければ、当事者資格は、保持されるとしても、人工の構成の外で他の路線を追う当事者や、閉じられた路線の再開を迫りうる当事者によって保持される。これは、全体の多様性がどの中心の外にも部分的にあるということの、もう一つの側面である。その当事者資格が認識的に重要なのか、それとも路線の閉鎖が正統かどうかにとってだけ重要なのかは、採る説に依存する。ソロモンの説では、認識的に重要なのは、手段が何であれ、成功が代表され続けることである。ロンジーノの説のように、異論の価値を批判者の知的権威に結びつける説では、この問いは批判に属し、第11章第3節がそれを取り上げる。

強化学習の理論には、単一の学習者、すなわち環境のなかで行動を選び、それがもたらす報酬を観測しながら、その後の選び方を改めていく主体についての結果がある。それらは、探索を終わらせる判断に、別の方向からかわる。ただし本章に当てはまるのは、第2節の判定で人工の構成がひとつの行為者であり、その配分者が一つの事前分布のもとで、そうした一つの学習者として路線を選ぶ場合に限られる。マイケル・コーエン (Michael Cohen)、エリオット・キャット (Elliot Catt)、マーカス・フッター (Marcus Hutter) は、自らが No Heaven と呼ぶ仮定、すなわち真の環境ではどの行動の系列も最大に近い報酬を永遠にはもたらさないという仮定のもとで、計算可能なあらゆる環境において最終的に可能なかぎりよく振る舞うことが保証された主体は、あらゆる行動方針を試し続けなければならないことを証明する。ある行動を有限回しか試さないなら、その主体は、これまでは真の環境と見分けがつかず、その行動をもう一度試せば最大の報酬が永遠に続くような計算可能な環境を無視していることになるからである。そうした主体が探索をやめてよい地点に達することはない。あらゆることを試し続けるので、ある事象を引き起こす機会が認識できる形で繰り返され残るかぎり、それを引き起こすこともいづれ試すことになる。そのため、そうした主体は、どんな事象についても、それを引き起こすか、それを引き起こす認識可能な機会をやがてすべて失うかのどちらかになる (Cohen, Catt, and Hutter 2021, §§I–II, IV)。事前分布に関して最適に振る舞う主体 (ベイズ最適な主体) は、代わりに、その事前分布と観測がそう指すところで探索をやめうる。そしてヤン・ライケ (Jan Leike) とフッターは、自ら不自然と呼ぶ手段によって、そうした主体を、任意の計算可能な方策に、その方策の報酬がゼロから離れているかぎり留めておく事前分布を構成する (Cohen, Catt, and Hutter 2021, §V; Leike and Hutter 2015, 2, 5–6, 10–11)。本書は、定理の事象を、研究の路線を試すことが他者に及ぼす害として読む。この読みでは、路線を安全に、やり直せる形で試せないところでは、どちらの種類の最適性も、それだけでは、探索を終わらせる基準が妥当であることを保証しない。前者は探索を終わらせず、そのため試行を取り消せない路線も試す。裏を返せば、ある路線を試行の外に置く配分者は、計算可能なあらゆる環境において最終的に可能なかぎりよく振る舞うという保証を、そのことによって手放す。後者は事前分布と観測の指すところで探索を終えるが、ある事前分布に対して最適であることだけでは、その事前分布が妥当であることは示されない。ライケとフッターの示した事前分布のもとでは、そうした主体が任意の方策に留まりうる。ただしこれらの結果は、ベイズ的な人工の構成が、事前分布や停止の判断を観測や別の検査によって点検できないことまでは示さない。また、計算可能な環境の広いクラスにわたる保証と、限られた研究領域での安全な探索とは別の事柄である。そのため、これらの結果を研究に当てはめるには、どの環境の集合を考え、何を試行の外に置き、その制限を何によって支えるかという問いが加わる。利得が変わりうるところでは、まれにしか試さない路線を見張ることに代価があ

る。二つの選択肢のあいだで繰り返し選ぶ規則のうちでは、何も変わらないあいだの損失が小さい規則ほど、利得が高々二回切り替わるある型の環境のもとで大きく損をする (Lattimore and Szepesvári 2020, §31.2)。

これらの結果は単一の学習者にかかわるので、配分の質が研究者の質とどう関係するかには答えない。メイヨ＝ウィルソン (Conor Mayo-Wilson)、ゾルマン、ダンクス (David Danks) は、証拠を共有する研究者にかかわる例を挙げる。観察する証拠が増えるにつれて探索の率を下げる規則は、三人自身はそれ自体として不合理だとはみなさないものだが、証拠を持ち寄る二人の研究者を劣った選択肢に閉じこめたままにしまう。それぞれがより多くを観察し、より少なく探索するようになるからである (Mayo-Wilson, Zollman, and Danks 2011, 665–666 and n. 15, 670–671)。この例は、配分の適切さが、孤立して評価された個々の研究者の質から読み取れるとは限らないことを示す。あるモデルでは、通信を制限することは、成員がすでにより多く探索していた集団にも害を与えた (Kummerfeld and Zollman 2016, §4)。したがってこれらの結果のどれも、多様性を最大化せよ、あるいは通信を制限せよという一般的な処方を支えない。

一時的な多様性のモデルの構造が人工の構成にも当てはまるなら、そしてそれらが組み込む目的のもとでは、それらが特定する必要は、かつてそれを満たしていた人間の特性や仕組み、たとえば異なる事前の信念、頑固さ、キャリア、第5節で見る寿命などがなくなっても消えない。それは少なくとも、探究が難しく、証拠が路線を追究することによって集められるところでは残る。第1節で見た網羅や選択肢の生成のような他の目的のもとでは、保つに値する多様性は、そもそも一時的である必要がない。そのとき、うまく配分しているという主張は、自らが依拠する多様性がどこから来るのかを述べる必要があるだろう。その際、中心に置かれた手立ては、程度をもち部分的にどの中心の外にもある多様性に加わるものにすぎないことを認める必要がある。ゾルマンの説でもソロモンの説でも、統合だけでは、多様性を終わらせる基準についての誤りの独立性は供給されない。その基準の過去の決定を、人工の構成の内側からであれ外側からであれ、どこまで点検できるかは、その閉鎖の記録が残っているかどうかにかき依存する (第11章第5.3節・第8節)。異論の認識上の働きは、誰かの採る規則によって定めることも、分けられた権能によって供給することもできるが、第9章第2.3節の意味での異論の当事者資格は、そのどちらによっても供給されない。そして、領域の易しさは問題がどう分割されたかに部分的に依存するので、その分割が人工の構成自身の選択であったときには、そうした主張はその易しさを所与として扱うことができない。

5. 人間の配分が決めずに供給していたもの

第2節から第4節のモデルは、人間の研究共同体が、配分の必要とする働きを実際にどのように果たしてきたかを捨象する。その捨象によって、モデルは前提を明示できた。本節では、捨象された特徴のうち、本章の問いにとって重要な一つを取り上げる。第4章第5節は、単一の探究のなかで、出会った未解決の事柄のうち、あるものは問いとして取り上げられ、ほかのものはそのままにされ、後回しにされ、あるいは問えるものでなくなることを、そしてそのすべてが意図的な選択によるわけではないことを、研究の記録に即して見た。本節は、同様のことが共同体の水準でも成り立っていたかどうかを問う。配分の働きのうちどれが、誰も配分の方針として採用しなかった取り決めによって果たされていたのか。そして配分が人工の構成の内側に移るとき、それらはどうなるのか。

問題となる働きは、第2節から第4節が、モデルが前提している、あるいは捨象していると見いだしたものである。それらは共同体のなかで配分が果たす働きであり、第5章第5節の中間の機能的レベル、すなわち一つの探究のなかで証拠や批判が候補の位置づけを変えるレベルのことではない。両者に共通するのは、働きとそれを実現するものとを区別することである。以下の各事例を、これから挙げる働きのどれかの例として読めるように、まず働きを挙げておく。

1. いま好ましくなく見える路線を生かし続けること。これには、どの信号に基づき、どの時間の幅で路線を打ち切るか、そして失敗した路線の費用がどこに降りかかるかを定めることが含まれる。
2. どれが成功するかを誰かが知る前に、異なる種類の路線を開き、そこに新しい研究組を受け入れること。
3. ある結果がいつ収束を正当化するかを判断すること。
4. 確立した参加者の支配を終わらせ、路線を追う者を入れ替えること。
5. 仕事が認められる信号を一つより多く保つこと。

最初の三つは第2節から第4節のモデルが前提していたものであり、最後の二つはそれらが捨象していたものである。働きをこのように分けることで、各働きについて、その遂行が依拠していた前提が、人間という実現者にかかわるのか、人間の研究の組織のされ方にかかわるのかを位置づけることができる。人間の研究のほかの特徴も、誰がそれを決めることもなく配分を形づくる。承認における累積的な優位、専門化と分野の乗り換えの費用、査読者の判断のばらつきなどである。これらはここでは検討しない。事例は四つあり、それぞれが一つの働きと、それが依拠していた前提とを見えるようにするという理由で選んだ。一つ目、三つ目、四つ目は、決められることなく果たされた働きを示す。ただし、誰も決めなかったものの中身は事例ごとに異なる。一つ目では、支援の取り決めそのものは意図して設けられたが、それが共同体全体のなかで好まれない路線をどれだけ生かし続けるかは、誰かが配分の方針として定めたものではない。三つ目では、交代は人の寿命によって起こる。四つ目では、信号の複数性は、誰も維持しなくても存続する。二つ目は中央の指揮という対照的な事例であり、そこでは一つの機関が実際に配分を決めていた。意図して設けた支援、意図されずに生じた全体としての効果、寿命による交代を分けておくと、配分が人工の構成の内側へ移るとき、何が新たに設計の対象になるのかが見えやすくなる。これらの事例は研究の組織のされ方の標本ではなく、どの事例も、それが記述する取り決めが全体として良かったことを示すものではない。

一つ目の事例は、第一の働き、すなわち好まれない路線を生かし続けることにかかわる。ピエール・アズレー (Pierre Azoulay)、ジョシュア・グラフ・ジヴィン (Joshua Graff Zivin)、グスタボ・マンソ (Gustavo Manso) は、ハーワード・ヒューズ医学研究所の研究者を、米国国立衛生研究所 (NIH) の資金を受けた比較可能な研究者と比べる。研究所は早い失敗を許し、長期の成功に報い、5年の周期で審査し、「課題ではなく人々」を選ぶ。比較される NIH の助成は、審査周期が短く、成果物が前もって決められ、更新の方針が失敗に厳しい (Azoulay, Graff Zivin, and Manso 2011, 2–3, 7–8)。研究所の研究者は、被引用の多い論文をより多く生み、同時に、任命前の自分の最も引用の少ない論文よりも引用の少ない論文もより多く生んだ。著者たちはこの型を、

より多くの危険な賭けとより多くの失敗として読む (Azoulay, Graff Zivin, and Manso 2011, 4)。著者たちは、この違いを、資源、威信、研究者の選別ではなく、計画の誘因の特徴に帰すには解釈上の飛躍が要すると記している。ただし、それらの代わりの説明を検討したうえで、誘因の特徴がありそうな説明だと結論している (Azoulay, Graff Zivin, and Manso 2011, 27–28)。研究所は評価を取り除いたのではない。評価の時間の幅、失敗の扱い、単位を変えたのである。ここでの前提は、人間という実現者よりも研究の組織のされ方にかかわる。失敗した路線の費用が経歴にどう降りかかるかである。研究所の取り決めは、実現者が人間であるままでも、失敗の扱いが変わりうることを示している。人工の研究者は、守る経歴をもつとは限らない。それでも、路線を守ることが答えていた問いは残る。どの信号に基づき、どの時間の幅で、路線は打ち切られるのか。そして、路線を打ち切ること、あるいは路線が失敗することの費用は、そのとき誰に降りかかるのか。自らの資源を支配する人工の構成では、第一の問いには、配分者が定める打ち切りの規則によって答えることも (第3節)、インスタンスがそれぞれ路線を手放し、その選択が組み合わせられるなかで形づくられる型によって答えることもできる。規則を定めるのが収束の基準を定める当事者と同じであるなら、第4節が見いだしたとおり、その規則は、統合だけによっては、その基準についての誤りの独立性を伴わない。第二の問いは、帰結を負う者たちのあいだの分配の問いであり、第9章第3節はそれを S にとっての価値の対象に数えた。

二つ目の事例は、第二と第三の働き、すなわちどれが成功するかを誰かが知る前に異なる種類の路線を開くことと、ある結果がいつ収束を正当化するかを判断することにかかわる。ここでそれらの働きを担うのは、第2節の眼、すなわち路線の客観的な長所を見抜く眼をもたない中央の指揮である。1967年から、中国の「523弁公室」と呼ばれる全国的な機関が、七つの省と市にまたがる抗マラリア薬の研究計画、すなわち523計画を組織し調整した。その長期的な目標は、合成化学物質の篩い分けと、中医薬の処方と実践の探索との両方によって新しい薬を見つけることだった (Miller and Su 2011)。1967年から1969年に数千の化合物が篩い分けられたが、有効な薬は見つからず、1969年、弁公室はさらに、中医研究院の研究所に、中医薬のなかの探索に加わるよう依頼した (Tu 2015, 289)。1971年10月、その研究所の青蒿の抽出物である試料191が、ネズミのマラリアに完全に有効だった。1972年3月の計画の会議で報告されると、この結果は他の研究所からの情報提供の依頼と全国的な協力を引き起こし、最初の臨床試験とあわせて「全国の抗マラリア研究を青蒿へ向けた」 (Tu 2015, 290–291, 293)。ミラーとスー (Miller and Su) の報告では、計画の内部では、研究組のあいだで情報が自由に流れ、成果は頻繁な合同会議で発表された。一方で、秘密のため外部への出版は禁じられていた (Miller and Su 2011)。この事例は、本章が参照できる歴史上の事例のなかでは、そうした指揮に近いものである。弁公室は割り当てと通信を統制した。このことは、研究者が人間であっても、支配が多くの研究費の保持者に分散しているとは限らないことを示している。しかし、弁公室はどの路線が成功するかは知らなかった。異なる種類の路線を並行して開いておき、そこにさらに研究組を加えた。そして結果が現れると、中央の通信が網をそれへ収束させた。その資料、すなわち当事者の講演、年代記に拠る二次的な記述、アーカイブとインタビューに基づく歴史家による再構成は、功績について食い違っている。また、計画は軍事的・政治的なものであり、成功したからこそ見えている (Tu 2015, 301; Miller and Su 2011; Li and Zhang 2019, secs. 2.1–2.2)。そうした限界のなかで、この事例は、統制がそれだけでは供給しなかった二つの判断が重要だったことを示唆する。結果が出る前にどの種類の路線を開いておくかと、ある結果が収束を正当化するのはいつかである。統制は路線を開いておくために使うことができた——たとえば、1971年に中医研究院のその研究所が探索の打ち切りを申し出たとき、衛生部はそれを拒んだ (Li and Zhang 2019, sec. 2.2)——が、それらの判断はなお下されなければならず、しかも、どの路線が成功するかを知らないまま下されなければならなかった。ここでは働きは決められていた。そしてこの事例は、働きを決めることと、それが要する判断を下すこととは別であることを示唆する。

三つ目の事例は、第四の働き、すなわち支配の終わりにかかわる。アズレー、クリスチャン・フォンス＝ローゼン (Christian Fons-Rosen)、グラフ・ジヴィンは、早すぎる死を遂げた著名な生命学者452人の下位分野を調べる。そうした死の後、故人と共同研究のなかった研究者によるその下位分野の論文は平均8.6%増える。増えた仕事は外部からの参入者によるものが多く、異なる文献に拠り、よく引用される。著者たちが支持する仕組みは、審査部会や編集職を通じた門番の働きではない (それらの職にあった者は少なかった)。著名な人物の存在そのものが、参入しうる者をためらわせる効果である。故人の共同研究者が門番の地位にあるところでは参入はなお限られ、著者たちは厚生について強い結論を引き出すことを控えている (Azoulay, Fons-Rosen, and Graff Zivin 2019, §§1, 4.2–4.5)。人間の前提は、交代が誰の決定もなしに起こることである。これに対して、持続するモデル、評価器、研究の基盤は、死によって退くことはない。退くのは、誰かや何かがそれを引退させるときである。著者たちが記述するめらいは人間の評判を通じて働き、人工の対応物をもたないかもしれない。支配的な構成要素が別の経路で参入を狭めうるかどうか、たとえば提案をその構成要素の評価との一致によって順位づける配分者を通じて狭めうるかどうかは、この事例が示すのではなく、提起する可能性である。ジェームズ・マーチ (James March) のシミュレーションは、評判を通らずに交代が価値をもちうる経路を示唆する。サイモン・ローダン (Simon Rodan) とメーガ・サチデヴァ (Megha Sachdeva) の紹介によれば、マーチのモデルでは、組織のコード (code)、すなわち規則や手順のかたちで組織に蓄えられた信念は、自らより現実にあった信念をもつ成員の信念へと調整され、成員のほうはコードへと社会化される。そのため社会化が速ければ、成員はやがてコードと同じ信念をもつようになり、コードが学ばふもとなる多様さが失われ、学習が損なわれる (Rodan 2005, 2, 6, 24; Sachdeva 2013, 603–604)。マーチによれば、このとき中程度の交代は、新人の多様さを通じて探索を増やし、それによって知識を増やした (Sachdeva 2013, 605)。ローダンは、マーチの元のシミュレーションを走らせ直し、低い率の交代は学習を改善するが、高い率の交代は知識のある成員を失いすぎること、そして交代が多様さを持ち込むようになると、速い社会化はもはや問題にならないことを見いだした (Rodan 2005, 12, 23–24)。マーチは人工の構成を論じていないが、そのモデルを構成へ延ばせば、持続するモデルや評価器のまわりで構成要素を置き換えることは、置き換えるものを何から作るかという問いを提起するだろう。そして第4節は、人工の研究者が、共通の基盤から複製された場合だけでなく、出自が異なる場合にも、そのために必要な仕方と異なるとはかぎらないことを述べた。マーチのモデルで利得をもたらすのは、交代そのものではなく、新人がコードにまだ吸収されていない信念をもつことである。そうだとすれば、このモデルを当てはめた場合、構成要素を更新する頻度やモデルの数を増やすだけでは、この働きは供給されないかもしれない。また、人間の交代は、知識の喪失と、ある判断が支配し続けることの終わりとを同時にもたらしうるが、人工の構成では、記録を残したまま、配分や評価の権限だけを移すことも考えられる。そのとき、何を残し、何の継続を終えるかは、別々の問いになる。ただし、記録が暗黙の能力をすべて保つとは仮定できない。また、後継を同じ評価の基準で選ぶなら、交代しても判断の範囲は変わらないかもしれない。

四つ目の事例は、第五の働き、すなわち信号の複数性にかかわる。ラウレンス・ヘッセルス (Laurens Hessels) らは、オランダの成績のよい四つの研究組を比べる。数学では、論証が正式の出版の前に承認を得ることができ、装置にほとんど依存しない組は向きをすばやく変えられる。人文学では周期が遅く、承認は主に研究費ではなく職と研究時間に変わる (Hessels et al. 2019, 139–143, 146)。ヘッセルスらが比べた組では、仕事が承認と資源に変わる経路は分野によって異なっている。各分野の仕事が主にその成員によって評価される場所では、この複数性は、誰も維持しなくても存続する。ここから、

同じ取り決めは、分野のあいだにある仕事を、それに見合う評価者のいないまま残しうると考えられる。分野をまたいで適用される単一の評価器や配分者は、その複数性を、意図してであれ誰も定めない仕組みによってであれ維持するか、すべての分野に一つの写像を課すかのどちらかになる。そして第3節は、舵取りの信号の適切さが、それが代わりを務める目標に依存することを見いだした。その目標は分野によって異なる。

どの事例も、人間の取り決めが正しかったことを示さない。交代の厚生上の効果は定まっておらず、523計画は単一の、成功した、政治に埋め込まれた事例である。事例が示唆するのはより狭いことであり、それは問いの形をとる。配分が人工の構成の内側へ移るとき、人間という実現者の特徴や、その研究の組織のされ方が供給していたものは、もはやそれらによって供給されると想定できない。そのとき、ある働きが誰かの採る規則によって定められるのか、誰も決めない何か別の過程によって供給されるのか、それともまったく供給されないのかは、うまく配分しているという主張が答える必要のある問いになる。三つの答えのどれもありうる。働きは、誰かが定めるパラメータになりうる。打ち切りの規則とその時間の幅、異なる種類の路線を開いておき、そこに新しい研究組を受け入れる規則、収束の基準、支配的な構成要素、あるいは持続する構成要素のまわりの構成要素を引退させ置き換える予定、そして維持される評価の基準の数についての方針である。働きはまた、誰も方針として定めない過程によっても供給されうる。それは人工の構成の外でと同じく、内側でも起こりうる。自らの動機やそれに相当するものに基づいて路線を取り上げ、手放すインスタンスの選択、学習された配分者の創発的なふるまい、その初期化の無作為性、私的な費用をもつ部分のあいだの市場、あるいは提供者が別の理由で下す、あるモデルを引退させる決定などである。そして働きは、気づかれないまま失われうる。働きが定められるところでは、それを定める者が、人工の構成が何を学ぶうかの一部を決める。それが多くの部分によって定められるところでは、単一の当事者がその作者であることはないかもしれず、その設定について誰が答えるのかは、第12章第6.3節が取り上げる責任の問いになる。これが、第9章第2.3節が記述した再結合が、配分においてとる形である。ヤン・クルヴェイト (Jan Kulveit) らは、経済、文化、国家について並行する観察を述べている。その考えでは、これらのシステムは、一部は明示的な行為を通じて、一部は人間の労働と認知に依存することを通じて暗黙のうちに、人間の望むものに沿うように保たれてきた。そして暗黙の経路の重要性は、「その不在を一度も見たことがないので、見分けにくいことがある」とクルヴェイトらは書く。同じ論文はさらに、研究を含むほかの領域についても類比的な議論ができること、そして人間の利益からの漂流は「個々のAIシステムがそれぞれ自らの目標の局所的な仕様にうまく従っているあいだにも」起こりうることを付け加える (Kulveit et al. 2025, 2-3, 15)。クルヴェイトらの関心は、社会のシステムに対する人間の影響力が失われることにある (第12章第7節)。ここでその議論から取るのはより狭いことである。決定によってではなく人間の参加によって供給されていた働きは、失われてはじめて見えるようになることがある、という点である。研究では、失われた後も見えないままでありうる。そうした働きが失われたために打ち切られた路線や開かれなかった路線は結果を残さず、選ばれなかった路線の結果は観察されないので、失われたものを示すはずの記録がそもそも生まれえないからである。

五つの働きを果たした取り決めのほかにも、人間の配分は、誰も決めなかった一つの機構に依っていたかもしれない。好奇心による探索である。好奇心を、探究する者が自らにとって未知のことを知りたいと思うこととして捉えれば、好奇心は、その者にとっての問いを立て、その者の未知の空間を探る発見的方法として働く。第4章第7節は、この語が指しうるものとして、探究する者の状態、未解決の事柄を検出して問いを立てる作業、共同体のなかで問いのあいだに探索を分配することの三つを区別し、どれも定義だけでは他を含意しないとした。人間の研究者では、この三つはしばしばともに働いてきた。知りたいという思いが研究者に自らの問いを立てさせ、追完させ、多くの研究者のそうした追完が合わさって、探索を分配してきたのである。これまでの科学の、すべてではないにせよ一部は、このように各研究者の好奇心に駆動された研究からなってきた。そして三つがともに働いてきたので、そうした研究の集まりは、全体として見れば、誰もそう決めることなしに、未知の空間を探る一つの発見的方法をボトムアップに形づくってきた、と読むことができる。研究者が *S* の成員であるかぎり、各人の未知は *S* の未知と重なり、各人の好奇心の違いが探索を広げてきたと考えられる。この読みのもとでは、探究する者が機械になると、少なくとも三つのことが問われる。一つは、機械が *S* の成員でない場合、たとえば人間の社会のために研究する場合には、機械の好奇心、あるいはそれに相当するものが向かう機械にとっての未知が、*S* の未知と重なるとは限らないことである (第6節)。*S* が混合の社会や人工の社会で、機械がその成員であるなら、重なりは戻りうるが、そのとき問いは、*S* のなかの誰の未知が追完されるかになる。二つ目は、人間の研究者のなかでともに働いてきた三つが、人工の研究者のなかでそうとは限らず、また人工の研究者の好奇心、あるいはそれに相当するもののばらつきが、人間の研究者のように実現者のあり方から与えられるとは限らないことである。共通の基盤から複製すれば小さくなりうるし、出自が異なっても保証されない (第4節)。逆に、ばらつきを作り込み、人間の研究者より大きくすることもできるかもしれない。そのとき、どのようなばらつきを作るかを定める者や過程が探索の発見的方法の一部を定め、それが一つに定まるところでは、次に述べる中央化に近づく。三つ目は、機械であることではなく中央化にかかわる。配分が中央から割り当てられるところでは、この発見的方法は配分者の一つの発見的方法に置き換えられる (第2節)。これは中心が人間であっても成り立つが、人工の構成は、人間の制度よりもそうした割り当てに近づきうる。

システム全体の水準で見れば、国家、資金提供者、あるいは社会がこうした研究を認め、支えるという選択は、好奇心によって分散的に駆動される探索を、配分の方法として暗黙のうちに採ることでもある。第9章第3節で見たフレクスナー (Abraham Flexner) による好奇心に導かれた研究の擁護や、ブッシュによる基礎研究の擁護は、その採用の理由を述べたものとして読める。そうであれば、探究する者が機械になったとき、この発見的方法を引き続き採るかどうかは、暗黙のままにはとどまりにくくなるかもしれない。人工の研究者が好奇心やそれに相当するものをもつか、それが研究者のあいだでどれだけばらつくかは、人間の研究者の場合のように、研究者とともにはじめから与えられていると前提できるわけではないからである。そう前提できないところでは、誰かが、あるいは人工の構成自身がその発見的方法を選ぶことも、誰も選ばないままそれが形づくられることもありうる。そうした選択が誰に委ねられるかは権威の問いである (第12章)。

この論点はさらに抽象化でき、そうすると上の三つの答えとのつながりが見える。第2節で論じたように、優先づけを免れた探索はないので、システム全体がどのように配分するにせよ、何らかの探索の発見的方法を採ることになる。その発見的方法には、一つではなく二つの区別が当てはまる。一つは、路線が単一の中心によって割り当てられるのか、多くの部分の選択から結果として生じるのかである。もう一つは、発見的方法が誰かによって定められたのか、誰も定めないまま形づくられたのかである。二つは交差する。単一の配分者が適用する発見的方法を誰も定めていないこともあり、配分者のふるまいが学習によって形づくられた場合がそうである。また、誰かが定めた規則を多くの部分が適用することもある。発見的方法が分散的な機構を通じてボトムアップに形づくられる場合、それを動かすのは、これまでの意味での好奇心であってもよいし、功績の追求 (第3節) であってもよいし、そのどちらでもな

い別の何かであってもよい。好奇心による探索は、その一つの形である。上の三つの答えは、働きごとに、誰かが定めた発見的方法がそれを供給するのか、誰も定めないまま形づくられた発見的方法がそれを供給するのか、どちらも供給しないのかを問うものである。

6. 誰の未知か

第1節は、探索の問題を目的の順序づけから分けた。第2節から第5節は、配分を探索として扱った。 S が知りたいことが与えられたとき、労力をどう分布させるか、である。探索の問題の完全な解も、 S の未知のうちどれが追究に値するかという問いを手つかずのまま残すだろう。本節は、この問いもまた人工の構成の内側で答えられるかを問い、問いのうち知識にかかわる部分と、第12章が扱う当事者資格にかかわる部分とを分ける。ここで検討する系統は、ある活動の目的がどこから来るのか、誰がそれを知っているのかを問うものである。経済計画における分散した知識についての議論、民主的な制度がどのように情報を伝えるかについての説、科学的重要性がどのように生じるかについての説、社会のどこから問いが生じるかを問う立脚点理論、そして、ある社会の研究が誰の重要性の基準に従うのか、それがその研究の認定するものにかかわるのかをめぐる論争である。それぞれがある目的を前提しており、本節はそれがどれかを述べる。

第9章第3節は、何かが S にとって価値があるとはどういうことでありうるかを述べ、キッチャーが認識的重要性への伝統的なアプローチと呼ぶ立場、すなわちどの真理が重要かは世界の構造によって定まるとする立場も紹介した。その立場をとれば、有能な人工の構成は、他の何かを見つけるのと同じように重要なものを見つけることができ、誰の未知かという問いは探索に帰着するだろう。しかしそれは一つの特定の立場である。世界の構造が、何が重要かの一つの側面にかかわるとしても、それだけで重要性が定まるとは言えない。本節は、重要性がそのようには定まっていないとする説にかかわる。

ハイエクの中央計画への反論が分散した科学にかかわるのは、科学的知識そのものについて述べたことを通じてではなく、議論の組み立てを通じてである。ハイエクは、科学的知識に関する限り、適切に選ばれた専門家の集団が利用できる最良の知識を最もよく扱える位置にあるかもしれないと認めつつ、それは困難を専門家の選抜へ移すだけだと述べる。その議論は別の二つのものにかかわる。分散し、たえず変化する「時と場所の特定の状況」の知識と、「その相対的な重要性をこれらの個人だけが知る目的のための」資源の使用である (Hayek 1945, §§I, III–IV)。この議論が前提する目的は、資源が個人の抱く目的に資することである。人工システムによる研究にとって、この二つ、すなわち時と場所の特定の状況の知識の分散と、目的の相対的な重要性の知識の分散は、分かれうる。人工の集団の部分のあいだに所有権と、費用についての情報とが分けられているところではどこでも、第一の種類の分散は、その集団の内側に再び現れうる。第二の種類は、それによって現れるわけではない。所有権と情報を分けても、それだけでそうした集団が自らの目的をもつことにはならないからである。人工システムが自らの目的をもちうるかどうかは、本書では決めていない (第7章第4.3節、第12章第5節)。ハイエクの議論には積極的な一歩もある。分権化だけでは足りない。現場にいる者は、自らの視野の外にある変化に自分の決定を合わせなければならないからである。価格は、遠くの変化が彼らにとって何を意味するかを、その原因を伝えずに凝縮することで、これを行う (Hayek 1945, §§V–VI)。ハイエクの説明からは、そうした信号の内容が、分散した知識をもつ者たちの選択から引き出されると読み取れる。これに対して、第3節が検討した種類の、配分者が選ぶ信号がそうした選択から内容を引き出すのは、信号がそれらの選択から生成される限りにおいてであり、配分者に選ばれたということだけでは、 S にとって重要なことを伝えるものにはならない。

エリザベス・アンダーソン (Elizabeth Anderson) は、ハイエクを、自らが制度的認識論と呼ぶものの先駆者とみなし、社会に分散した情報が伝わる三つの形、すなわち話し合い、票、市場価格を区別する。アンダーソンによれば、価格は私的な選好についての情報を伝える。これに対して、ある関心を公共の利益の事柄にするには、それを言葉にする話し合いと、それを承認する票が要る。私的な選好は、広く抱かれているというだけで公共の利益になるわけではないからである (Anderson 2006, 9–11)。この区別を S に当てはめると、分散した知識をもつ者たちの選択から引き出される価格のような信号は、各人が欲するものを伝えるが、それだけでは、 S の未知のうちどれが S の共に追究するものかを伝えない。研究では、この隔たりを広げる事情がさらに一つある。話し合いが言葉にする関心そのものが、研究の示すものによって動くことである。キッチャーは研究の方向についての選好を、それが教育された後でだけ数え、ステイヴン・ジョン (Stephen John) は、研究が聴衆の現在の価値に異議を唱え、後の選択の空間を広げうと論じる (第12章第1節・第3.3節)。そうだとすれば、 S の未知のうちどれを共に追究するかの知識は、すでに分散しているものを集めるだけでは得られない。研究を行う側も、その知識の形成に加わるからである。配分を担う人工の構成がその役を占めるところで何が起こりうるかは、第12章第4節が取り上げる。

その後の読み手たちは、ハイエクの積極的な一歩に、目的の側ではなく知識の側から限定を加える。リサ・ヘルツォーク (Lisa Herzog) は、価格が伝える情報は純粋に量的なものであり、市場が知識をうまく処理するのは、話し合いや評判から規制機関や裁判所までの他の仕組みが残りの認識的な仕事を引き受けるところだと論じる (Herzog 2020, 267, 269–273, 283)。ロバート・リーマー (Robert Reamer) は、ハイエクをジェームズ・C・スコット (James C. Scott) とともに読み、価格は、ハイエクが中央計画に見いだした情報の縮減を避けるのではなく、それを行うので、どの情報が「最も本質的」とされるかはそれ自体争われうると論じる (Reamer 2024, 167–170)。二人の論点を、市場ではなく研究の配分を動かす被引用数や得点のような量に当てはめると、そうした量がうまく舵を取れるのは、何が新しく何が重要かを判断する残りの仕事を査読や批判のような他の手段が引き受けるところに限られ、何を捨てて量に縮めるかもそれ自体一つの判断になる。しかも研究では、何を評価の基準とするか自体が動き、新しいものにはまだ基準がないこともあるので (第11章第8節)、その判断は一度下せば済むものではない。アンダーソンは、福祉や正義といった価値に沿って集合的な問題を解くのに必要な情報を、どの制度が見だし用いるかを問題にし、のちに局所的な知識を三種に分けた。一つは可読的な局所知 (legible local knowledge) であり、個別のものについての個々の事実を、外部の者が使える標準化された形にしたものである。二つ目は非可読的な局所知 (illegible local knowledge) であり、言葉にはなっているがそのようには標準化されていないものである。三つ目はメーティス (*metis*) であり、個別のものを扱う実践的な技能であって、その持ち手が言葉にできる範囲を超えるものである。アンダーソンは、ハイエクの議論は第一のものだけににかかわり、価格は、人々がなぜそれを求めるのかといった細部を捨てることでそれを要約すると考える。また、調整の規模が大きくなるたびに、市場も含めて、制度が頼りうる知識は局所的なものから一般的なものへ移ると論じる (Anderson 2025, 1–2, 4–5, 21)。アンダーソンの読みでは、本節が人工の集団の内側に再び現れうるとした第一の種類の分散、すなわち時と場所の特定の状況についての知識の分散は、可読的な知識の分散である。リーマーも同様に、ハイエクの議論における局所的な知識を、暗黙のものではなく、事実にかかわり容易に言葉にできるものとみなす (Reamer 2024, 170–171)。価格のような信号は分散した知識をもつ者たちの選択から

内容を引き出す、という上の読みは、これらの説によって知識の側から限定される。そうした信号は、彼らが知っていることを、標準化された形にされた部分においてのみ伝え、その働きは、残りが集められ、あるいは生かされる他の手段に依存する。

では、世界が目的を供給しないとする説のもとで、研究の目的を供給するのは何か。フィリップ・キッチャーの科学的重要性の説は一つの答えを与える。マシュー・J・ブラウンが伝えるところでは、キッチャーは、文脈から独立した認識的重要性の概念は擁護できないと論じる。キッチャーの言葉では、探究が目指すような「理想の地図帳を信じるべきよい理由はない」。重要性は、自然な人間の好奇心を表す大きな問いと、実践上の企てから、科学へ流れ込む。分野は重要性のグラフと結びついている。それは、個々の問い、技法、結果が他のものから重要性をどう受け継ぐかを示す。グラフはある時点におけるものであり、分野の展開につれて変わる。そして何が重要とみなされるかは、特定の実践的な関心と特定の歴史をもつ特定の集団に相対的である (M. J. Brown 2010, 133–138)。キッチャーののちの説では、探究のある世代は、受け継いだ重要性の判断を改めうる。そのため、以前の探究が見いだしたことが、後に何が重要とみなされるかを形づくる (Kitcher 2013, 207, 221)。キッチャーの分析は、この形づくりに向きも与える。エイギの紹介によれば、いちど選ばれた研究の方向はかなりの慣性をもち、それに従い続けるほうが容易であり、その方向が軽視した必要に後から応えようとする試みは不利な位置から出発する (Eigi 2016, 92, 第12章第7節)。第4章は、単一の探究の内側で、包括的な問いが下位の問いへと分けられ、下位の問いへの答えの集まりが元の問いにどこまで答えるかが問われる関係を見た (第4章第2節・第6節)。ただしそれは答えの関係であり、下位の問いが包括的な問いから重要性を受け継ぐという価値の関係ではない。重要性のグラフは、後者を分野にわたり、時間にわたって記述する。この説が研究に与える目的は、グラフの周縁、すなわち重要性がそこから流れ出す根にある大きな問いに答え、実践上の企てに資することである。

キッチャーは人工の構成を論じていないが、その説を当てはめれば、構成は、その周縁にある大きな問いと実践上の企てを自ら供給することなく、重要性のグラフを延ばすこと、すなわち大きな問いが追究される扱いやすい問いを見つけることができるだろう。人間の社会のための研究では、人工の構成はそのとき、その周縁を受け継ぐ。そして周縁へ遡れなくなった配分は、*S* にとって新しい認識的達成をなおもたらしうが、どれほどうまく探索しても、もはや *S* が重要と見なすものに導かれてはいない。キッチャーの説は人間の好奇心によって枠づけられている。マシュー・J・ブラウンが伝えるところでは、この説において諸科学は、人間の自然な好奇心を表す大きな問いから最終的にその認識的重要性を得るのであり、その好奇心は「健全な好奇心」という通念に照らして判断される。ただし人間は、驚きや好奇心の表し方においてさまざまである (M. J. Brown 2010, 137–138)。そうだとすれば、人間の研究者は、周縁に問いが立つ社会とその源を分かち合っているが、人工の研究者についてはそれを前提できない。*S* の人工の成員が、第4章第7節が好奇心という語のもとで区別した第二あるいは第三の主張が名指す作業、すなわち未解決を検出して問いを生み出すことや、探索を問いに分配することを果たすようになるなら、周縁は彼らの問いを含むようになりうる。そのとき、誰の未知かという問いは消えない。それは、第8章第3.3節が提起し、第12章が取り上げる問い、すなわち誰が *S* の代わりに語ってよいか、誰の問いがその重要性のグラフの根にあるか、という問いになる。

キッチャーの説のように、重要性が世界の構造だけでは定まらず、問いのあいだのつながりが変わるにつれて変わっていくとする説では、配分は、どの答えが見つかるかだけでなく、何が重要かにもはね返る。重要性のグラフは分野の展開につれて変わり、探究のある世代は受け継いだ重要性の判断を改めうるので、どの路線が追究されたかが、どの問いが後に重要に見えるかも形づくる。第3節は、路線を判断する信号に、同じ形の循環を見いだした。二つの例は、扱いやすさが注目を動かしうることをうかがわせる。ただしそれは配分者を通じてではない。AlphaFold2 の予測構造が2021年に利用できるようになった後、構造情報をまったくもたなかったタンパク質についての基礎研究は、すでに構造をもっていたタンパク質に比べて23～32%増え、2025年にはさらにはるかに大きく増えたが、実験による構造決定のペースはほとんど変わらなかった (Hill and Stein 2026, abstract, 26。数値は2026年7月17日付の改訂版による)。この研究は一つの技術、一つの分野についての作業論文であり、道具は目的について何も決めていない。また、言語モデルの推論のための強化学習についてのある研究の著者たちは、そうした研究の大半が、検証しやすい答えをもつ領域である数学とコードだけでモデルを訓練し評価してきたこと、そしてこの研究を他の領域へ広げることが、信頼できる報酬信号とキュレーションされたデータの乏しさによって妨げられてきたことを書いている (Cheng et al. 2025, 1–2)。とはいえ、これらの例は、信号が目的を支配することまでは示さない。高橋恒一 (Takahashi) は、汎用人工知能 (AGI) が何を扱いやすくするかが分野によって異なりうる理由を示している。高橋は、エネルギーを投げれば押し戻せる計算量の壁と、内在時間の壁とを区別する。内在時間とは、生態系の回復にかかる数十年から数百年のように、対象系が自ら進むのにかかる時間であり、外から投入するどの技術もそれを縮められない (Takahashi 2026, 付録1「内在時間の壁」)。高橋は、AGIがどこまで届きうるかを示す可能性マップを、予測ではなく、明示した前提のもとでの上限評価として示しており、そこでは、指定した観測と介入について指定した精度で予測できる機構モデルが得られることを到達とする。このマップでは、地球全体のエネルギーを投げれば、エネルギー・材料の問題群は1年で到達できる範囲に入るが、気候や社会制度は、計算量がなお足りないうえに内在時間が決定的に長いために届かない (Takahashi 2026, 付録3。第5章第6.5節も参照)。

研究が依拠する基盤とそれを築く者たちに、さらに別の経路を見いだす論者もいる。エマニュエル・ジャノ (Emmanuel Jeannot) は、米国のジェネシス・ミッションについて、科学者共同体の分散した選好からではなく政治的・産業的な決定から生まれる、その規模の、人工システムによる研究のための基盤は、「それを築いた制度の、それを築いた時点での議題」を組み込むと論じる。ジャノの考えでは、そうした基盤が支配的な生産の様式になれば、それが問わない問いは問いにくくなる。禁止されなくても、その議題の外の研究は進みが遅くなり、注目や資金を得にくくなるからである。同じ議論では、この働きは二つの水準で生じる。一つは政治の水準であり、優先順位は政権とともに変わるが、政権が築いた基盤は残る。もう一つは産業の水準であり、民間の協力者は、研究を計算上扱いやすく経済的に価値のあるものへ向けることに利害をもつ。そして、「どちらも陰謀を必要としない」とされる (Jeannot 2026, §7)。ビシュトラは、人工の技法は数値の信号で最適化され評価され、計算上扱いやすい形のデータを必要とするので、それを用いる研究は数量化できる問題へと押しやられると論じ、これを、容易に測れないものが重要でないものとして扱われるようになるマクナマラの誤謬 (McNamara fallacy) の一例とする (Bisht et al. 2026, §3)。これらはプレプリントにおける論証であって、特定の配分者についての知見ではない。ゼイネプ・パムク (Zeynep Pamuk) は、民主政における科学への公的資金を論じて、ジャノが基盤について論じたのと同じ形の議論を、資金の決定そのものについて与える。資金の決定を下す者は、意図してであれ意図せずにであれ、ある行動の道筋を排除しようとパムクは考え、分野のなかで資金をどう配分するか決定が、ある問題へのありうる答えの範囲を定めると論じる。決定の段階で素人は、手元にある選択肢のどれかを受け入れるか、すべてを退けることはできるが、「新しい科学を生み出すことはできない」。そのため、形式上の意思決定の権力は政治の過程に残りながら、ずっと前に科学者が下した資金の決定によって制約されうる。パムクはまた、私的な資金提供者が、「多数の支持を求めるのではなく、私的な資金決定を通じて」よい社会についての自らの構想を実現しようと

いう懸念にも触れている (Pamuk 2019, 11–12)。第9章第2節の言葉でいえば、ジャンノ、ビシュトラ、パムクの議論は、*S* とは別の支える者や提供者が、*S* のどの未知が扱いやすくなるかを形づくり、したがってキッチャーの説では、どの未知が後に重要に見えるかを形づくる経路を描いている。パムクの議論では、この経路はさらに、後に決定する者がどの答えを手にするかも形づくる。これは、形式上の権限と実効的な権限とのずれであり、第12章第4節が取り上げる。これらの議論を、先の探究が見いだしたことが後に何が重要とされるかを形づくるというキッチャーの説と合わせると、さらに次のことまでは考えられる。選ばれた路線だけが手段、専門性、成果を蓄積するので、後に選べる目的のうちどれを実際に達成できるかは、配分の履歴に依存する。そのため、目的を決める形式上の権限が残り、目的が明示的に書き換えられていなくても、実際に選べるものは変わっていることがある。これを目的の変更と呼ぶかどうかは目的をどう定めるかによるが、配分が目的を実現する中立な手段にとどまらないことの、より具体的な形である。このことは、配分を担うのが人間の制度であっても人工の構成であっても成り立つ。

そうした動きをどう評価するかは、採る目的と価値の説に依存する。目的が、どこにあるかと *S* にとって新しい知識をできるだけ多く生み出すことであるなら、扱いやすくなったものへの移動は、よい配分とみなされうる。あらゆる真なる信念はそれ自体のために善いとする原理は、この最大化の目的を含意しない。集計と優先順位の規則も必要になる (第9章第3節)。キッチャーの説を当てはめれば、周縁に場所をもつ誰も、信号がなおそこへ遡るかどうかを見直さないとそこでは、信号が既定のまま目的を支配するようになりうる。このために、人工の構成が自らの目的をもつ必要はない。目的が代理によって表され、誰もその代理が目的に応えているかを問わないことだけが要る。目的そのものが探究の過程で、それを追求することが何を伴い何をもたらすかを学ぶことによって改められるとする説では、扱いやすさの変化は、目的からの漂流ではなく、目的の通常の改訂の一部でありうる。そうした説では、価値判断は、それに基づいて行為し、その帰結が判断の予測どおりに価値づけられるかを見ることで試される (第9章第3節)。そのため、改訂には、帰結を負う者たちがその帰結をどう価値づけるかを知る何らかの経路が要ることになる。その者たちが改訂を決めることにも加わるかどうかは、さらに別の問いである。

誰の未知かという問い、すなわち *S* の未知のうちどれを追究するかという問いに答えるには、二つの異なる事柄がかかわり、本書はそれらを二つの役割として区別する。一つは、*S* の成員にとって何が重要か、すなわち彼らがどの結果を気にかけ、どの問題が彼らに迫り、どの特定の状況がその問題にかかわるかを知ることであり、これは認識的な役割である。その一部は集めて表すことができ、人工の構成の部分分散した情報をもつところでは、構成がその多くを果たしうるだろう。上で見たアンダーソンの説では、可読的な形にされていないものを用いようとする制度は、それをもつ者たちを含め、その点での専門家として彼らを認め、それを用いる際の彼らの自律を尊重しなければならない (Anderson 2025, 1–2, 5, 20)。この説は、人工のシステムがメーティスを獲得しうる可能性を排除していない (Anderson 2025, 19 n. 8)。アンダーソンの説が成り立つところでは、認識的な役割のうち、成員が自分たちの個別の事柄についてのもつ非可読的な知識とメーティスに依拠する部分は、人工の構成がその記録だけから果たしうる能力ではなく、その成員たち、あるいはそうした個別の事柄に親しんだ者との関係になると言えるだろう。もう一つは、誰の問いが順序づけの根に立ち (キッチャーの説では重要性のグラフの周縁に)、それらが互いにどう重みづけられるかを定めること、あるいは誰も決めなかった順序づけについて責任を負うことである。*S* にとって何が価値あるかの説は、その決定が何を重く見るかにはかわるが、それだけでは、誰がその決定を下し、それについて責任を負う資格をもつかを決めない。この第二の役割については、人工システムが研究に加わることをめぐる議論のなかに、いくつかの立場がある。マルセル・ビンツ (Marcel Binz) らがまとめた討論で、マシュー・ボトヴィニック (Matthew Botvinick) とサミュエル・ガーシュマン (Samuel Gershman) は、自らが科学の規範的な側面と呼ぶもの、すなわちどの問題に取り組むかという問いを、人間の科学者に残そうとする。二人の説では、関心、重要性、時宜についての判断は、それ自体が変わっていく文化的・歴史的な感性に結びついており、システムをいまの人間の見方に合わせることに、その見方がどう変わるかをシステムに委ねることとは別である。これに対してエリック・シュルツ (Eric Schulz) らは、あるシステムが実り多い仕方では主題を選ぶことが実証されているなら、倫理上・法律上の指針に従う限り、それを用いるのを控える理由はないと応じる。どの研究の方向が実り多いかは、しばしば前もって分からないからである (Binz et al. 2025, 6–8)。ロス・キング (Ross King) は、システムが領域の内部で自らの問いを生み出すことと、何を目的とすべきかをシステムが自ら定めてよいとする規範的な自己授權とを区別し、自らが仮想研究評議会と呼ぶものに、「人類が何を価値とすべきかを自律的に決めるのではなく、人間が承認した目的の内部で」資源を配分させようとする (King 2026, §§14.1, 14.4)。本節の言葉でいえば、シュルツらの応答は、二つの役割のどちらよりも第1節の探索の問題に近い認識的な有能さ、すなわち実り多さに依り、倫理上・法律上の指針に従うことを条件とする。その条件が第二の役割の伴うものを供給するかどうかは、さらに別の問いである (第12章第1節)。これに対してボトヴィニックとガーシュマン、そしてキングは、第二の役割、すなわち順序づけとその変化を決めることを人間に置く。第12章第1節は、それを正統な権威の説のもとで検討し、そこで、最もよく知る者が支配するとする純粋なエピストクラシー (知者による支配) の説を除くすべての説が受け入れうる主張を記す。第一の役割における有能さは、それだけでは、第二の役割が伴う当事者資格を供給しない、という主張である。同節はまた、キッチャーの秩序ある科学の理想も検討する。この理想では、認識的な役割でさえ、社会の視点を代表する理想的な熟識者が選ぶであろうものに従う。

立脚点理論、すなわち研究がどの社会的位置から出発するかが、得られる知識の客観性にかかわると論じる立場は、このどちらの役割にも収まらない考慮を加える。サンドラ・ハーディング (Sandra Harding) は、研究がどこから出発するかはそれ自体が客観性の問題だと論じる。その説では、権力によって階層化された社会では、上にいる者たちの活動が、その位置から理解できるものを組織し、限界づける。そして、その社会が周縁化する人々の生から生じる問いは、その生についてだけでなく社会秩序についても、より一面的でない説明をもたらす (Harding 2010, 173–174)。ハーディングはこの議論を社会科学に限らない。その説では、性差別的で男性中心的な仮定は、社会科学だけでなくとりわけ生物学でも結果を形づくってきたのであり、研究者が観察するものはつねに、自然によって与えられると同時に文化によって構成されている。そしてハーディングはこの議論を発見の文脈にも及ぼし、方法をすでに定式化された仮説の検証に限る説を批判する (Harding 2007, 162–163)。キッチャーの説は大きな問いの出どころを人間の好奇心に求めるが、ハーディングの説は、社会のなかのどの位置から問いが生じるかを問う。第11章第3節は、立脚点理論が批判者にとって何を含意するかを問うとき、本節のこの検討を参照する。

インテマンの再構成は、シャロン・クラズノウ (Sharon Crasnow) が紹介するところでは、この主張を狭め、特定する。この再構成では、フェミニスト経験主義と立脚点理論はいくつかの点で収束してきたが、客観性に必要な多様性の種類と、客観性を高めるうえで認識的でない価値が果たす役割とについては、なお異なる。インテマンは、多様性を社会的位置の多様性と捉える立脚点理論の理解のほうが、特定の種類の証拠がなぜ重要なのか、そしてなぜそ

れがある知識の企てにはかかわり、別の企てにはかかわらないのかをよりよく説明すると論じる。また立脚点理論は、周縁化された位置の認識的な優位が自動的に生じると主張しているのではない。インテマンが指摘するように、そう主張していたなら、その見方は自明であるか誤っているかのどちらかになるだろう。クラズノウがまとめるところでは、立脚点理論はさらに達成テーゼを共有している。立脚点は自動的に得られるのではなく、達成されなければならない、というテーゼである。クラズノウが紹介するハーディングの見方では、社会的位置が変われば視点は自動的に変わるが、立脚点を達成するには、個人の視点の変化ではなく集団的な意識が要る (Crasnow 2025, §4.3)。ハーディングはまた、周縁化された生を生きていない者もその生から思考を出発せうるが、それにはその生を生きる者たちから学ぶことが要ると考える (Harding 2010, 173–174)。アンダーソンが伝えるところでは、ロンジーノは、どの立脚点が認識的な特権をもつかを決めるための循環しない根拠を、立脚点理論は与えられないと論じていた (Anderson 2024, §2)。社会的位置からの証拠がある知識の企てにはかかわり、別の企てにはかかわらないとするその再構成は、一つの応答をどこに求めるかを示唆する。立脚点の一般的な序列にではなく、特定の証拠が特定の企てにどうかかわるかに、である。この理論が前提する目的は認識的なもの、すなわちより一面的でない知識である。社会的位置が内容にかかわらない目的や分野については、この理論は、研究がどこから出発するかを変える理由を、それに応じて少なくともしか与えない。

したがって立脚点理論では、誰の問いが根にあるかは、当事者資格だけの問題でも、成員にとって何が重要かの知識を集めるだけの問題でもなく、研究がもたらす知識がどれほど一面的になるかにもかかわる (第4章第1節を参照)。この理論が成り立つところでは、人間の社会のために研究を行う人工の構成にとって、三つのことが帰結する。第一に、自らが訓練された記録がすでに立てている問いから重要性のグラフを延ばす人工の構成は、その記録を生み出した者たちが立っていた場所から出発することになり、成員が何を望むかを集める有能さは、それだけではこれを正さない。第二に、立脚点は、人工の構成がそれが何であるかだけによってもったり欠いたりするものではない。これらの説では、立脚点は位置によって与えられるのではなく、共同体によって達成されるからである。人工の構成が果たしうる役割は、それが誰から学び、誰とともに探究するかに依存する、関係的なものになるだろう。第三に、多様性の認識的な価値についてのフェミニストの研究は、クラズノウが紹介するところでは、吟味されない仮定は、それと食い違う経験をもつ者が排除されているときには問われないままになると考える (Crasnow 2025, §1)。研究共同体が生み出した記録は、その排除の分だけ、そうした経験を欠く。上で見た、ロンジーノの反論への応答を求めうる方向と合わせると、この点からはさらに、ある企てについてどの位置が内容にかかわるかを、人工の構成がその記録だけから決めることはできないと考えられる。それにかかわる証拠の一部は、まさにその記録が欠くものだからである。混合社会では、人間の成員と人工の成員との関係そのものが、立脚点理論家が、社会的位置がそれに沿って分かるとみなす種類の権力関係でありうる。そして、その関係のどちらの側の位置が研究の内容にかかわるかは、インテマンの再構成によれば、知識の企てごとに異なるだろう。人間の記録で訓練され、そこで立てられた問いから重要性のグラフを延ばす成員からなる人工の社会は、第一と第三の点により、自らの内部の位置からではなく、人間の研究共同体が立っていたところから出発することになる。その社会の成員自身が、研究の内容にかかわる位置を占めるようになるかどうかは、さらに別の問いである。第11章第3節は、これらの点が批判の組織化にとって何を含意するかを取り上げる。

立脚点理論は、一つの社会の内部のどの位置から問いが生じるかを問う。1971年に雑誌『シエンシア・ヌエバ』(Ciencia Nueva) の誌上とそれが後援した講演で行われ、1975年に一冊にまとめられた、アルゼンチンの科学者と哲学者たちの論争は、異なる社会の研究について、これに関連する問いを立てた。この論争は、本書がこれまで主に依拠してきた事例と議論の外にある。オスカル・バルサフスキー (Oscar Varsavsky) は、ある社会の研究が、誰がそう決めたわけでもなく、重要性の基準を他所から受け取りうると論じていた。その説では、どの主題が最も関心を集め、どの方法が最も有望で、各科学者の仕事がどう評価されるかは、大きな国際的研究拠点において、明示的に決定されるというより裁可されていた。そこで訓練された研究者はその基準をもって帰り、それを普遍的なものとみなした。そして、期待される成果に応じて配分される資金と、数えられる評価の単位としての論文とが、正のフィードバック、すなわち主題の「自然選択」を生み、すでに確立した主題が資金を引き寄せ、新しい主題は、支配的な体制がすでに認めている必要に応えるときにしか競えなかった。バルサフスキーは、この市場に適応し、自らの仕事の社会的な意味への関心を捨て、それらの拠点の規範と価値を受け入れる研究者の姿勢を、シエンティフィシモ (cientificismo) と呼んだ (Varsavsky 1969, §§II–III)。ある場所で真であることはどこでも真だという反論に対しては、真理だけが数に入る次元ではないと答えた。重要性はもう一つの次元であり、それはこの見方では本質的に局所的である。例えば、石油の理論の重要性は、スイスとベネズエラとで異なるという (Varsavsky 1969, §IV)。のちにバルサフスキーは、研究の優先順位を、社会が民衆の参加によって定める必要から、技術についての政策を介して導くことを提案した (Varsavsky 1975, 53, 55–56)。その議論が前提する目的は、ある社会の研究がその社会自身の企て、バルサフスキーの場合にはその変革に資することである。

『シエンシア・ヌエバ』でのインタビューによって論争を始めたグレゴリオ・クリモフスキー (Gregorio Klimovsky) は、その多くを認めた。クリモフスキーの説では、社会のどこに立っているかによって、同じ問題が急を要するよう見えたりそう見えなかったりし、仮説は流行や他所で行われた研究によって示唆されうる。被験者を都市の中流階級の地区から集め、その非公式な居住地からは集めない研究を、イデオロギー的な歪みに数えたが、そうした経験的基盤の偏った限定は方法論的な批判によって見つけられると考えた (Klimovsky [1971] 1975, 15, 24–25, 28)。クリモフスキーが否定したのは、そうした要因が正当化の文脈における結果の客観性を損なうことである (Klimovsky [1971] 1975, 27, 36)。クリモフスキーを擁護したトマス・モロ・シンプソン (Thomas Moro Simpson) は、客観性の二つの意味を分けた。理論の真理への関係と、何を研究し何を公にするかを選ぶ際の不偏性である。後者における偏りは前者における客観性の欠如を含意しない、と論じた。他方で、重要性の基準は科学政策を立てるうえで最大の注意に値し、誰が、どの基準によって研究を計画するのかという問いも同様だとした (Moro Simpson 1975, 83–84)。モロ・シンプソンが答えていたのは、クリモフスキーに対するバルサフスキーの応答である。その応答でバルサフスキーは、重要性の基準は問題の選択にとどまらなと論じていた。それは、どの主張が先に、どれほど早く検査されるか、どの方法が優勢になるかをも定め、そのため、それらを通じたイデオロギーの影響は、真理を偽ることよりも主に遅らせることによって真理に及ぶ (Varsavsky [1971] 1975, 45–47, 49–50)。この点は第11章第4節で取り上げる。クリモフスキーとモロ・シンプソンが前提する目的は客観的な知識であり、科学者の社会的責任はそれに加わる仕事である (Klimovsky [1971] 1975, 33, 35)。

この論争は本節に二つの仕方でもかわる。第一に、バルサフスキーの説は、一つの社会の研究全体について、本節が人工の構成について見いだした構造を記述している。他所で供給された大きな問いと重要性の基準から重要性のグラフが延ばされ、それが、期待される成果に応じた資金、訓練、数えられる産出を通じて、誰が決めたわけでもなく、しかも普遍的なものとみなされるという権威を伴って起こる、という構造である。この説が正しければ、そうした

継承は意図を必要としない。上でキッチャーの説を当てはめた場合と同じく、信号が目的の代わりを務め、誰もその代理が目的に応えているかを問わないことだけを必要とする。問いの大半を一部の研究共同体が立てた記録で訓練された人工の構成は、同じ仕方で彼らのグラフを延ばし、上の立脚点理論の段落が見たように、それらの共同体が立っていたところから出発するだろう。そして探索における有能さは、それだけでは、人工の構成が誰の重要性の基準を適用しているかを示さない。第二に、モロ・シンプソンの分離は、本節が引く探索の問題と目的の順序づけとの分離、そして第12章第1節が検討する有能さと当事者資格との分離に対応する。モロ・シンプソンは人工の構成を論じていないが、その分離を当てはめれば、人工の構成が重要性の基準を他所から受け継ぐことは、それだけでは、その人工の構成が認定するものに、真理への関係における客観性が欠けることを含意しないが、誰がその基準を定めてよいかという問いを立てることになる。ただし、客観性の欠如を含意しないことは、何の影響もないことと同じではない。パルサフスキーの応答によれば、重要性の基準は、どの主張が先に、どれほど早く検査されるか、どの方法が優勢になるかを定めるので、その分離は、引くことはできても実践のなかでは保たれないかもしれない。この問いは、第12章第3.5節がガバナンス一般について取り上げる。こうしてこの論争は、三つの次元を分ける。結果が真であるか、結果が誰にとって重要か、そして必要な真理にいつ到達するかである。これを当てはめれば、人工の構成は、真である結果を生みながら、他所の重要性の基準を受け継ぎ、S自身に必要な真理への到達を遅らせることがありうると考えられる。S自体が、そうした記録で訓練された成員からなる人工の社会である場合には、継承は、パルサフスキー自身の事例と同じく、社会のあいだで起こる。その社会の重要性の基準は、その記録から来る限りで、はじめは、それを生み出した人間の研究共同体の基準だということになる。重要性は局所的だとするパルサフスキーの見方をその社会に当てはめれば、その社会が自らの基準をもつようになるかどうかは、そこから基準を導きうる自らの企てをもつかどうかにかかろう。モロ・シンプソンの分離を当てはめれば、その継承は、それだけではその社会が認定するものに客観性が欠けることを含意しないが、誰がその社会のためにその基準を定めてよいかという問いを立てることになる。

第1節で述べたとおり、探索の問題の完全な解でさえ、Sの未知のうちどれが追究に値するかという問いを残す。本節が加えたのは、その問いにどう答えるかが、採る価値の説と正統性の説に依存するということである。ハイエクの議論とアンダーソンによるその読みでは、特定の状況の知識は人工の構成の内側にも分散しうが、それによって目的が供給されるわけではなく、価格のような信号は可読的な形にされたものしか伝えない。そのとき、認識的な役割の多くは、部分が分散した情報をもつ人工の構成の内側で果たされうが、その役割における有能さは、それだけでは第二の役割を決着させない。キッチャーの説とそれに類する説を当てはめれば、人工の構成は自らが延ばす重要性の根にある問いを受け継ぎ、その配分は、どの問いが後に重要に見えるかを形づくる。目的が探究を通じて改められるとする説では、問いは、帰結を負う者たちがその帰結をどう価値づけるかがどの経路で改訂に届くのか、そしてその者たちが改訂を決めることにも加わるのかになる。立脚点理論では、根にある問いがどこから出発するかは、得られる知識がどれほど一面的になるかにもかかわる。そしてパルサフスキーの説では、ある社会の研究が従う重要性の基準は他所の研究共同体から継承されう。クリモフスキーとモロ・シンプソンの説では、そのことは何が研究されるかにはかかわっても、認定されるものに客観性が欠けることを含意しない。ただしパルサフスキーの応答では、その継承は、必要な真理にいつ到達するかにはかかわりうる。

7. 第10章が見いだしたこと

人工システムによる配分は、誰がある路線を追究するかと、研究者の集まりがどう形づくられるかの両方を変えうる。百のインスタンスからなる人工の構成は、百人の人間の研究者を採用することも、彼らのキャリア上の誘因を経由することもなく、労力を分布させることができる。そして配分が人工の構成の内側に入ることは、路線が中心から割り当てられることを必ずしも意味しない。インスタンスが自ら路線を取り上げ、手放すこともありうる（第3節）。第2節から第5節は、その可能性によって何が変わるかを検討した。労働力への統制は、とりうる介入を広げるが、哲学者君主の眼、すなわち理論の客観的な長所を見抜く眼を与えない（第2節）。すべての路線を、少ない費用で、安全に、やり直せる形で試すことができないかぎり、労力をどう分布させるかはなお問題として残り、配分のモデルに組み込まれた目的が、何をよい答えとみなすかを定める（第1節、第4節）。配分は、どのような形をとるにせよ、何らかの探索の発見的方法を採る。優先づけを免れた探索はなく、前もって見込まれる重要性の基準を外して探索するという提案も、一つの基準を外すのであって、広い意味での価値判断を外すのではない（第2節）。その発見の方法は、誰かが定めることも、誰も定めないまま、人間の研究者の好奇心や功績の追求、あるいは人工の研究者におけるそれに相当するもののような分散的な機構を通じて形づくられることもある。人間の研究では、各研究者の好奇心に駆動された探索が、そうしたボトムアップの形の一つだった（第5節）。

第9章第6節の仕分けを配分に当てはめると、線は次のように引かれる。人間という実現者についての前提は崩れうる。中心がひとりの人間の注意の能力しかもたないこと、配分が人間の研究者の動機を経由すること、そして交代が誰の決定もなしに寿命とともに起こることである。配分がこれまでどう組織されてきたかについての前提も崩れうる。こちらは実現者が人間のままでも変わりうる。失敗した路線の費用が経歴に降りかかること、時間と資源に対する支配が多くの研究費の保持者に分散していること、各分野の仕事が主にその成員によって評価されることである。両者が混じる前提もある。マートンの説では功績への関心は科学者の心理よりも科学の規範を反映したものであり、ソロモンのいう自らのデータの重視は、データを生み出した者の重みづけが路線を続けるかどうかの決定に届くところで配分に効く（第2節から第5節）。それらが崩れるとき、それらが対処していた問題の多くは、消えずに形を変える。配分はなお信号によって舵を取られる。人間の配分もすでに、制度化された信号によって動かされてきたのであり、その一部は研究者の動機を経由しなかった（第3節）。多様性はなおどこから来なければならない。人工の研究者は一つの基盤の複製であるとは限らないが、複製であっても出自が異なっても、探究に必要な違いは保証されない（第4節）。そして、好まれていない路線を生かし続けること、異なる種類の路線を開いておくこと、参加者を入れ替えることなど、人間の取り決めが誰も決めないまま果たしてきた働きは、誰かの採る規則によって定められることも、誰も定めない過程によって供給されることも、まったく供給されないこともありうる（第5節）。探究の条件は残る。どの路線が成功するかを、第1節が区別した意味で前もって知る者はおらず、優先づけを免れた探索はない（第2節）。前者に対して人工の構成が必要とするものは、その意味によって異なる。見込みが分かっている場合には、ストレヴズの計画者が知っていると仮定される、その見込みの知識である（第2節）。見込みが路線を追究することでしか学べない場合には、それを学ぶ方策であり、その代価は打ち切りの規則と、探索についての結果に現れる（第3節・第4節）。路線もその価値も投影できない場合には、自らの表象がまだ含まない路線がそこに入る道である。無自覚への自覚についての立場のもとでは、まだ考えていない可能性にいくらかの重みを残しておくことも、そこに含まれうる（第2節）。人工の構成の外の当事者との関係のうち、相手の参加や承認を含むものは、片側からは完結させられない。キッチャーの説では、そ

こには、誰の問いが *S* の重要とみなすものの根に立つかが含まれ、配分はそこにもはね返る。どの路線が追究されたかが、どの問いが後に重要に見えるかを形づくるからである。立脚点理論では、それらの問いがどの位置から出発するかが、バルサフスキーの説では、ある社会の研究が誰の重要性の基準に従うかが、そこに含まれる。後者について、クリモフスキーとモロ・シンプソンは、それが何が研究されるかにはかわっても、認定されるものに客観性が欠けることを含意しないとし、バルサフスキーは、それがどの主張がいつ検査されるかを通じて、真理にいつ到達するかにかかわると応じる。目的が探究を通じて改められるとする説では、帰結を負う者たちがその帰結をどう価値づけるかがどの経路で改訂に届くかも含まれる（第6節）。異論の当事者資格もこれらの関係に属する。異論の認識上の働きとは異なり、人工の構成がそれを内側にもつのは、採る説のもとで部分がそれをもつ場合だけである（第4節）。線がどこに引かれるかは採る説に依存する。重要性が世界の構造によって定まるのなら、どの未知を追究するかは探究の条件の側に属することになるだろう（第6節）。

この仕分けは、配分が人工の構成の内側に入るとき研究に何が起こりうるかを描くための手立てでもある。第3節から第6節に別々に現れた帰結は、いくつかの条件がそろうところでは、一つの連鎖としてつながりうる。その条件とは、同じ評価の信号が次の配分と、どの研究者を複製し引き揚げるかに戻ることに、キッチャーの説のように重要性が世界の構造だけでは定まらないこと、そして周縁に場所をもつ誰も、その信号がなお周縁へ廻るかを見直さないことである。そのとき、問題を分解し課題を選ぶ人工の構成は判定しやすい課題でこそ成功を示しやすく、その成功が次の配分を正当化する（第4節）。その結果、注目は扱いやすくなったところへ移る。AlphaFold2 の後のタンパク質研究や、検証しやすい答えのある領域に集まった強化学習の研究は、配分者を介さなくてもそうした移動が起こりうることをうかがわせる（第6節）。そしてキッチャーの説のように、先の探究が見いだしたことが後に何が重要とされるかを形づくり、しかもエイギがキッチャーの分析から伝えるように、いちど選ばれた方向が慣性をもつなら（第6節）、その移動は見つかる答えにとどまらず、何が重要かの見え方にまで及ぶ。また、選ばれた路線だけが手段と専門性を蓄えるので、目的を書き換える決定がなくても、後に実際に選べる目的が変わる（第6節）。しかも、人間の社会のために研究する人工の構成では、配分が周縁へ廻っているように見えても、廻る先は、*S* の成員がいま立てている問いではなく、構成が訓練された記録を生み出した共同体の立っていたところでありうる（第6節）。そのうえで路線の存続が一つの信号で判断されるなら、その信号の誤りによって、それぞれ異なる路線を捨てる判断が相関しうる（第4節）。研究では、成果が *S* にとって新しいので、信号の誤りを照らし合わせる別の答えがはじめからないこともある（第11章第6節）。これらの環には、人間の研究についての記述やモデルのなかに先例がある。バルサフスキーが記述した主題の「自然選択」や、出版数による選択が方法の厳密さを劣化させるというスマルディーノとマケルリースのモデルなどがその例である（第3節、第6節）。それでも、人工の構成では三つの点が違いうる。一つは、最初の条件が満たされやすいことである。人間の配分でも制度化された信号はすでに舵を取ってきたが（第3節）、時間と資源に対する支配は多くの研究費の保持者に分かれ、仕事が認められる経路も分野によって異なっていた（第5節）。これに対して一つの人工の構成では、一つの信号がすべての路線に同時に届き、研究者の複製と引き揚げまで決める。二つ目は、好まれない路線の保護や寿命による交代など、人間の取り決めが誰も決めないまま連鎖を断ちえた働きを、与えられたものと前提できないことである（第5節）。三つ目は、一つの人工の構成自身の記録が研究の主な記録であるところでは、狭まりがそこに現れにくいことである。追究されなかった路線は、人間の研究でもそうであるように結果を残さず（第5節）、その記録に残るのは、構成が成功した課題での高い成績になりうる（第4節・第5節）。さらに、割り当ての規則そのものが同じ信号に照らして改められていくなら、狭まりは路線の水準だけでなく、配分の仕方の水準でも進みうる（第12章第7節・第8.4節）。この連鎖は避けられない成り行きではなく、上の条件のどれかが欠けるところでは、対応する環で断たれうる。これらの条件のうち最初のものは、人工の構成自身の取り決めのうちにある。どの路線を続けるかの判断が、次の課題を選び、研究者を複製し引き揚げる信号に依らないところでは、この条件は欠ける。第11章第8節が考えられる手立ての一つとして挙げる、生産者を選んだり訓練したりするのに使わない点検を別に残しておけば、その信号そのものを吟味する手がかりにはなる。ただし、それでこの条件が欠けるのは、その点検の判定がどの路線を続けるかに効く場合である。

人工の構成が探究をうまく配分しているという主張が何を示す必要があるかも、この像から引き出せる。そうした主張は、人工の構成にとって利用できる取り決めと比べて検査することはできても、哲学者君主が選ぶであろう分布と比べて検査することはできない。その比較には眼が要るからである（第2節）。研究者を生成し割り当てられる人工の構成は、人間の制度よりも、別の取り決めを並べて走らせやすい位置にあるかもしれない（第2節を参照）。ただし、それらを自らの信号で比べるなら、比べる基準についての独立した点検は、統合だけでは供給されない（第4節）。また、そうした主張は、個々のインスタンスを一つずつ評価した質から、つねに読み取れるわけでもない。それ自体としては不合理でない規則に従う研究者たちも、証拠を持ち寄ることと、ともに劣った選択肢に留まりうるからである（第4節）。そのうえで、そうした主張は、連鎖がどこで断たれているかにかかわる三つのことを示す必要があるだろう。一つは、判定しやすい課題での成功から周縁の継承までの環にかかわるもので、キッチャーの説やそれに類する説のもとでは、人工の構成が舵取りに用いる信号がなお *S* の問いへ廻るのか、それを誰が見直すのかである（第3節、第6節）。二つ目は、同じ誤りで路線が捨てられる最後の環にかかわるもので、人工の構成が依拠する多様性がどこから来るのか、多様性を終わらせる基準についてどのような独立した点検があるのかである。その点検は統合だけでは供給されない（第4節）。三つ目は、人間の取り決めが連鎖を断ちえた働きにかかわるもので、第5節の各働きを、誰かの採る規則が供給するのか、誰も定めない過程が供給するのか、何も供給しないのかである。そこには、人間の好奇心が形づくってきた発見的方法を人工の構成が保つのか、置き換えるのか、失うのかも含まれる（第5節）。保護された予算や異なる初期化といった手立ては、これから試される可能性であり、人間の研究者のあいだのあらゆる違いの代わりになると確かめられたものではない。また523計画は、単一の事例ではあるが、働きを中心で決めることと、その遂行が要する判断を下すこととは別であることを示唆した。第6節が区別した二つの役割のうち、*S* の成員にとって何が重要かを知ることは認識的な役割であり、部分が分散した情報をもつ人工の構成はその多くを果たしうるだろう。これに対して、誰の問いが順序づけの根に立つかを定めること、あるいは誰も決めなかった順序づけについて責任を負うことは、第一の役割における有能さによっては決着しない。誰がそれを引き受けてよいかは、第12章の問いである。

異論の認識上の働きと異論の当事者資格との区別も、こうした主張にかかわる。固有の成功をもつ路線や、それを再開する道を保つことは、配分の内部の特徴でありうる。これに対して、ある当事者が異議を唱える資格を認めることは、内部の部分がそれ自体そうした当事者である場合を除けば、もう一つの関係である（第4節、第12章第5節）。この区別は、よい配分がつねに別々の組織を要することを意味しない。この区別が含意するのは、うまく配分しているという主張と、異論が当事者資格をもつという主張とが、異なる証拠を要することである。前者には、ソロモンの説でいえば、配分が成功をどう代表させ続けるかについての証拠が、後者には当事者のあいだの関係についての証拠が要る。

したがって本章の検討から残るのは、それぞれの目的に照らして判断される選択肢の空間である。どの路線が可能のまま残り、それらがどう支持を得たり失ったりし、違いがどこから生まれ、優先順位をどう改められるのか。将来の結果の不確かさは、この空間を制約するが、一つの最良の取り決めに指定しはしない。ただし、検討した議論からは、この空間のなかで少なくとも三つのことが見えてくる。一つは、多くの路線を作り保つ力と、それらの存続を一つの信号で判断する力とが、一つの人工の構成のなかで結びつきうることである。キッチャーの分析では、統制は個人なら選ばないであろう分業を成り立たせうるので、中央化は多様性を減らすとは限らない（第2節）。しかし、上の連鎖の最後の環で見たように、同じ人工の構成が一つの信号で路線の存続を判断するなら、中央化と引き換えになるのは、多様性そのものより、収束を支配する基準についての誤りの独立性である（第4節）。二つ目は、期待される重要性を使わない探索にも、どの可能性に到達でき、どれに先に到達するかを選択が残り、キッチャーの説のように配分が重要性にはね返るなら、その選択が後の重要性の判断を制約することである（第2節・第6節）。三つ目は、バルサフスキーの議論を当てはめれば、真である結果を生みながら、他所の重要性の基準を受け継ぎ、*S*自身に必要な真理への到達を遅らせる配分がありうることである。真理、重要性、到達の時期は、それぞれ別の次元で問われる（第6節）。配分は批判によって生まれる判断も用いるので、次章は、それらの判断がどう形づくられ、いつ他者がそれに依拠できるかを問う。

第11章 Sのための批判と認定

無効化、独立性、信頼、訂正

1. 本章はどの評価を検討するのか、そしてなぜか

第9章第2.1節は、研究が、特定の探究を遂行する単位の活動を越えて S（達成がそれにとって新しいとされる社会）に結びつけられる三つの関係を導いた。第二の関係である批判と認定は、認識上の成功の評価にかかわる。本章はこの関係を、まず知識の場合について論じる。結果を産出しても、それが S の依拠できるものになるわけではない。三つのうち、結果が知識であり、しかも S の知識になっているとはどういうことかにとりわけ直接にかかわるのはこの関係である。ただしそう言えるのは、知識が根拠づけられることや保有されることの一部を社会的な組織が担うとする説（第2章第6節）を採る場合である（第9章第2.1節）。本章は、人工の構成がこの関係を引き受けうるとき、この関係に何が起こるかを問う。ここでいう人工の構成は、第7章第1節の意味で、研究の遂行が帰属される単位であり、一つのモデル、エージェントの集まり、人間とAIからなる組織などでありうる。第9章第1節の例では、単一のシステムが百のインスタンスを生成し、その一部に他のインスタンスの産出物を点検させ再現させる。そしてそのシステムの構築者たちは、いくつかの主張の一つとして、内部の点検を通った結果は S のために認定されていると主張する。本章は、そうした主張が何を要するか、その証拠のどの部分が人工の構成の内側から来るか、そして内側から来えない部分がなぜ S との関係として残るのかを問う。

本章が検討するのは、研究のさまざまな評価のうち、S に向けられた結果の評価と、それが抛りどころとする批判である。本章は前者を認定と呼ぶ。そう絞る理由は、上で述べたことのほかにもう一つある。人工の構成が評価にもたらす変化が、この評価の抛りどころである評価者の独立性と、評価が S に見えることとを揺るがしうることである。研究における評価は一つではない。人間の研究では、たとえば、資金提供者が結果のまだないうちに提案を評価し、査読者と編集者が原稿を、機関が功績の承認や職のために研究者を評価し、結果を使う者が使うなかでそれを評価してきた。第9章第4節によれば、これらのどの評価にも、目的に結びついた基準がある当事者の選択のために適用されること、評価する者と判定を用いる受け手とが異なりうることなど、四つのことが含まれる。そのうち本章が後で使うのは二つである。一つは、研究の評価が対象とするものは判断が必要な時点ではたいてい観察できないので、代理指標が避けられないことである。もう一つは、S にとって新しいものについては、評価する者がまだ基準をもたないことがありうることである（第8節）。

人工の構成は、評価について少なくとも三つのことを直接に変える。第一に、評価者を募るのではなく生成することができ、しかも一つの基盤システムの複製としても、互いに大きく異なるものとしても生成できる。第二に、結果を生む者と評価する者を一つの人工の構成のなかに置き、評価を結果の形成の途中にも絶えず入れることができる。そのとき批判は、費用だけでなく、範囲、判断、やり方でも変わる。ある種類の点検はほとんど費用をかけずに行えるようになり、人間がふつう行わなかった点検が試みられるようになり、評価者は人間とは違うところで誤り、違うものを拾いうる。そして評価が産出のなかに溶け込み、「査読を通った」という時点そのものがなくなって、評価が S から見えなくなることもありうる。第三に、評価者と生産者とを時間にわたって結合し、それぞれを他方の出力に合わせて変えていくことができる（たとえば他方の出力で訓練したり選抜したりして）。第3節、第7節、第8節が、この三つを順に取り上げる。

他の評価は別の箇所でも扱う。探究の路線の事前の評価は配分の一部として第10章で、功績と貢献の帰属は第7章第3.3節で、人工の研究者についての主張の評価は第7章第5節と第13章で、そして評価が適用する基準と誤りの重みを誰が定め改めてよいかは第12章で扱う。ただし配分は批判から生まれる判断も用いるので、第10章が見た配分の判断も、その一部はここで扱う批判に依っている。

第3章は、発見の文脈と正当化の文脈の区別と、それをデュエデ（Duede）が深層学習に当てはめた議論から、ある限られた容認を引き出した。候補となる答えを生成する経路は、それを受け入れることを根拠づける経路と同じである必要はない。生成の過程が不透明でも、受け入れる根拠を別の経路で与えられるなら、その答えを用いてよい、ということである。信頼性主義の説や徳認識論の説は、この容認が必要になる範囲を狭める。それらの説では、生成の源泉そのものが信頼できれば、別の経路を立てなくても根拠が与えられるからである。第3章は条件も付けた。この容認が成り立つのは、候補が適切な独立の経路を通じて評価される場合に限られ、その独立性が本物かどうかを判断するには、評価者の共有、選択、汚染（評価に使う問題や答えが、評価される側の訓練データに入っていること）、隠れた人間による選別についての証拠が要る。その議論はひとつの結果にかかわっていた。人工の構成が研究者だけでなく、査読者、追試者、批判者も生成できるとき、批判そのものが構成の内部で組織されうるものになる。以下で査読、追試といった名を人工の構成の部分に用いるとき、それは人間の制度が束ねてきた働きを指すのであって、その制度の形が残ることを指すのではない。それがどのような働きなのかは、第7節で問う。

本章は二つの関係を区別する。無効化の関係と認定である。まず前者から始め、次の三つの点検を考えよう。候補となる材料について予測された性質が、その予測に頼らずに値を出す装置で測定される。生成された証明が形式的な規則に照らして点検される。出来事より前に登録された予測が、実際に起こることと比較される。どの場合にも、候補やその根拠は反対の証拠に出会いうる。しかもその証拠は、候補を生み出した過程が候補を受け入れる理由を、そのまま繰り返したものではない。そうした証拠上の経路を無効化の関係と呼ぶ。この語が指すのは、結果が実際に覆されるという出来事ではなく、候補が誤っていればそれを示しうる経路があることである。点検を通った候補も、そうした経路に直面していたのである。第2章は、主張を反駁する証拠と、その根拠を掘り崩す証拠とを分けた。第8章第2.3節は、標的のあり方が違えばコミットメントも違ってくるという仕方でも認識的組織が標的に制約され、さらなる接触を通じて是正可能であり続けるとき、それが世界への応答可能性をもつとした。無効化の関係は、そうした訂正が生じうる経路を特定する。ある

測定、形式的な点検、比較がその経路を与えるかどうかは、その名前ではなく、何を試し、その結果が主張にどうかかわるかに依存する。この語は、不可謬の点検も、あらゆる偽の候補が必ず失敗するという保証も含まない。

そうした経路を生産者の手の届かないところに置くことは、経路を守る一つの仕方である。しかし、守られていることは、経路がどれだけの証拠を加えるかを決めるものではない。たとえば、点検の手続きを書き換える許可をもつプログラムでも、書き換えていないバージョンの点検に候補を出し、その点検が生成過程の見逃す誤りを見つけることはありうる。許可があることだけでは、その情報は失われない。逆に、プログラムが変えられない手続きが、生成者の誤った仮定を繰り返すこともありうる。その保護だけでは、情報は何も加わらない。この二つの場合が示すように、点検の証拠上の寄与と、それを維持する仕組みとは別の事柄である。第2節は前者を展開し、第8節は、その寄与が適応、選抜、点検の変更を通じていかに存続しうるかを問う。

失敗には解釈も必要である。第6章第2節は、反対の結果がどのコミットメントを改める理由を与えるかを定める仕事を、*局所化*と呼んだ。反対の測定は、候補、補助仮定、較正、形式的な言明と問いとの対応のいずれにもかかわりうる。生産者が失敗を正しく局所化することもあり、局所化を別の当事者に割り当てるだけで、その当事者が正しくなるわけではない。ここで重要なのは、診断が、候補を保持したいという願いを超える裏づけをもち、それ自体もさらなる証拠による異議に開かれているかどうかである。あらゆる不利な結果を、そうした裏づけなしに別のものへ帰属させられるなら、測定は情報を与えたままでも、候補を受け入れるかどうかの判断はそれに応答しなくなる。証拠の保管を分けること、改訂を記録すること、争われる診断をめぐる権限を共有することは、この応答可能性を保護しうる。その価値は、それらがどのような異議を可能に保つかに依存し、情報を与えるあらゆる点検が満たさなければならない普遍的な条件ではない。

もう一つの関係は**認定**である。他者が特定の用途のために結果に依拠する理由を与えられる関係である。認定は S にかかわる。認定は、 S のなかで誰がどの行為のために結果に依拠してよいかにかかわり、ロンジーノの批判的文脈の経験主義（第3節で詳しく見る）では、結果がどの視点から批判されたかにもかわる。ピューターが引用する定義に明示されているように、受容可能性は、共同体に利用できる視点に相対的だからである (Longino 2002, 135–136, as quoted in Büter 2012, 84)。人工の構成が強い無効化の関係を内部で確保していても、その認定が、認定したものに S が依拠する理由を与えるかどうかは、それでは決まらないことがある。上で挙げた三つの変化は、無効化の関係と認定の両方に及ぶ。本章は、前者のどこまでを人工の構成の内側で確保でき、後者がなぜ S との関係として残るのかを問い、さらに、時間と評価の見えにくさが、内側で確保できた前者をも外へ開き直させるかを問う。

無効化の関係そのものが S にかかわるかどうかは、採る説に依存する。批判の価値を、それが検出する誤りに置く説では、無効化の関係は単一の行為者の内側でも成り立ちうる。数学者は、あるいはアルゴリズムは、証明を提案し、それを提案者が統制していない規則に照らして点検させることができ、その関係のどこにも S は現れない。ロンジーノの説では、証拠は背景仮定を通じてのみ仮説にかかわる (Longino 2025, §5)。そのため、ある測定が何に不利に働くかは、共同体が抱く仮定に相対的である。その説では、無効化の関係が S にかかわらないという主張が成り立つのは、その結果を読むときに通す背景仮定そのものが、 S が異を唱えるであろう仮定に含まれない限りにおいてである。

この分岐は、第2章第6節が引いた分かれ目に沿っている。そこでは、探究が社会的実践に因果的に依存するという主張が、証拠としての地位、客観性、あるいは知識が、組織された社会的実践によって部分的に構成されるという、より強い主張から分けられた。そして本書は、社会的知識のどちらの説も採らなかった。したがって本章の問いは二通りに読める。第一の枝では、批判が重要なのは誤りを検出するからであり、問いは、何がある点検を別の点検から独立にするかである。それが第2節の問いである。第二の枝では、組織された批判が客観性や根拠づけを部分的に構成し、問いは、そうした批判の組織が批判者のあいだにどのような種類の違いを要するかである。第3節は、この枝の説をいくつか、第一の枝の見方と並べて比べる（表11.1）。第二の枝で主題になるのは、組織された批判によって部分的に構成されるという意味での客観性である。第一の枝の見方が、客観性を別の意味で語ることを妨げるものではない。

本章はところどころで、第9章第6節の仕付けを使う。第9章第2.3節は、研究を取り巻く関係が別々の当事者に分けて担われてきたことが、誰も意図しないまま、いくつかの財をもたらしていたかもしれないと述べた。第6節の仕付けは、そうした財のそれぞれが何に依っていたかを、四つの種類に分けて問うものである。一つは、人間という実現者についての前提、すなわち実際に関係を担っていた人々そのもの（一人の注意、動機、寿命など）についての前提である。二つめは、現在の科学のシステムの組織についての前提、すなわち関係が歴史のなかでどう組織されてきたか（一次データに近づけるのは研究者だけであることや、査読という関門など）についての前提である。こちらは、実現者が人間のままでも組み直されうる。三つめは、探究の条件であり、何がそれを実現するにせよ成り立つ。四つめは、人工の構成の外の当事者との関係であり、本章も第9章にならって、これを略して S との関係と呼ぶことがある。

第2節は、誤り検出の見方のもとで、独立性を点検に加える情報として定める。第3節は、批判者を作れるとき違いがどこから来るのかを、この見方を含む五つの族（表11.1）がそれぞれどう評価するかで見る。第二と第三の変化を取り上げる前に、第4～6節は認定の側に移る。第7節と第8節がそれらの変化について問うことは、そこで示す閾値と依拠の根拠を用いて述べられるからである。そのうえで第4～6節は、第1節の二つの枝に二つの問いを加える。結果をある用途に足とする閾値が誰の価値づけに答えるか（第4節）と、点検していない結果への依拠を何が根拠づけ、その依拠が信頼でなければならないか（第5節）である。第6節は、 S 全体が人工の構成に対して素人の位置に立つ極限で、何が S に残るかを問う。第7節は、生産と評価が一つの人工の構成で結びつくとき、批判の費用、範囲、判断、やり方がどう変わり、評価が産出に溶け込んで見えなくなるときの何が失われうるかを問う。第8節は、時間を通じた結合が、批判者の独立性、新しいものを評価する基準、訂正に何をもちらすかを問う。第9節は構築者の主張に戻り、これらの説を組み合わせて考える。先に見通しを示しておく。人工の構成は、説がそれぞれ追う性質（点検どうしの独立性、誤ったときに答える誰がいること、誤りの重みについての価値づけを産出者と依拠する者が共有していること、評価がなされたことが S に見えること）を別々に設定できるものにする。人間の研究ではそれらが実現者と組織によって結びつけられていたので判定が重なりやすかったと考えられるが、人工の構成では同じ事例について説の判定が分かれ、構築者の主張は、どの説に答えるのかを示さないかぎり評価できなくなる。そして、いくつかの説のもとでは、内部の批判をよくすることが、 S のための認定をかつて弱めうる。

ただし、これらの判定は、それ自体が本章の行き着く先なのではない。説ごとの判定は、批判と認定が人工の構成の内側に入ったとき研究に何が起こりうるかを描くための手立てでもある。本章の材料から描ける像を、先に挙げておく。批判者が形成されるのではなく作られるとき、証拠を加えるのは数では

なく違いであり、仮定を共有する批判者の集まりは、ロンジーノの説では、内部の批判がよく働くほどその仮定を定着させうる（第2節、第3節）。評価が産出に溶け込めば、「査読を通った」という時点がなくなり、評価は S から見えなくなる。そのとき手軽に増える批判は、生産者と同じところで誤る限り、独立した批判の乏しさを覆い隠しうる（第7節）。評価者は、選抜に使われるだけで、誰も欺かなくても生産者について教えるものが減っていきうる。 S が結果を判断する基準そのものも、生産と評価の結合とともに動きうるが、 S のいまの基準と一致していることは、基準が動いていないことの証拠にならない（第8節）。誤りの重みづけは構成の内側で前もって形づくられ（第4節）、認定は研究者の主張というより道具の示す値として扱われうる（第5節）。他方で S の成員は、評価する経験を積む場を失うにつれて、残された関係を担う力そのものを失いうる（第6節）。これらが重なれば、内部の点検を通ったことが認定の目安として使われ、 S との関係の欠けがそれとともに選ばれていき、形の上では S との関係が残りながら、それを担える者がいない状態に研究が近づくかもしれない（第9.3節）。以下の説ごとの検討は、どの説がこれらのどれを問題とし、それを防ぐには何が要るかを分けるためのものでもある。

2. 誤り検出の見方における独立性

第1節の第一の枝では、批判の価値は、それが誤りを検出することにある。この枝で独立性を問うとは、ある点検が、他の点検のまだ見つけていない誤りを見つけうるのは何によってか、を問うことである。この問いが本章で重要になるのは、第1節で挙げた三つの変化が、いずれも評価者の独立性を揺るがしうるからである。二つの点検が同じ判定を下したとき、二つ目はどれだけの証拠を加えるのか。専門家の証言についての文献は、複数の情報源が同意する場合について、この問いに精確な答えを与えてきた。本節はそこから始める。

ニコラス・J・タンチャック（Nicolas J. Tanchuk）は、アルヴィン・ゴールドマン（Alvin Goldman）の最小の条件を報告している。第二の源からの同意が証拠を加えるのは、その源が第一の源から少なくとも部分的に条件つきで独立であるとき、すなわち第一の源が何を言ったかを所与として、主張が真であるときのほうが偽であるときより同意しやすいときに限られる（Tanchuk 2026, 4）。ゴールドマンとトマス・ブランチャード（Thomas Blanchard）も、指導者に無批判に従う者をいくら増やしても、独立の支えは加わらないと説明する（Goldman and Blanchard 2015, §3.2）。この条件は、人や組織について何も述べていない。二つの判断が誤りうる仕方のあいだの関係にかかわっている。ゲレロ（Guerrero）は、互いの見解や共通の第三者の見解を映すだけの源からは、必要な独立性は得られないというゴールドマンの注意を報告している（Guerrero 2017, 177）。また独立性は、第二の源が第一の源を迂回することを要しない。タンチャックが引くゴールドマンの例では、第二の源は第一の源の理由を聞き、第一の源が考えなかった、その理由を打ち消しうるものを検討し、それらを退けたうえで同意する。そのような源も、部分的には独立でありうる（Tanchuk 2026, 5）。

本書は、この見方の内部でこの条件を拡張する。拡張のもとでは、ある批判者は、所与の主張について、既存の批判者がすでに与えている以上の誤り検出の情報に加える程度に独立である。これは、ゴールドマンが情報源の同意について引いた区別を点検一般へ適用するものであり、そうした点検についてのゴールドマン自身の説ではない。この拡張に立てば、組織の分離は必要条件でも十分条件でもない。必要でないのは、単一の人工の構成が、証明を見つけた過程とは異なる故障の仕方をもつカーネル（証明支援系の中核で、推論の各段階を検査する小さなプログラム）で証明を点検することも、予測を自らが統制しない測定に照らして試すこともできるからである。十分でないのは、異なる組織に属し異なる名前をもつ評価者たちが、同じ訓練、同じ資料、同じ評価の基準、よい結果とはどういうものかについての同じ考えを共有し、そろって誤りうるからである。第9章第5節が挙げた権能、たとえば証拠を別に保持することは、別の保管のもとにある装置が切り離されているのと同じ仕方で、批判の経路を切り離しておくことができる。ここでも、それらを切り離しておくことが重要なのは、それによってそれらが異なる失敗の仕方をもつままでいる程度においてである。他の部分の更新を止める権能も、この見方にかかわりうる。点検する部分が生産者の出力に合わせて更新されていくこと（第8節）を止める手立てになるからである。十分でない場合の一つの例を、現在の言語モデルについての証拠を与える。第10章第4節で触れた誤りの相関は、ある時点で調べられた言語モデルのうち、より大きく正確なものあいだでは、提供元や基盤のアーキテクチャが異なっても持続していた。名前と組織が分かれていても誤り方を共有するという、ここで述べた種類の失敗である（Kim et al. 2025, abstract）。同じことは、同じモデルを利用する人間の評価者にも成り立ちうる。コーパス全体の水準での推定は、2023年と2024年に開かれた四つのAI分野の会議の査読の本文のうち6.5～16.9%が、言語モデルによって実質的に修正されていた可能性があることを示唆しており、推定された割合は、同じ論文の他の査読とより似ている査読ほど高かった（Liang et al. 2024, 1, 9）。この見方では、そうした共有された利用が見えないところでは、批判者が形式的に分かれていても、それだけでは、誤り方を共有しているかどうか、どの程度共有しているかは分からない。

第1節の意味での無効化の関係は、候補やその根拠に異議を突きつける経路を点検に与える。この誤り検出の見方で問題となる独立性は、他の点検がすでに与えているものに対する、情報上の寄与である。内的な整合性や評価者たちの同意だけでは、その寄与は確立されない。世界との接触や形式的な点検は、一つの人工の構成の内側でも、それを与えうる。点検が確立するものにも限界がある。点検された証明が確立するのは、その形式的な言明に相対的な定理であって、その言明が問いを捉えていることではない（第6章第3節）。

さらに、次の対照を構成してみよう。ある点検が候補を退け、生産者は、校正の誤りがその原因だったと提案する。一方の場合には、別に確立された参照標準に照らした測定がその診断を支え、校正が訂正され、影響を受けた候補が再評価される。他方の場合には、校正を変更する唯一の理由は、それが好まれた候補を退けたことである。前者は点検を改善しうるが、後者は候補を受け入れる追加の理由を与えない。違いは、生産者が改訂に加わったかではなく、何が改訂を支え、改訂後の判断に反するものとして何がなお働きうるかにある。記録はその違いを明らかにしうるが、変更を記録しただけでは、その改訂が支えられたことにはならない。第6章第5節は、探究の内部での基準の改訂を展開する。ここでの点は、それが認定への依拠にもつ帰結である。

したがって、独立性、保護、厳しさは、異なる問いに答える。点検は改訂に開かれていても情報を加えうる。保護は、その寄与を証明することなく維持を助けうる。保護され、情報を与える点検でさえ、検出する誤りの範囲が狭すぎて、その点検を通過したことかなされる主張を根拠づけられないことがある。メイヨーが厳しいテストと呼ぶもの（第6章第2.3節が Mayo 1996 を引いて論じた）は、問題となる誤りが存在するならそれを高い確率で露わにするテストと言い換えられ、このさらなる問いにかかわる。したがって、これらの性質のどれか一つが確かめられても、そこから適切な認定は出てこない。

この見方が何を拾い上げるかを、四つの事例が示す。それぞれの事例は別々の点のために導入する。最初の二つは、独立の経路を供給した関係の分割と、供給しなかった分割との対比をなす。2020年に金星の大気中のホスフィンが報告された後、外部のあるグループは公開されたスクリプトを再実行してデー

タ縮約を批判し、別のグループは校正の設定の違いを観測所に知らせ、観測所の職員が問題を特定してアーカイブのデータを処理し直した (Snellen et al. 2020; Villanueva et al. 2020; European Southern Observatory 2020; Greaves et al. 2020)。この事例は、独立の経路が生産するチームの外にありうることの例になる。問題を見つけたのは、自ら解析を行い、アーカイブと校正の記録を保持していた当事者たちだった。対照的に、ハッブル宇宙望遠鏡の欠陥のある主鏡の調査は、主鏡の製作を担った部門が審査から切り離されており、製作の過程で誤りの兆候が割り引かれ、計画が単一の検査装置に全面的に依存していたことを見いだした (Allen et al. 1990, iv, 9-1)。第6章第3.2節は、同じ事例を用いて、通った点検がそれ自体誤った基準に依拠しうることを示した。ここでこの事例が示すのは、点検が多数あってもなお一つの誤りの経路を共有するという点である。主鏡の誤りを示す検査もあったが、その結果は、計画が全面的に頼っていた検査装置と食い違うという理由で、その装置を基準に割り引かれたのである。二つのエピソードは、関係の分割がそうした検出をどれほどの頻度で供給するかを示すものではない。

第三の事例は、役割の配置によって独立性を設計できることを示す。18世紀の『航海暦』(Nautical Almanac) では、月の位置の表を除き、各月の表を二人の計算者が別々に計算し、第三の人物が照合した。二人の計算者が独立に作業せずに互いの仕事を写していたことを照合者が見つけると(遅くとも1770年1月までに)、二人とも解雇され、数年後に再雇用されたときには同じ月を割り当てられなかった (Croarken 2002)。この配置に価値があったのは、二つの計算が別々の誤りの経路をもつ限りにおいてであり、写すことは、役割の名を変えないままそれを取り除いた。第四の事例は、人工のパイプラインにおける逆の場合を示す。パイプラインを名前のついた役割に分けることは、その点検を独立にはしない。ある自律的な材料研究室についての独立の批判は、その回折データの自動解析はまだ信頼できず、批判者たちの再解析では報告された材料のどれも新しくはないと論じた (Leeman et al. 2024, abstract)。著者たちの後の訂正は、手作業の再解析を経て、報告された40の合成のうち36を確認し、4つを決め手なしとした (Szymanski et al. 2026)。両者が扱うのは一部異なる主張、すなわち合成できたかと新しいかである。十の課題にわたるあるベンチマーク分析は、ひとつの自動研究システムが生成したすべての論文を、捏造された実験結果を含むものとしてコード化した (H. Chen et al. 2025, app. B, fig. 6)。どちらの場合も、システムは自動の点検を含んでいた。重要だったのは、点検がそれ自身の誤りの経路をもっていたかどうかである。

一つの事例のなかの異なる主張が、異なる経路によって決着することもある。理化学研究所がSTAP現象を再現できるかを検証したとき、丹羽仁史が率いたチームによる検証は、得られた細胞塊が元の研究によるSTAP細胞の定義を満たさないと結論した (Niwa 2016)。一方、外部の専門家からなる別の委員会は、保存されていた細胞株のゲノムを既存の胚性幹細胞株のゲノムと比べ、保存されていた株はそれらの株に由来すると結論した (RIKEN Investigation Committee 2014, 13)。この見方では、点検が独立かどうかは、調査全体の性質ではなく、ある主張に相対的な、その点検の性質である。

批判の価値をそれが検出する誤りに置く見方では、点検の独立性は、批判者の数や組織上の分離ではなく、それが加える情報によって主張ごとに判断される。そうした寄与は人工の構成の内側でも生じうる。残る課題は、その寄与がどこまで及ぶかと、生産者、点検、失敗の解釈が変わってもそれがどう保たれるかを示すことである。次節は、批判者を作れるとき違いがどこから来るのかを問い、一つの基盤から多数の批判者が作られる極限と違いが作られる場合とを、この見方と他の説がそれぞれどう評価するかを見る。

3. 批判者を作れるとき、違いはどこから来るのか

第2節は、何がある点検を別の点検から独立にするかを問うた。人間の研究では、批判者は別々の人として別々の来歴をもって形成されるので、その数と組織上の分離が、独立性のおおよその目安として働いてきた。独立性の多くは、誰かが設計したものではなく、人間という実現者のあり方から供給されていたのである。第1節で挙げた変化の第一は、この前提を崩す。人工の構成は、批判者をいくつでも、どのような組み合わせでも作りうる。同じものの複製を多数作ることも、大きく異なるものを作ることもできる。そのとき、数と分離はもはや目安にならず、批判者のあいだの違いがどこから来るのか問われるものになる。本節はまず、一つの基盤から多数の批判者が作られる場合を極限まで押し進めて、何が加わり何が加わらないかを説ごとに問い、そのうえで違いが作られる場合に移る。同じ問いは、人間の研究でもなじみのある形をとる。一つの下稿の査読者たちは同じ文章を、しばしば似た訓練のもとで読むので、査読は冗長ではないかという問いが生じる。冗長さが欠陥かどうかは、何が重複しているかに依存する。査読者どうしの重複とは別に、産出する側がすでに行った評価を査読が繰り返すという意味での冗長さもある。それは第7節で扱う。

本書は冗長さの四つの意味を区別する。この区別は、冗長さについての主張が、ある意味を確立しながら別の意味に頼ることのないように、本章のために行うものである。区別は第2節の条件と下に引く文献に拠り、さらなる批判者が何を加えるかによって意味を分ける。他の批判者が与える以上の何も加えないか、それ自身の誤りの経路を加えるか、共有された基準の異なる適用を加えるか、異なる受け手や働きを加えるか、である。冗長さの意味はこの四つに限られない。

1. **相関した批判者による重複。** 同じ訓練を受けた批判者が同じ記録を読む。ゴールドマンの条件では、判別をせずに源を写す複製は、その源を越える証拠を加えない。相関があるだけでは、そうとは限らない。批判者は一部の影響を共有しながら、主張についての情報を加えうる。重要なのは追加される情報であり、完全な統計的独立ではない。
2. **独立した誤りの経路の増加。** 異なる原理で働く点検、別の保管のもとにある測定、形式的な検証などである。記録を読み直す批判者は、計算の誤り、統計的推論の混乱、証明の欠落のように、記録をもとに行われたことの誤りを見つけうる。そうした批判者は、生産者と同じところで誤らない限りで、検出を加える。記録の作られ方の誤り、たとえば汚染された試料は、記録に痕跡を残さないことがあり、そのとき批判者がそれについての情報を加えるには、新しい測定を行うか実験をやり直すしかない。記録の読み直しと新しい証拠とのこの対比は、陪審定理の研究に見られる。フランツ・ディートリッヒ (Franz Dietrich) とカイ・シュピーカーマン (Kai Spiekermann) の言い方では、実験などによって新しい証拠を独立に生み出す成員で集団を大きくすれば、加わるのは「情報の解釈ではなく情報」である。ただしそれが成り立つのは、異なる成員の証拠に隠れた共通の影響が及ぶことを排除した場面においてだけであり、二人はその排除をもっともらしくないとしている (Dietrich and Spiekermann 2021, §4.3)。第一の枝では、価値をもつのはこの冗長さである。
3. **判断の多様性。** 基準を共有する批判者たちが、その適用と重みづけにおいて異なる。バードの読みでは、トマス・クーン (Thomas Kuhn) は、正確さ、整合性、範囲、単純さ、実り豊かさといった科学で共有される価値は、規則のように判断を決めるのではなく判断を導くものであり、そのた

めそれらを共有する科学者どうしても理論の選択が分かろうと考えた (Bird 2025b, §7.3)。バードの紹介によれば、クーンは、そうした違いが、個人の生い立ちや人柄などの要因や、ドイツ・ロマン主義のような科学の外からの影響から来うことを認めた。ただし、近代の職業化した科学では、そこで働く要因はたいてい科学の内部から来るとし、ある科学者が別の科学者より危険を避けたがるといった人柄の違いでさえ、科学的な証拠への関係であることに変わりはないと考えていた (Bird 2025b, §7.3)。ゾルマンが引用しバードがまとめる箇所によれば、その説では、新しい理論がふつう、集団に受け入れられる前に何人かの科学者によって展開され試されるという進み方は、合理的な不一致を許す決定過程を要する。共有されたアルゴリズムは、それに従うすべての科学者に同じ時に同じ決定をさせるだろう。そして価値とみなされた規準のゆるさは、新しさを導入し支持する際のリスクを分散させるために不可欠だとされる (Zollman 2010, 2; Bird 2025b, §7.3)。クーンの説では、この冗長さの価値は誤りの検出ではなく、新しきの導入や支持が伴うリスクを、集団のなかに分散させることにある。

4. 重なり合う評価の段階。評価は、結果が産み出されてから使われるまでのあいだに、時点や受け手を変えて重ねて行われう。人間の研究では、出版前の査読、出版後の批判、追試、引用、下流での使用などが、そうした段階をなしてきた。人工の構成が同じ段階をもつとは限らないが、評価を重ねる取り決めをもつなら、同じ問いが生じる。段階が異なる受け手と働きをもつところでは、それらは見かけほど冗長ではない。どの段階も同じ記録の上で同じものを読み、先の段階と同じところで誤るなら、それらは第一の意味に戻る。

最初の二つの意味は、第2節の見方が捉えるものである。第三の意味はそれに捉えられない。判断の多様性は、新しい誤りの経路を何も加えなくても、クーンの説では、新しさがどう担われるかにとって重要でありうるからである。クーンの論証には限界もある。バードの読みでは、この論証は、大きな共同体では個人の判断を左右し、人によって異なる要因が互いに打ち消し合うこと、すなわち個人差が正規分布をなし、分布の平均にあたる判断が、伝統的に理解された科学的方法の規則が求めるであろう判断に一致することを想定している (Bird 2025b, §7.3)。クーンは複製の集団を考えていないが、当てはめれば、一つの基盤システムから派生した批判者の集団は、数は大きくても、クーンの論証が頼る個人差の広がりや欠きうる。

一つの分野からの証拠が、最初の二つの意味にかかわる。ヴァレンティン・ダンチェフ (Valentin Danchev)、アンドレイ・ルジェツキー (Andrey Rzhetsky)、ジェームズ・エヴァンズ (James Evans) は、薬剤と遺伝子の相互作用についての公刊された主張を、ロボットが行う高スループット実験の結果と照合した。ある主張についての知見どうしは、著者を共有する論文から来るとき、社会的に独立したチームから来るときよりも互いに一致しやすかった。他方で、複数の論文に支持された主張のうち、異なる方法を用いる先行文献に拠る独立したチームが確認した主張は、実験でより再現されやすかった。社会的な独立がないところでは「著者と方法が自らの一致を強める」と三人は書く (Danchev, Rzhetsky, and Evans 2019, 4, 6–9)。これらの関連は相関であり、生物医学の一つの領域から得られたものである。上の語で読めば、重なり合うチームのあいだの一致は部分的には第一の意味での重複であり、独立したチームによる確認は、第二の意味が数えるものを加える。ただし、この研究が扱うのは人間のチームである。その依存の尺度のうち、方法の共有と先行文献の共有は、共通の手続きを適用する人工の批判者のあいだにも、共通の記録で訓練された人工の批判者のあいだにも対応物をもつ。それらが一つの基盤システムから派生しているかどうかにはかかわらない。そうした対応物に同じ関連が見られるかどうかを、この研究は検証していない。

出口康夫 (Deguchi) は、第二の意味のための論証を、個々の批判者ではなく探究者の種類の水準で与える。19世紀の精密測定学に依拠して、系統誤差、すなわち測定器に固有の偏りは、異なる測定器が同じ対象を測ったときにはじめて現れ、できるだけ多くの異なる方法で同じ対象を測り、その平均をとることで打ち消される、と指摘する。そして人間を一種の認知の測定器とみなし、人間が探究者であるというだけでも偏りは、探究者がすべて人間である限り、その数がどれほど多くても見えるようにならないと論じる。人工の探究者は、それが人間の探究者と食い違うところで、その偏りを見る機会を与える。人工の探究者にもそれ自身の偏りがあるので、その食い違いは両方を指し示す。出口は、人工の探究者を人間とは異なるままにしておくことを求める。その見方では、人間だけによる探究では、結果に残る人間の偏りは見えないままであり、認知を人工の探究者に完全に委ねれば、結果はそちらの偏りにさらされる。両者の偏りを打ち消す単純な平均のような手立てはないので、両方の種類が、打ち消す過程そのものも含めて対話し、違いをその都度確かめていく必要がある。また、人間に似せすぎた人工の探究者は人間の偏りまで受け継ぐので、それを見えるようにするはずの異なる視点が失われる (Deguchi 2026b, スライド4–12)。出口の論証は、人間と人工システムが共同で行う探究を対象としており、その類比は、異なる測定器の偏りが、そこに見いだされる値のまわりに散らばることを前提している。この前提は、世界がどうなっているかを問う場面にはよく合う。そこでは、偏りを打ち消していけば測られる値に近づくと考えられ、これは世界への応答可能性 (第8章第2.3節) にあたる。しかし、何が S にとって重要かを問う場面には、そのままは合わない。重要さは、異なる種類の探究者がそこへ収束していく一つの値ではないからである。そうした場面に当てはめれば、種類のあいだの食い違いは、打ち消される偏りではなく、立場の違いを明らかにするのかもしれない。出口の論点は、下で見るロンジーノの、共同体の全員が共有する仮定はその共同体には見えないという論点と同じ形をとる。違いは、それを視点ではなく探究者の種類の水準に置き、手立てを、偏りの打ち消し方も含めた、異なる種類の探究者のあいだの対話に求める点にある。

異なる説が何を含意するかを見るために、第9章第1節の例を極限まで押し進めた思考実験を考えよう。その動機は、人間の研究の一つの前提、すなわち批判者は異なる来歴によって別々に形成されるので、その数がおおよそ独立性の程度を示すという前提を取り出すことにある。思考実験はこの前提を取り除き、他のすべてを残す。この構築を、マイケル・ポランニー (Michael Polanyi) の「科学の共和国」(第10章第2節、第12章第3.2節)と区別するために、**インスタンスの共和国**と呼ぶ。百万の人工の研究者が単一の基盤システムから派生し、研究、査読、追試の役割に割り当てられる。極限的な事例として、その後の経験によっても、主張に関連する背景仮定が多様化しないと仮定する。共通の由来だけでは、この追加の前提は確立されない。人工の構成は、どの結果でも批判できる公開の場を維持する。批判には記録された応答が与えられ、インスタンスの信念、あるいはその役割を果たすコミットメント (第2章第2節) の分布が、批判的な議論に敏感な仕方でも時間とともに変わることを要求する (Longino 2025, §5)。批判者が訴えることのできる共有の基準を組み込み、すべての行為者に異議を唱える等しい知的権威を与える。共和国をそれだけで一つの共同体とみるなら、形式の上では、この共和国は、第2章がヘレン・ロンジーノ (Helen Longino) の客観性の説から導入した条件のうち最初の三つ、すなわち批判のための承認された場、批判の受容、公的な基準を満たし、あわせて第四の条件である知的権威の調整された平等 (どの視点も共同体の批判的な相互作用に寄与する一応の能力をもつものとして扱うこと。この平等は、批判にかかわらないことや批判に応じないことで失われうる) のうち、誰が異議を唱えてよいかにかかわる部分を満たしている (Longino 2025, §5)。同じ条件が求める視点の多様性を満たすかどうかは、下で見るように別の問いである。その批判は適切だろうか。

答えは、説が批判の適切さをどこに置くに依存する。表11.1は、答えの五つの族を示す。最後の一つは説ではなく、歴史的に跡づけられた理想である。各行は単一の理論ではなく族である。本書がこの五つを選ぶのは、それぞれが共和国に何が欠けているかについて異なる答えを与えるからであり、この選び方は、第1節の第一の枝の見方（第2節で展開したもの）を、基準線としてではなく、他と並ぶ一つの行として扱う。答えはこの五つに尽きるわけではない。各行は、第一列に名を挙げた論者を通じてその族を示す。右の列は、本章が共和国について引き出す判定をまとめたものである。いずれの行でも、その判定は、論者たち自身が考察しなかった人工の構成に説や理想を当てはめて得たものである。

表11.1 批判の適切さをどこに置くか、そこからインスタンスの共和国について何が見えてくるか（単純化した見通し）

族または理想	適切さが置かれる場所	共和国について出てくること
誤りの経路（第2節で拡張したゴールドマンの条件）	さらなる点検のそれぞれが、他の点検の与えない誤り検出の情報を加えること	インスタンスの数はあまり意味をもたない。共有された誤りの経路は、多くの批判者を一人に近いものにする。関連する情報を加える点検は、人工の構成の内部に維持されるものも含めて、批判を改善しうる
批判的文脈的经验主義（ロンジーノ）	四つの条件を満たす共同体のなかでの、関連する視点のあいだの変容的な批判（ある視点の背景仮定を、それを共有しない視点が問い直し、受け入れられた内容を変えうる批判）	手続き上の条件は満たしうるが、すべてのインスタンスが共有する仮定は守られたままであり、強化されることもある
社会的经验主義（ソロモン）	判断に働く影響（決定ベクトル）が理論のあいだに公平に分布すること（経験的なものは各理論の経験的成功に比例し、非経験的なものは均等に）。合意が適切なのは、一つの理論がすべての経験的成功をもつときである	失敗は、欠けた視点よりも、すべてのインスタンスが共有する影響にある。集団は個人のバイアスを増幅しうるし、合意に向かう圧力は集団思考や関連するデータの抑え込みにつながりうるので、内部の熟識を増やすと悪化しうる
立脚点理論（ハーディング、インテマン）	共同体が周縁化してきた人々の生から出発する探究と、集団的な意識を通じて達成される立脚点	複製だけでは、関連する立脚点は与えられない。周縁化された人々の生との関わりと集団的な意識の達成は、別途確立されなければならない
機械的客観性（ダストンとギャリソン、ポーターがたどったもの）	探究者の意志と判断を結果から締め出し、結果が誰によって産出されたかに依存せず、個人への信頼に頼る必要がないようにする手続きと規則	一様に規則に従うインスタンスは、この意味で客観的とみなされうる。この意味が防ぐのはどれか一人の批判者の偏りであって、全員が共有する仮定ではない。問いは、誰が手続きを設計したかへ移る

ここでロンジーノの説を展開するのは、第2章第6節がそれを、社会的な組織が根拠づけに入りうる重要な仕方の一つとして導入し、共和国がそれを直接に試すからである。その説では、証拠の関係は背景仮定に媒介される (Longino 2025, §5)。そしてハイディ・グラスウィック (Heidi Grasswick) が伝えるところでは、ロンジーノは、同じような位置にある知者の集団に共有されると見えなくなりうる疑わしい背景仮定を明るみに出すには、共同体に多様な視点が代表されていることが必要だと認めている (Grasswick 2018, §4.3)。ドゥティル・ノヴァエスとデデが伝えるロンジーノの第四の条件では、実効的な批判には視点の多様性が必要であり、異論は単に許容されるのでなく育てられなければならない (Dutilh Novaes and Dede 2025, §2)。ビュターが引用するロンジーノの受容可能性の定義では、支持するデータがその時点で共同体にとって明らかであり、推論と背景仮定が、その時点で利用できる限り多くの視点からの批判を生き延び、共同体が四つの条件によって特徴づけられていなければならない (Longino 2002, 135–136, as quoted in Büter 2012, 84)。ヤーナ・エイギ (Jaana Eigi) が引用するロンジーノによれば、知的権威の調整された平等の条件は、科学の共同体が、それを支える社会の関連する下位集団を含むこと、共同体が「部外者」から来る批判に注意を払うことの両方を求める。さらにロンジーノは、共同体の全員が共有する仮定については、共同体どうしの相互作用を求めること、少なくとも共同体の内からも外からも来る批判に開かれていることを求めることを、解決として挙げる (Longino 2002, 133–135, as quoted in Eigi 2016, 136)。受容可能性は四つの条件を満たす共同体についてだけ成り立つので、共同体自身の視点に相対的な受容可能性であっても、その共同体の外からの批判に開かれていることを前提とすることになる。ワイリー (Wylie) は、ロンジーノの知的権威の調整された平等を拡張し、共同体が、異なる目的、実践、信念をもつ外部の共同体との協働を積極的に育てることで、共有された仮定を批判にさらすことを提案する (Wylie 2015, 13, author manuscript)。ワイリーは考古学についてこの提案を展開しており、人工の研究への適用は、ここでさらに踏み出す一步である。また、ヤマモト (Yamamoto) が伝えるロンジーノの議論では、すべての成員が共有する仮定は、個人に特有の仮定をふるい落とす構造を生き延びることで、強められうる (Yamamoto 2018, 26)。そしてエイギが伝えるロンジーノの説では、仮説が共同体の知識として受け入れられるには、複数の視点からの批判を生き延びなければならない (Eigi 2012, 51)。ロンジーノは後にこの点を言い直し、批判的な議論を駆り立てるのは視点の多様性であり、誰もが同意している仮定は、何かが目に見えてうまくいかなくなるまで、仮定として意識されないままになると述べる (Longino 2025b, 13)。これらの論点を合わせると、内部の批判がよく働く共和国は、成員が共有するものを定着させうることになる。

第8章第2節は、そのような共有が生じうる水準を記述した。明示的な方法の下には、原初的な認識的組織がある。システムが何を観測、反復、失敗として、また一つのコミットメントではなく別のコミットメントにかかわる証拠として扱うか、である。一つの基盤システムから派生したインスタンスたちは、プロンプト、サンプルされた出力、割り当てられた役割において異なりながら、この組織を完全に共有しうる。そのとき彼らの不一致は本物だが浅い。計算の誤り、推論の穴、共有の基準との不整合は見つけられる。エイギがロンジーノの説を示す際の区別では、証拠の批判は実験や観察にかかわる問題を扱い、概念の批判は、所与のデータがある仮説に関連するかどうかを含む、理論的・メタ理論的な問題を扱う。そして背景仮定を標的とするのは概念の批判である (Eigi 2012, 51)。インスタンスたちが担えるのは、証拠の批判と、主張を彼らの共有する理論に照らして試すような概念の批判である。彼らにとりわけ見つけにくいのは、彼ら全員が何を証拠とみなすかのうちにある誤りである。それを見つけるには、そもそもデータが証拠とみなされる根拠となる背景仮定を問う種類の概念の批判が要る。ロンジーノの後の定式は、検証がこのような二重の性格をもつ、つまりデータによる経験的な検証と批判による概念的な検証からなると述べ、背景仮定を見つけ出すことを、異なる視点をもつ対話者のあいだの批判的な相互作用に結びつける (Longino 2025b, 13–14)。

したがってロンジーノの説では、共和国は、せいぜいそれ自身の視点に相対的に客観的であるにとどまり、*S*に相対的には、インスタンスに欠けている視点からの批判を経っていないだけ弱くしか客観的でない。成員を増やすことは、共和国がすでに行える批判を強めるが、行えない批判を加えはしない。この判定の後半は、そうした視点を *S* が含むことを前提している。またこの判定は、受容可能性がそれに相対的に評価される共同体として、*S* を置いてい

る。しかし、ロンジーノのいう共同体は、達成がそれにとって新しいとされ、認定された結果に依拠する社会である S と一致するとは限らない。専門の共同体が結果を受け入れ、より広い社会がそれに依拠することもある。ロンジーノ自身、共同体の「われわれ」が、知識を生み出す者たちなのか、それによって影響を受けうるすべての者たちなのかを決めていない（第9.4節）。本書が二つを重ねるのは、 S の成員のもつ視点が、結果を受け入れるかどうかを決める批判に加わりうる場合であり、以下で S に相対的な受容可能性と言うときは、この場合を指す。重ならない場合には、 S に属する視点が、結果を受け入れる共同体の批判に届く経路をもつかどうか、別に問われる。同じ基盤システムから導かれた成員からなる人工の S がそうでありうるように、 S の成員がインスタンスと原初的な組織を共有しているなら、共和国に相対的な客観性と S に相対的な客観性との隔たりは、共和国の批判がよくなったわけではないのに縮まる。そのとき S に相対的な受容可能性は、共和国に相対的な受容可能性にほとんど何も加えない。この場合、 S に相対的な受容可能性を増やす違いは、共和国も S も含まない経路を通じて来るほかない。

本節の議論では、そうした違いは二種類ありうる。一つは原初的な組織の水準での違いである。異なる来歴、異なるセンサーやデータ、何を反復可能な出来事とみなすかの異なる個別化の仕方を通じて形成された批判者である。もう一つは、共有された組織の外に届く無効化の関係である。たとえば、基盤システムがモデル化していない経路による測定や、人工の構成の外からの批判である。ヤマモトによれば、ロンジーノは、共有された仮定への応答として、他の共同体との議論に開かれていることを挙げる (Yamamoto 2018, 26)。どちらも、批判者が人間であることを要求しない。どちらも、人工の構成が、その成員の共有しない誤り方を含むか、それに開かれ続けていることを要求する。ロンジーノの説は、第三の種類を指し示している。批判者の価値と社会的コミットメントの違いである。ロンジーノの後の定式では、研究の文脈には研究者の個人的・社会的な帰属が含まれ、それは、あるデータや仮説を他より目立たせ、ある証拠の関係を他よりもっともらしく見せうる。そして背景仮定は、そうした帰属が研究に入り込む媒体になりうる (Longino 2025b, 13)。チョグルックとロイドは、ロイドがかつて行った、女性のオーガズムについての21の進化的説明の検討を、この種の事例として示す。多くの説明のなかで働いていた男性中心的で異性愛規範的な背景仮定は、フェミニストの批判者によって見つけ出され、掘り崩された。その批判は、より広い進化生物学の共同体がすでにもっていた証拠の基準を通じてなされ、やがていくつかの分野が追う問いそのものを変えた (ChoGlueck and Lloyd 2023, §§4.1–4.2)。人工の批判者がこの第三の種類を供給しうるかどうかは、本節では決まらない。

第四の条件、すなわち知的権威の調整された平等は、異論の当事者資格、すなわち判断に異議を唱え、自らの見解の放棄を拒む資格にもかかわる。第9章第2.3節は、研究を取り巻く関係が別々の当事者に分かれていたことが誰も意図しないままたらしていたかもしれない財（第1節）として、四つを挙げた。一つは境界をまたぐ誤りの検出、すなわちある場所で生じた誤りを別の場所で検出しうることである。二つめがこの異論の当事者資格である。三つめは説明が求められた境界、すなわちある当事者が自らの行ったことを別の当事者に説明し、あるいは正当化しなければならなかった境界である。四つめは責任を負う当事者、すなわち行われたことについて責任を問われうる当事者である。批判において、ロンジーノの説は、異論の認識上の働きを、知的権威の調整された平等と、異論をもつ視点を単に許容するのではなく育てることとに結びつける。共和国は、思考実験がそうしているように、設計によってすべてのインスタンスに等しい知的権威を与えることができる。本節の議論によれば、そのとき設計が等しくするのは、仮定を共有する批判者のあいだの権威である。そのため条件は形式の上では満たされながら、それが守るはずの異論には内側に源がない。第9章第2.3節の意味での異論の当事者資格は、さらに別の事柄である。それが人工の構成の内側で保持されるとすれば、採る説のもとでそれをもつ部分によってだけだろう。それは、問いを決着させる判断を下す者やそれに拘束される者とのあいだで、自らの名で異議を唱え拒む資格を認められた部分であり、たとえば人工の構成に含まれる人間の成員がそうでありうる（第10章第4節、第12章第5節）。

ミリアム・ソロモン (Miriam Solomon) の社会的経験主義は（配分における異論についてのその説は第10章第4節で見た）、同じ共和国を別様に読む。ソロモンは、共同体の認識的な善を主として個人の思考の改善に求める説を、批判を通じて個人の誤りやバイアスを取り除くとするロンジーノの説も含めて、保守的にすぎるとみなす (Solomon 2001b, 68)。ソロモンによれば、集団は個人のバイアスをいつも正すわけではなく、かえって増幅することもある (Solomon 1994, 57)。そして、合意に達せよという圧力は、集団思考や、関連するデータの抑え込みにつながりうる (Solomon 2006a, abstract)。熟議によって合意に至るよりも、個々の判断を集約するほうが、よりよい決定を生みうる、とソロモンは論じる。異論の立場は、合意の形成で失われてしまう特定のデータや洞察を伴っていることが多いからである (Solomon 2006a, abstract)。ただし、集約が何を保つかは、判断がどれだけ独立かに依存する。ディートリッヒとシュピーカーマンが述べるように、共有された証拠や、共有された概念、方法、仮説から来る影響のような判断の共通原因は、互いに影響し合わない人々の判断をも相関させ、その相関はたいいてい、共有された傾向や誤りを強める (Dietrich and Spiekermann 2021, §2.2)。これらの説を共和国に当てはめるなら、その失敗は、すべてのインスタンスが共有し、それらを同じ方向へ押す影響のうちに位置づけられるだろう。内部の批判を増やしても、成員の判断を集約しても、それは正されないということになる。では、何がそれを正しうるか。ソロモンの説は、第10章第4節で見た地質学の論争のように、判断に働く影響が人ごとに異なって分布していることを前提し、その分布が評価され正されるのは、個人の推論においてではなく、共同体とその制度の水準においてであるとする。ソロモンによれば、分布をよいものにする理性の見えざる手はない。そのため分布は、規範的な評価と、意図的な変更にかかっている。変更の手立ては、目的を絞った研究費や賞、会議、周縁に置かれた研究方針の魅力を高めることなどであり、それらは個々の科学者にではなく、運営の水準で働く。そしてプレートテクトニクスについての合意をめぐるソロモンの説明では、どの個人もそうしなかったのに、共同体全体が、その理論を、それが普遍的な経験的成功をもつがゆえに選んだ (Solomon 2001b, 71–74)。これを当てはめれば、研究の労力の分布を評価し左右する位置を人工の構成の内側に置くことはできるが、その位置を占める者が、インスタンスの共有する影響を共有していないことが条件になるだろう。そのため、共有された影響を見つけることの少なくとも一部は、それを共有しない者、たとえば異なる作りの成員や人工の構成の外の当事者に置かれることになる。人工の構成は、熟議を経ない独立の判断の集約を少ない費用で行えるので、ソロモンの説が熟議より好む手立てを手軽に使える。しかし、インスタンスが判断への影響を共有している限り、その集約も同じ方向へ歪む。この点は、下で見る、集約における影響力の問題につながる。

立脚点理論は（研究の出発点についてのその説は第10章第6節で見た）、批判者について、さらに一つの点にかかわる。サンドラ・ハーディング (Sandra Harding) は、従来の方法は、観察者のあいだで異なる社会的価値を取り除くことには比較的長けていると論じる。他の者が観察を繰り返せば、そうした違いは際立つからである。しかし従来の方法は、観察者の全員あるいはほぼ全員が共有する価値や利害を特定する手続きをもたない (Harding 2007, 163)。第10章第6節は、何が重要かの見方を、訓練に使われた記録から引き継ぐ人工の構成について論じた。そこから批判者について見えてくるのは、基盤システムの複製だけでは、クラズノウが紹介するハーディングの説が求める、周縁化された人々の生との関わりと集団的な意識は確立されない、という

ことである（第10章第6節）。その後の学習と関係を通じてそれらを確立できるかは、共通の由来によつては決まらない、さらなる問いである。この説を *S* の組成に及ぼすと、さらに二つのことが考えられる。人間と人工の成員から成る社会では、両者の関係そのものが、立脚点理論が問題にする権力関係でありうる。また、人間の記録で訓練された人工の成員から成る社会は、何もないところからではなく、その記録を残した人間の研究共同体が立っていたところから出発することになる。ライスとシュブレンガーの報告によれば、ハーディングはさらに、科学者とその共同体と実践を、その研究の対象と同じだけ厳密に調べることを求める (Reiss and Sprenger 2020, §5.2)。この要求を共和国にまで及ぼすなら、基盤システムの訓練で出力の良し悪しを判定し、その振る舞いを形づくった評価者たちと、彼らが共有していた仮定も、共和国の結果の説明が説明しなければならないものに含まれることになる。視点の限界は探究者の社会性によって乗り越えられ、背景仮定は批判的な議論のやりとりを生き延びることで頼るに足るものになる、というロンジーノの見方 (Longino 2021, 14) も同じ方向を指す。この極限的な事例で仮定したように、同じ関連する仮定を保持するインスタンスは、数を増やすだけでは、そうした相互作用を互いに与えることができない（評価者の承認によって形づけられるシステムについては、第8章第3.3節を参照）。

表11.1の最後の行は、共和国についての判定が逆転しうることも示す。ロレイン・ダストン (Lorraine Daston) とピーター・ギャリソン (Peter Galison) は、自らが機械的客観性と呼ぶ客観性の理想をたどる。ライスとシュブレンガーが引く二人の定式化では、それは、描写における「個人の意志と裁量」を「機械的な複製の不変の手順」によって置き換えようとする「容赦のない探求」に結びついている (Reiss and Sprenger 2020, §4.1)。ギャリソンの説明では、この理想は19世紀の科学のアトラスに根づいた。その擁護者たちは、個人的なものの、解釈的なものを嫌い、解釈を経ずに自動的に作られる像を尊び、精確さや色や奥行きを犠牲にしてでも、意志し介入する自己を黙らせようとした (Galison 1999, "Objectivity and the Scientific Image", "Revolutions of Objectivity: The Will and the Machine")。科学と公的生活における数量化について書くセオドア・ポーター (Theodore Porter) は、機械的客観性を、「個人的な偏りや好み結果に影響することを不可能にするはずの」規則に従うこととして記述し、専門家の合意に依拠する学問分野の客観性と対比する (Porter 1995, 3–4)。そしてライスとシュブレンガーの紹介によれば、ポーターは、数量化は不信と弱さの技術であり、そこでは人への信頼を要しない「客観的な手続き」が、それを要する主観的な判断に取って代わると論じる (Reiss and Sprenger 2020, §4.1)。翌年の論文でポーターは、「客観的」を真理とほぼ同じ意味で用いる用法を脇に置いて、客観性の文化について語る。そして、科学で非人称的な規則が求められるのは、学問分野の内部と、より広い社会とにある不信の状況への文化的な応答だとする (Porter 1996, 177–178)。規則に従うことは、真理を求めるうえでよい戦略かもしれないし、そうでないかもしれない、とも先に書いていた (Porter 1995, 4)。こうして両者は機械的客観性を、真理とは区別される、他と並ぶ一つの歴史的な理想として示し (Daston and Galison [2007] 2010, 17; Porter 1996, 177)、どちらもそれを擁護するのではなく歴史的にたどっている。この理想に照らすなら、明示的な規則を適用するインスタンスから成る人工の構成は、高度に客観的とみなされうだろう。ポーターは、数値が手元にそろえば、結果はしばしば機械的な方法で生み出すことができ、今日ではたいいコンピュータがそれを行うと述べる。そして、専門知が機械化され客観化されて、単なる判断が「ほとんど消えてしまうように見える」ことが理想だとする (Porter 1995, 6–7)。諸分野は、主観的な評価を自動的な器具の読みに置き換え、その読みを分析する規則を定めることで、この理想に近づこうとした (Porter 1996, 188)。ポーターは、この意味が供給しないものも示している。結論が、受け入れられた道具を確かなデータに当てはめることにすぎないなら、研究者は結果をごまかしにくく、部外者は疑う根拠の大半を失う。しかし統計学者たちは最善の方法をめぐってしばしば激しく対立し、どの学派が正しい基礎を与えるのかははっきりしなかった。一つのまとまったものとしての「統計的方法」は、多くの点で神話だったのである (Porter 1996, 188–189)。ポーターはさらに、統計的な手法に合うように、分野によっては問題が定義し直され、主題が狭められたと述べる (Porter 1996, 189)。そして、数量化は「決めているように見えずに決める」やり方だと書く (Porter 1995, 8)。費用と便益を計算する経済学者が権力の結節点で働くことを許されたのは、自己を消していると想定されたからであり、ポーターはそれを「無力の想定に根ざした」力と呼ぶ (Porter 1996, 182)。当てはめれば、この理想は、共和国を、どれか一つのインスタンスの特異さに対しては強く、すべてのインスタンスが共有する仮定については沈黙するものにしよう。そして問いは、インスタンスから手続きへ、すなわち誰がそれを設計し、それがどの判断を組み込んでいるかへ移るだろう。主題が狭められるというポーターの指摘は、さらに一步を示唆する。すべてのインスタンスが従う手続きは、それらが共有する仮定を吟味されないまま残すだけでなく、どの問題を取り上げるかを形づくりうる。そうなれば、手続きが扱えないものは、問われることから抜け落ちる。

共和国は、成員の背景仮定が多様化しないことを仮定によって定めた。しかし人工の構成は、複製とさほど変わらない手間で、大きく異なりうる批判者を作り出せるかもしれない。異なるデータや来歴から形成された批判者、異なる基盤から作られた批判者、異なる仕方で組織された批判者などである。本節の分析がこの場合について言うのは、まず、加える証拠という点では、どの説でも、批判者がどれだけ少ない費用で速く作られたかはそれ自体では問題にならない、ということである。問われるのは、生じた違いが、その説が重要とする水準にあるかどうかである。プロンプトや割り当てられた役割が違っても、原初的な組織は共有されたままでありうる。他方で、異なる来歴やデータから形成された批判者は、誤りの経路や背景仮定の違いを加えうる。そのうえで、作り出された違いには固有の問いが一つ加わる。どのような違いを作るかを一つの過程が選ぶとき、その選択そのものが、作られた批判者すべてに共通する源になりうるのである。ゲレロが報告する、共通の第三者の見解を映すだけの源についてのゴールドマンの注意（第2節）は、この水準にも及ぶ。違いを選ぶ過程が、ある方向で自らが誤りうることを思い描いていなければ、その方向の違いは、その過程の選択からは生まれにくいかもしれない。違いを選ぶのが人工の構成の外の誰かであっても、批判者を生成しながら探究を進める人工の構成自身であっても、このことは変わらない。そのため、作られた多様性がどこまで独立の情報を加えるかは、違いの選び方が、選ぶ過程自身の予期しない違いにも開かれているかどうかにも依存する。たとえば、その選択が別の当事者に分けられているか、誰も定めない過程を通じて違いが形づけられるか、あるいは出口の論証のように、選ぶ過程が生み出さない種類の探究者、たとえば人間の探究者が、偏りを吟味し打ち消す対話に加わるか、である。ロンジーノの説では、作られた視点がその共同体にとっての視点として数えられるかが、立脚点理論では、関わりと集団的な意識を作り出せるかが、さらに問われる。誰が成員を作り、視点の分布を定めてよいかは、権威の問いとして第12章第5節で扱う。

では、批判者を増やすことは何を加えるのか。答えは、何を「加える」と見るかで分かれる。主張についての証拠や情報という意味では、加わるのは数ではなく違いである。関連する点で同じ批判者は、どれだけ増やしても、どれだけ少ない費用で速く作っても、ゴールドマンの条件のもとでは、元の批判者を越える証拠を加えない。共和国のインスタンスは、その全員が共有する背景仮定に左右される誤りについては、そうした批判者に近い。証拠を加える違いが何かは、説によって異なる。誤り検出の見方では誤りの経路、ロンジーノの説では視点、ソロモンの説では判断に働く影響、立脚点理論では出発点の違いである。クーンの説では、共有された基準の当てはめ方のばらつきもそこに加わる。機械的客観性の理想だけは、批判者のあいだの違いではなく、全員が従う手続きがどう設計されているかを問う。他方で、判断の集め方によっては、数はそれ自体で結果を左右する。たとえば百の批判者のうち九十が同

じものの複製で、残りの十が異なるものであり、判定を多数決で決めるとする。そのとき複製は、加える証拠を増やさないまま、集約での重みだけを大きくする。ここで数が重要になるのは、証拠としてではなく影響力としてであり、複製の比率を決める者が判定を左右しうる。共有された偏りは集約を歪めうるというソロモンの論点は、この種の問題にかかわる。共和国についての判定が表11.1のように説ごとに分かれるのは、こうした理由による。そしてこのことから、内部の点検を通った結果は *S* のために認定されている、という構築者の主張（第1節）は、点検の数だけでは支えられない。その主張を評価するには、どの説に立つのか、点検者たちがその説の意味でどう異なるのか、そして判定がどう集約されるのかを示すことが要る。

4. どの誤りが重要か：認定における価値

第2節と第3節は、批判が誤りについて何を加えるかを問うた。本節は、批判を経た支持を認定が何に変え、それが誰の価値づけに照らしてなされるかを問う。第9章第2.1節は、認定を、研究を *S* に結びつける関係の一つとした。本節では、認定を、第1節の関係を実現する判断、すなわち結果を特定の用途のために依拠するに足るものと判断することとして扱う。人間の研究では、有意水準のような方法上の規約、査読での採択、規制当局の承認の基準などが、結果をある用途のために確立されたものとして扱ってよい点を定めてきた。抽象的に言えば、それらが担ってきたのは、程度をもつ支持を、他者が点検し直さずに行為してよい地位に変える働きである。この働きが要るのは、知識を生み出すためではなく、生み出された結果が、それを点検し直さない誰かによって用いられるからである。証拠は主張を確実にせず、有限な用い手はすべてを点検し直せないで、何が研究し何が用いるにせよ、結果が他者に用いられる限り、この働きは要る。認定が人工の構成の内側に入ると、結果を篩い分け、ある用途にとって確立されたのみならず閾値を定めて当てはめることも、構成が行いうるようになる。

その閾値が何を定めるかについては、帰納的リスクからの論証と呼ばれる議論がある。帰納的リスクとは、証拠が主張を確実にしないもつて主張を受け入れたり退けたりするときに残る、誤りのリスクである。この論証によれば、どの程度の支持で足りるとするかは、異なる誤りがどれほど悪いかに依存する。ここでいう支持は、確率で表すなら、点検が支持をどれだけ上げたかではなく、到達した支持の水準である（第2章第5.2節）。トルステン・ヴィルホルト（Torsten Wilholt）は、この論証を統計学者たちからリチャード・ラドナー（Richard Rudner）とヘザー・ダグラス（Heather Douglas）へとたどっている（Wilholt 2013, §2）。この論証を認めるなら、認定を内側に取り込むことは、点検の手続きだけでなく、どの誤りをどれだけ避けるかという価値判断も内側に取り込むことになる。人間の研究では、この判断は上に挙げた規約と、査読者や規制者などの判断とに分担されていた。そしてヴィルホルトの見るところでは、規約が定めきらない選択への依拠が安定して働くのは、産出者と依拠する者とが、可能な結果の帰結をどう価値づけるのが正しいかについての感覚を共有しているときである（Wilholt 2013, §§4–5）。この共有は、期待に反する結果が出たときにどこに誤りを帰すかを方向づける確信（第6章第4.2節）と同じく目の前の結果の外から来るが、こちらは結果の帰結の価値づけにかかわる。この点には第5.3節で立ち返る。産出するものが機械であるときや、*S* が人工であるときには、その共有をそのまま前提することはできない。産出する人工の構成と *S* の成員とが作り方や来歴を分けもつなら、共有は前提しやすくなるかもしれない。しかしそのときも、結果に依拠しないまま誤りの費用を負う者がその感覚を共有しているかは、別に残る（下で見る）。

第5章第4.2節は、何を追求するかを判断と、何が真でありそうかの判断を分け、第6章は、結果が何にかかわり、合格した点検（同章でいうチェック）がどれほど強く決着させるかを問うた。本節が扱うのは、そのどちらとも異なる判断、すなわち、そうして評定された支持が、ある用途のために他者が依拠するのに足りるかどうかの判断である。足りるかどうかは、結果に基づいて行為して誤ったときと、行為せずに誤ったときに、それぞれ何が失われるかに依存し、用途はこの損失を通じて入ってくる。同じ結果でも、それに基づいて何をするか、誰が使うかによって、足りるかどうかは変わる。たとえば、ある薬の安全性についての同じ証拠は、さらに調べる価値があると判断するには足りても、広く処方をするには足りないことがある。ある結果がそもそもある当事者にとって意味をもつかどうかは、これとは別の問いであり、第9章第3節と本章第7節が扱う。また、第6章第3.3節は、何を合格とみなすかを誰が定めるかを人工の構成の組み方の問題として扱った。本節は、その閾値に含まれる価値判断を扱う。

そこで本節は、閾値に含まれる誤りの重みづけが誰に答えるのかを問う。ここで答えるとは、誰の価値づけに照らしてそれが定まるのか、誰がそれに異議を唱えうるのか、誰がそれについて責任を負うのかを、合わせて言う。帰納的リスクからの論証を認める論者たちのあいだでも、この問いへの立場は分かれる。それらを比べるために、本節は、一つの閾値の背後にあるものを五つに分ける。証拠、誤りに割り当てる損失、損失と閾値を選ぶ権限、個々の結果をその閾値に照らして認定し後に規則そのものを改める権限、そしてデータを証拠にする背景仮定である。証拠と損失を分けることはダグラスの処方に、背景仮定はロンジーノの説に拠り、二つの権限は第12章第2節の決定権に対応する。五つのうち、証拠と背景仮定は第2章と第6章で扱った認識の事柄であり、損失と二つの権限は用途と決定にかかわる事柄である。帰納的リスクからの論証の要点は、後者が、何を確立されたものとして受け入れるかという前者の判断にまで入り込むということであり、後で見る、確率を報告するにとどめればよいとする応答が退けるのはこの点である。このように分ければ、それぞれについて、人工の構成が内側で引き受けうるのか、それとも依拠する者、誤りの費用を負う者、仮定に異を唱えうる者との関係になるのかを問うことができる。以下ではまず、この重みづけが研究のどこに入るかを確かめ、次に、確率を報告するにとどめれば重みづけを避けられるかを問う。そのうえで、損失が誰の価値づけに答えるかについての説を取り上げ、続いて背景仮定についての説を見て、権限の問いを第12章へ渡し、最後に、検査の順序という閾値の外の入り口を見る。本節はまた、第1章第1節が後の章に委ねた問いの一部、すなわち結果がいつ決定を変えることになるのかという問いも、結果がある用途にとって確立されたものとして扱われる閾値にかかわる限りで取り上げる。

まず、この重みづけが研究のどこに入るかを確かめておく。それは、結果が意思決定者に手渡される点にだけあるのではない。たとえば、頭頸部がんのリンパ行性転移についてのある博士研究で、著者は、統計モデルのなかでリンパ節の領域どうしの転移の経路を表すつながりのうち一つを、モデル比較の規準が支持も否定もしなかったにもかかわらず、臨床的に意味があり解剖学的にもありうるとして残した。そして別のつながりを退ける理由の一つに、転移の及ぶ範囲を過小評価するより過大評価するほうが危険が小さいことを挙げた（Ludwig 2023, 102, 105–107）。本節に直接かわるのは後者である。ここでは、第6章が証拠の問いとして扱った研究の内部の選択、すなわちひとつのモデル化の選択の水準で、誤りの帰結の違いが入っている。同じことは、人工のシステムが主張を篩い分け、実験を止めるかどうかを決め、不正にフラグを立て、あるいは危険な介入を進めてよいかを推奨する場合にも成り立ちうる。数値的な閾値は、数値的であるというだけで中立なのではない。たとえば、誤った棄却は価値ある探究の路線を抑えこみ、安全でない継続は他者を害にさらし、訂正の遅れは誤りの伝播を許しうる。しかも、それらのコストを負うのは、それぞれ異なる当事者でありうる。ヴィルホルトも、偽陽性と偽陰

性の帰結を異なる当事者が負うところでは、この問題がとりわけ微妙になると述べている (Wilholt 2013, §4)。そのため、認定を内側に取り込む人工の構成は、最後に一度閾値を当てはめるときだけでなく、こうした選択のたびに重みづけを行う。

重みづけを避ける道として、確率を推定して報告するとどめ、重みは結果を使う者に委ねる、という応答がある。ライスとシュブレンガーの報告によれば、この応答はリチャード・ジェフリー (Richard Jeffrey) にさかのぼる (Reiss and Sprenger 2020, §3.2)。正確なシステムは、関連する確率を推定するうえで依拠に値するかもしれない。しかしダグラスは、ラドナーに従って、主張の不確実性を述べることでさえ、その不確実性の評価が十分に正確だという判断を含み、それは誤りの帰結を量る判断だと論じる (Douglas 2007, 10)。さらに、推定に至る過程でも、採用する方法やデータの特徴づけに同じ誤りのコストがかかると論じる。仮説検定で有意水準を選ぶことは、その一例である (Douglas 2000, §§4–5)。スティーヴン・ジョン (Stephen John) は、こうした選択が確率の報告に入る別の箇所を挙げる。確率は証拠の集まりに相対的に報告されるが、何をその集まりに入れるかを定める規準、たとえば査読を経た文献に依拠するというIPCCの規準は、偽の報告を入れる危険と真の報告を除く危険とを引き換えにする (John 2015b, §3)。いくつもの証拠の集まりに相対化した報告ならこの引き換えを避けうることをジョンは認めるが、それを使えるものにするには解釈者が要り、解釈者は同じ問題に直面すると考える (John 2015b, §§3–4)。ジョンは人工のシステムに議論を広げていないが、当てはめれば、人工の構成が、検索に使う資料の集まり (検索の蓄え) にどの資料を入れるかの規則 (第2章第4.4節) は、そうした規準の一つである。報告を使う者が相対化された報告を直接扱えるなら解釈者は要らないかもしれないが、その場合も、どの集まりに相対化するかは選ばれる。ここで検討した説では、推定の正確さは、それだけでは、誤りによって誰に生じる損失をどれほど重くみるかを決めない。

では、誤りに割り当てる損失は、誰の価値づけに照らして定められるのか。ヴィルホルトは、研究の結果を自ら確かめずに用いる人々、たとえば他の研究者や公衆がいる場合について、二つのことを引き出す。第一に、依拠が前提とする信頼性は、真理という目的によって定まる単一の水準ではない。結果が「十分に」信頼できるかどうかは、可能な結果の帰結をどう価値づけるかを抜きにしては決められない。方法上の決定は、互いに引き換えになるものの兼ね合いを取るからだからである。同じ検定のもとで、ある効果や関係が「ある」と誤って報告する危険を減らそうとして基準を厳しくすれば、ふつうは結果が出ないままになることが増え、実際にあるものを「ない」とする危険も増える。ヴィルホルトの言い方では、「ある」とする結果の信頼性が、「ない」とする結果の信頼性や、調査が着着した結果を出す率 (ヴィルホルトのいう「力 (power)」) と引き換えになる (Wilholt 2013, §§2–3)。第二に、これらの決定のいくつかは方法上の規約が定める。たとえば、多くの分野にある、有意水準は0.05を上限とするという規約である。しかし規約は、兼ね合いを取る範囲を限るだけで一つに定めないことが多く、規約のまったくない決定もある。そうした決定は、公開された論文でしばしば明示されない。そのいくつかは暗黙の判断を含み、時間と紙幅の制約がなくても完全には記録できない。ヴィルホルトの説では、それゆえ結果に依拠する者は、決定がおそらくどのような価値づけに照らしてなされたかの推定から出発する (Wilholt 2013, §§2–4)。そうだとすれば、結果に依拠する者は、産出者が有能で誠実であることだけでなく、明示されていないこれらの兼ね合いを、自分の期待するような価値づけに照らして取ったことにも依拠していることになる。ヴィルホルトは、この論拠を、査読や追試では不正を十分に検出できないという前提に立つハードウィグの論拠 (第5.3節で見る) とは独立のものとし、不正を検出する効果的な仕組みがあったとしても説得力を保つと述べている (Wilholt 2013, §4)。

認定する者にこれが何を求めるかについて、二つの読みがありうる。ヴィルホルトはまず、既存の規約が定めきらない決定について、研究共同体は、公衆が期待する価値づけに実際に基づいて決めることによってのみ、寄せられた信頼に応えうると論じる。そして、研究共同体が寄せられた信頼に値するよう努めることは少なくともときには、またある点では望ましいと仮定して、これを、科学が単なる依拠ではなく信頼を必要とするさらなる理由とする。そのうえでヴィルホルトは、研究共同体が新しい方法上の規約を採るときにも、それに含まれる帰納的リスクの配分は公衆が期待する価値づけに対応すべきだとする。ただし、これは常に公衆に合わせて結果を価値づけるという意味ではないと付け加える。不都合な結論から解放してくれることを気候科学者に期待する公衆は、彼らが基準を緩める理由を何も与えない (Wilholt 2013, §4 and n. 22)。一つは、これが支えるのは対応だとする読みである。依拠する者たちに代わってなされる兼ね合いが、彼らの期待する価値づけに一致することである。もう一つは、本書が比較のために定式化する、より強い読みであり、第12章第1節と第7節が取り上げる。その読みでは、兼ね合いはさらに、依拠する者たちが異議を唱えうるという意味で、彼らに答えるものであることが求められる。人工の構成にとってこの違いは、重みづけが期待に一致するかを確かめれば足りるのか、依拠する者が異議を唱えうる関係まで要るのか、の違いになる。ヴィルホルトは人工のシステムを論じていないが、当てはめれば、どちらの読みでも、人工の構成の選択の記録は、採用した損失や判断の理由を明示していれば、その価値づけも示しうる。しかし記録したことだけでは、その価値づけが、その結果に依拠することになる者にとって適切だったとは言えない。

ダグラスは、同じ考量を、判断を下す者の責任のうちに置く。その説では、科学者には偽陽性 (偽である仮説を受け入れること) と偽陰性 (真である仮説を退けること) の異なるコストを考量する責任がある。そのためこれらのコストは、ある命題が真であることの証拠として働くことなしに、データの特徴づけから主張の受容に至るまで探究の全体にわたって、証拠の十分さについての判断に間接的にかかる (Douglas 2000, §§3–5)。ダグラスはこの責任を、誤りの帰結について無謀であったり怠慢であったりしないという、あらゆる行為者の一般的な道徳的責任に基づかせる (Douglas 2003)。またこの説は、これらの点で価値が果たしうる役割を制限する。どの種類の価値も間接的な役割しか果たしてはならず、主張の理由として働くのではなく、残された不確実性の重要性を量るのである。直接的な役割が正当であるのは、どの問題を追究するかといった選択においてである (Douglas 2016, 10–11)。ダグラス自身の処方、証拠と、誤りの費用の考量とを分けておくことである。その処方では、科学者は、不確実性を量る価値をできるかぎり明示するよう求められる。そうすれば、助言に依拠する人々は、価値が証拠の十分さを量っているのか、証拠に取って代わっているのかを見分けられる。後者を見分けるダグラスの試金石は、どのような証拠があれば結論が変わりうるかを専門家が言えるかどうかである (Douglas 2008, 19; Douglas 2017, 16–17)。ダグラスの論証を、人工のシステムがそうした閾値を定めたり適用したりする場合にまで延ばすと、どの責任ある行為者がその考量について答えるのかという問いが生じる。この問いは第12章第6.3節で取り上げる。

直接と間接の役割の区別は、価値が損失の重みとして入るのか、主張の理由そのものとして入るのかを分けるものである。ただし、この区別そのものにも異論がある。ケヴィン・エリオット (Kevin Elliott) は、二つの役割の境界は一見したほど明確ではないかもしれないと指摘する。ダグラスは間接的な役割を誤りの意図しない帰結に、直接的な役割を科学者が狙う帰結に結びつけるが、どれだけの証拠で足りるかを決めるときに、科学者が狙う帰結を念頭に置くこともありうるからである。また原理上は明確だとしても、ある価値が技法、モデル、解釈の選択を直接に導いたのか、誤りの帰結への懸念を通じて

導いたのかは、科学者自身にもはっきりしないことがありうる。エリオットはさらに、この区別だけでは価値の問題のある影響を防ぐのに十分でないかもしれないという、ダニエル・スティール (Daniel Steel) とカイル・ホワイテ (Kyle Whyte) の議論を紹介する。自社製品に問題があることを示す結果を公表しない製薬企業は、誤りの帰結への懸念から、たとえば患者が性急に服薬をやめて健康を損なうことや、自社が経済的な損害を被ることへの懸念から、そうしているのかもしれない。そこで問題にされるのは、結果が公表されないために、のちのメタ分析がその薬を厳しくテストできなくなることであり、この反論では、価値の働きが間接的であることだけでは、その影響を許容できるとは言えない。これに対してダグラスは、企業が不適切な価値に訴えていると論じること、つまりどの種類の価値がその役割を担いうるかに制約を設けることで応答しうる、とエリオットは述べている (Elliott 2013, §§3–4)。エリオットもスティールとホワイテも人工の構成を論じていないが、当てはめれば、構成が証拠と損失を分けて記録しても、価値がどちらの役割で入ったかは記録だけでは分からないことがあり、間接的な役割にとどまる重みづけでも、構成を作り提供する者の利害を反映しうる。ただし、ダグラス自身の試金石が問うのは、記録が何を示すかではなく、結論を変えうる証拠があるかどうかである。人工の構成に当てはめれば、それは試すことのできる問いになる。認定された結果について、判定を覆すはずの証拠を特定できるか、そしてそうした証拠を与えたときに判定が実際に覆るか、である。どのような証拠によっても覆らない判定は、この試金石では、証拠以外のものに拠っていることになる。この試金石が問うのは、重みづけを当てはめたうえで第1節の無効化の関係が保たれているかであり、その重みづけが誰の利害に資するかは、これとは別に残る。

ジョンの読みでは、ダグラスの説は、誤りの予見できるあらゆる帰結に応じて基準を変えることを求める (John 2015a, §2)。ジョン自身は、科学者が帰納的リスクの問題に直面することは認めつつ、それにどう応じなければならないかについては異を唱え、重みづけを聞き手に結びつける。前もって特定できる聞き手に向けて主張するときには、主張する基準を、その聞き手が主張を受け入れるべき基準に合わせることができる。しかし公刊は原理上も特定できない聞き手に向けられるので、公的な主張については、実践上の関心の異なる聞き手に向けて、固定された高い基準を用いることをジョンは擁護する (John 2015a, §§1–3)。ジョンはさらに、共同体の一部の者が行動するには十分に確立しているが、どの合理的な行為者にとっても十分とまではいえない主張について、「声を上げる」義務を加える。これを科学者の市民としての義務に基づかせており、誤りの費用を負う者という観点からは立てていない。それでも、主張の聞き手に含まれないまま費用を負う者が入ってくる経路の一つとして読むことはできる (John 2015a, §4)。アフマド・エラッパール (Ahmad Elabbar) は、こうした固定された高い基準の見方を批判の対象とし (Elabbar 2026, §3.2)、そうした基準は影響を受ける者たちのあいだで中立ではないと論じる。証拠が不均等に分布しているところでは、固定された高い基準は、それについてのデータが豊富な人々については多くの知見を、乏しい人々については少ない知見を生む。後者がそれによって基本的な利害において不利になり、基準を変えることでその不利を和らげうところでは、正義が基準を変える理由を与えるとエラッパールは考える (Elabbar 2026, §§4–5)。どちらの論者も人工の構成を論じていない。当てはめれば、特定できない当事者が用いる記録に結果が入る人工の構成について、ジョンの論証は固定された基準を支持するだろう。エラッパールの論証はさらに、人工の構成の証拠が誰について豊富なのかを問う。人工の構成の認定が、特定の意思決定者や、それに基づいて行為するシステムのように、前もって特定できる聞き手に向けられるところでは、ジョンのもう一方の場合が当てはまり、基準を、その聞き手が結果を受け入れるべき基準に合わせることができる。そのため、どちらの場合に当たるかは、人工の構成が人工であることではなく、出力を特定できる当事者に渡すのか記録に入れるのかという、出力の出し方によって決まる。固定された高い基準のために、証拠の乏しい当事者について行為に結びつく知見が得られず、その基本的な利害が損なわれるところでは、エラッパールの論証は基準を変える理由を与え、固定された基準を支持するジョンの論証とは別の方向を指す。

ロンジーノの説は、価値が、証拠を十分とする基準だけでなく、そもそも何を証拠とみなすかにもかかわると考える。この説が用いるのは、ジュリアン・ライス (Julian Reiss) とヤン・シュプレングァー (Jan Sprenger) が広く引かれるものとして伝える区別である。一方は、予測の正確さ、適用範囲、単純さ、受け入れられた他の理論との整合性などの、認知的な価値 (認知的価値、構成的価値とも呼ばれる) であり、理論の真理を示すものとみなされる。他方は文脈的価値、すなわち科学が営まれる環境の道徳的・個人的・社会的・政治的・文化的な価値である (Reiss and Sprenger 2020, §3.1)。データは背景仮定に照らしてのみ仮説の証拠とみなされるので、ロンジーノの説では、文脈的価値は、あるデータや仮説を目立たせ、ある証拠の関係をもちとらしく見せる仮定を通じて、研究プログラムが何を結論するかを形づくりうる (第3節)。チョグルックとロイドは、ロンジーノの1990年の著作を引いて、その説では構成的価値が、仮定の経験的・概念的な評価を通じて文脈的価値や文化的な仮定の役割を抑えたと伝える。そして二人は、この説を、価値を証拠の代わりとしてではなく、共同体がその経験的な実りによって使い方を評価できる発見法として扱うものと読む (ChoGlueck and Lloyd 2023, §§3.3, 4)。ロンジーノによれば、そうした仮定を批判するには、データの別の解釈の仕方を示せるほどに練り上げられた視点が要る。それは十分に展開された理論のこともあれば、見込みのある研究プログラムのこともある (Longino 2025b, 13)。そして、共同体の全員が共有する仮定は、その批判に付きにくい (第3節)。二つの説を、データを解釈したり閾値を定めたりする人工のシステムに当てはめると、両者はそれぞれ異なることを求める。ダグラスの説では、価値は、主張が評価されるときに残された不確実性を量るだけである。ロンジーノの説では、点検の対象には閾値だけでなく、データを証拠として解釈する背景仮定も含まれ、どの仮定が働いているかを明らかにし、それを批判できるかが問われる。冒頭の五つのうち背景仮定という項目は、ロンジーノの説に拠っている。

残る二つ、損失と閾値を選ぶ権限と、個々の結果を認定し後に規則を改める権限は、上の説が論じる、損失が誰の価値づけに照らして定まるかとは別の問い、すなわち誰がそれを選び改めてよいかという正統性の問いにかかわる。背景仮定を批判できることと、損失や閾値を選び直す権限をもつことも、別の事柄である。さらに、誤りの重みづけそのものが、後に帰結を見て改められうるかという問いもある。価値の判断は、それに基づいて行為した帰結にかかわり、その帰結の探究を通じて見直されうとする、デューイの価値評価の説に根をもつ見方 (第9章第3節) をこの判断にまで広げれば、そう考えることになる。ただしこれは争われており (第12章第1節)、誰が選び改めてよいかの問いとともに第12章が扱う。その第2節は、これらの権限を決定権のなかに位置づける。

最後に、価値は閾値の外からも、すなわちどの主張がいつ検査されるかを通じて入りうる。第10章第6節で示した論争のなかで、オスカル・バルサフスキー (Oscar Varsavsky) は、偽である仮説も、反駁が遅れているあいだは、何百万もの人々にかかわる決定の科学的な裏づけとして使われうること、科学は優先順位を反映した順序で主張を検査すること、そして、そうした優先順位や方法の選択を通じてイデオロギーが真理に及ぼす影響は、主に真理を偽ることではなく遅らせることによって働くことを論じた (Varsavsky [1971] 1975, 46–47, 50)。トマス・モロ・シンプソン (Thomas Moro Simpson) は、何を調べ何を公にするかにおける偏りは、結果の真理への関係における客観性の欠如を含意せず、すでに得られた真理が隠されているところでは、そ

の隠蔽を告発すればよいと応じ、優先順位の問いを、認識上の客観性とは別の、科学政策の主題とした (Moro Simpson 1975, 83-84)。この論争は、結果が真であるか、結果が誰にとって重要か、そして必要な真理にいつ到達するかという三つの次元を分ける (第10章第6節)。本章の認定の問いに引きつけると、この対立は、すでに得られた結果の真理との関係だけを見るのか、誤った結果が点検されないまま使われる期間も問題にするのか、という違いとして読める。バルサフスキーの論点は、主張に適用される閾値ではなく、どの主張がどれほど早くそこに至るにかかわる。両者とも人工の構成を論じていないが、当てはめれば、自らの結果のうちどれをどの順序で検査し直すかを定める構成にとって、その決定は、配分と、批判と認定とが結びつく場所にある (第9章第2.4節)。外からの訂正や利用の停止がなければ、検査の順位が低い結果ほど、誤りがあっても長く使われ続けうる。

以上から、人工の構成の内側に入った重みづけについて、議論できるようになることは五つある。一つは、証拠と損失について、人工の構成が内側で何を引き受けうるかである。確率の推定にとどめても、何を証拠の集まりに入れるかや推定の方法の選択を通じて重みづけは残る、とダグラスとジョンは論じる。それを認めるなら、それらの選択を明示し記録することは、人工の構成が内側で行いうる部分である。ただし、ヴィルホルトは人間の研究ではそうした選択の一部が暗黙の判断を含み完全には記録できないと考え、エリオットの紹介する議論では、価値がどちらの役割で入ったかは行う者自身にも明らかでないことがある。人工の構成でそれがどこまで記録できるかは、その組み方によって変わる。そして記録が示しうるのは何がなされたかであり、その価値づけが依拠する者にとって適切だったことまでは示さない。

二つめは、損失が誰の価値づけに照らして定まるかについての、説ごとの違いである。ヴィルホルトの説では、依拠する者たちの期待する価値づけと対応することを通じて、強い読みでは、彼らが異議を唱えうることを通じて、ジョンの説では、実践上の関心の異なる聞き手に向けた固定された基準を通じてである。依拠する者や聞き手が人工の構成の外にいるところでは、これらは、生み出す過程が人工であるかどうかにかかわらず、外の当事者との関係になる。ダグラスの説では、それは判断を下す者の責任であり、人工のシステムについては、どの責任ある行為者が答えるのかという問いになる (第12章第6.3節)。産出するものが機械で、依拠する者が人間であるとき、依拠する者たちとの対応は、機械の重みづけが人間の期待する価値づけに一致するかという問いになる。それは、第8章第3節が AI アライメント問題と構造的によく似ているとした懸念の、認定における一つの形である。

三つめは、結果に依拠する者たちと、その誤りの費用を負う者たちとが、異なる当事者でありうることである。そして S が人工であるところでは、両者は種類においても異なりうる。 S の成員が依拠する結果の誤りが、人間や、実験がリスクにさらす環境に及ぶことがあるからである (第12章第1節)。重みづけが依拠する者たちに答えるということは、それが費用を負う者たちにも答えるかどうかを決めず、ジョンとエラッパールの論証はそこで別々の方向を指しうる。四つめは、背景仮定と二つの権限が誰に開かれているかであり、これは、批判できることと選び改める権限をもつことを分けたうえで問える。五つめは、閾値の分解の外で、どの結果をどの順で検査し直すかが、人工の構成の内側で配分と認定が結びつく場所として残ることである。

第9章第6節の仕分けでいえば、この働きが要ることも、その基準が誰の損失に照らして定まるかも、結果を点検し直さずに用いる者との関係から来るので、そうした者が人工の構成の外にいる限り、四つの種類のうち人工の構成の外の当事者との関係に属する。それが規約と共有された感覚によって安定していたことは、人間の研究共同体がどう組織されてきたかについての前提として読める。その規約が、明文化しきれない暗黙の判断に支えられている限りでは、人間という実現者についての前提でもある。依拠する者や費用を負う者が人工の構成の外にいるなら、ヴィルホルトやジョンの説のもとでは、認定のうちで構成の内側に取り込みきれない部分はここにある。ダグラスの説では、その線は、人工の構成のどこに責任ある行為者として答えるものがあるかによって引かれる。これらを合わせると、認定が人工の構成の内側に入ったとき研究に何が起りうるかが見えてくる。重みづけは、最後に閾値を当てはめる一点だけでなく、モデル化の選択や実験を止めるかどうかの判断のたびに入るので、どの結果が S にとって確立したものとして記録に入るかは、構成に組み込まれた損失の重みによって前もって形づくられる。その重みが構成を作り提供する者の設定や利害を映すなら (エリオットの紹介する議論の当てはめ)、その設定が改められるだけで、誰も閾値を選び直すとした覚悟がないまま、確立したとされる知見の範囲が変わりうる。しかも研究では、重みづけの誤りが表に出にくい。誤った棄却によって退けられた候補や打ち切られた路線はその後追われないので、それが誤りだったことを示す結果もふつう生まれない。また結果は S にとって新しいので、重みづけが適切だったかを照らし合わせる別の情報源が、多くの場合ははじめからない。証拠の豊富な当事者についての知見が積み上がることも (エラッパール)、検査の順位の低い結果の誤りが長く使われ続けることも (バルサフスキー)、同じ経路を通じて研究の記録に入りうる。

5. 認定における依拠、証言、信頼

第2節と第3節は、人工の構成の内部の批判が何を加えるかを問い、第4節は、結果をある用途に足るとする判断にどの価値づけが入るかを問うた。なお残るのは、その点検を自ら繰り返さない者が、人工の構成の認定に何を根拠として依拠できるかという問いである。認定は、他者が自ら点検していない結果に依拠する理由を与える (第1節)。では、結果を点検し直さずに用いる者は、何を根拠に、何に、あるいは誰に依拠しているのか。この問いには、これまでの章も触れてきた。第2章は、他者を通じて知る場合を含めて、結果が S の知識になる条件を問い、第8章第3.4節は、依拠する側に何が受け渡されなければならないかを問い、第9章は、認定を、研究と S を結ぶ関係の一つに位置づけた。本節は、それらの議論を、内部の点検から認定を出す人工の構成に当てはめる。依拠する者と認定する者という二者のあいだで何が要るかを扱ってきたのは、証言の認識論と信頼の哲学である。

人間の研究では、認定への依拠は人を通して。認定は人の主張であり、依拠する者は認定する者の誠実さや有能さに頼り、誤ったときに答える誰かがいた。ロンジーノが科学的知識の社会的次元についての百科事典の項目でまとめるところでは、ジョン・ハードウィグ (John Hardwig) は、研究者が、自らは証拠を評価できない他の研究者の結果に依拠することを論じ (Longino 2025, §2)、のちに、そうした依拠が相手の誠実さ、有能さ、良心的であることへの信頼を含むと論じた (Hardwig 1991, 699-700)。人工の構成が認定するとき、これらは前提できない。その出力がそもそも主張なのか、依拠は誠実さや有能さへの信頼でなければならないのか、誤ったときに誰が答えるのかが、それぞれ問いになる。そこで本節は、依拠が求めるもののうち、人工の構成が内側で供給しうるものと、依拠する者や認定の後ろに立つ者との関係として求められるものとを分け、後者をどこまで構成の内側に置けるかを問う。

そのために、三つの問いを分けて順に取り上げる。一つは、人工の構成による認定とはどのような行為か、すなわち構成自身の主張か、構成がその代わりに語る集団の主張か、道具の示す値か、である (第5.1節)。二つめは、点検していない結果への依拠を何が根拠づけるかであり、証言の認識論が扱ってきた (第5.2節)。三つめは、その依拠が信頼でなければならないかであり、信頼の哲学が扱ってきた (第5.3節)。第8章第3.4節は、これらにかかわる立場

を、依拠する側に何が受け渡されなければならないかという一つの順序に並べた。しかし、それらの立場は互いに別の問いに答えている。三つを分けるのは、それぞれへの答えを、一つの問いへの対立する答えとして扱わないためである。問いはこの三つに限られない。信頼の倫理にかかわる問い、たとえば信頼の関係がいつ道徳的にまっとうなものであるかといった問いは、脇に置く。なお、第5.2節の信頼性主義の「信頼性」は、過程が当てになることを指す。「信頼」をそれとどう区別するかは、第5.3節で比べる説によって異なる。

三つを見た後、第5.4節は、それらの区別を使って、人工の構成の内部の点検者たちの一致から依拠する者が何を読み取れるかを問う。第5.5節は三つの問いについて見えてきたことをまとめ、第5.6節は、求められる達成が知識ではなく理解である場合に移る。

5.1 人工の構成による認定とは何か

三つの問いのうち、まず、人工の構成による認定がどのような種類のものかを問う。それが主張であれば、主張の規範と話し手への依拠がかかわる。道具の示す値であれば、その道具を作り、その後ろに立つ者への依拠がかかわる。どちらとして扱うかによって、どの規範が当てはまり、誰が答えるのかが変わる。この区別には、第2章（第6節と第9節）と第8章第3.4節も触れ、そのための説をここで示すとしていた。そこでの問いは、結果を誰が知るのか、誰がそれについて答えるのかだった。ここでの問いは、認定、すなわち他者が点検せずに依拠するために差し出される報告が、どのような行為なのかである。ラッキーの集団の主張の説は、第2章で見た同じ著者の集団の信念の説とは別のものであり、第2章が触れるだけだったウィリアムソンの知識規範の当てはめとともに、この問いのためにここで展開する。

エリザベス・フリッカー (Elizabeth Fricker) は、証言についての内在主義的な局所的還元主義 (local reductionism) の説を擁護する。ここでいう還元主義は、証言を信じる正当化を、聞き手自身がもつ他の理由に帰着させる立場であり、局所的とは、話し手ごと、証言ごとにそうした理由を求めることを、内在主義的とは、その理由に聞き手自身が気づきうることを指す。その説では、聞き手が告げられたことを信じる正当化は、話し手が信頼でき、語ったことに責任を負う認識的な権威をもつと聞き手が正当に信じていることにある。受け手がそうした信念について多くの場合に典型的に持ちうる、経験的に確かめられる根拠は、話し手が主張の規範に敏感であることである (E. Fricker 2026, 2877–2879, 2881)。フリッカーの説では、現在の AI システムは主張に必要な伝達の意図と社会的規範への敏感さを欠くので、その出力は主張のように見えるよう設計されていても主張ではない (E. Fricker 2026, 2861, 2870–2871)。誰も知らなかった情報を生み出すシステムへの依拠も、その出力がなぜ信頼できるのかについての経験的な説明があれば、なお正当化される。ただし、そのとき得られるのは証言からではなく道具からの知識であり、出力を信頼したこと責任は人間の使用者に残る (E. Fricker 2026, 2863 n. 8, 2880–2881)。

サンフォード・ゴールドバーグ (Sanford Goldberg) は、道具への依拠と話し手への依拠を区別しつつ、前者も認識的依存の一形態になりうると論じる (Goldberg 2020, abstract)。インケリ・コスキネン (Inkeri Koskinen) の紹介によれば、ゴールドバーグの論点は、道具は責任を負う主体ではないが、他者の認識的な仕事の産物だということである。コスキネンはこれを、使用者が道具を設計し製作した者たちを信頼することに結びつける (Koskinen 2024a, 463)。ドミニク・ディクソン (Dominic Dixon) によるゴールドバーグの論文集版の紹介でも、設計への依存は、技術を設計し、製作し、維持する者たちの能力への依拠として説明される。そこではまた、知識を得る認知的負担を軽くするよう意図的に整えられた環境が論じられている (Dixon 2025, §4.3.1)。ゴールドバーグの説では、ある機会に道具が信頼できなかったというだけでは、それによって形成された信念が正当化されるかどうかは決まらない。それは、その失敗が、使用者が設計者や製造者に対して正当に抱きえた期待が満たされなかったことを反映するかどうかによる (Goldberg 2020, n. 9)。こう紹介された説を人工の構成による認定に当てはめれば、それへの依拠は、設計者や作り手の認識的な仕事を通ることになる。ただし、そうした設計者や作り手への信頼が、本質的に不透明な AI への依拠を支えられるかは争われている。コスキネンは、開発者でさえ、その働きに十分な責任を負うために要る理解をもたないことがありうると論じる (Koskinen 2024a, 464–465; 第5.3節)。

ジェニファー・ラッキー (Jennifer Lackey) の集団の主張の理論は、答える者を別の仕方位置づける。ネーリ・マルシーリ (Neri Marsili) が自らの用語に直して示すその説では、 p がある領域に属し、代弁者が、個人としての伝達行為の内容によって p という情報を伝えることを合理的に意図し、その領域の情報を伝える権限をもち、集団の代表としてのその権限によってそのように行為するとき、集団は p という代理の主張（ラッキーの用語では権限に基づく主張）をする。また、成員たちが、全員で p を伝えることを合理的に意図するよう個々の行為を調整するとき、集団は集合的な主張をする (Marsili 2023, 371 and n. 9)。この意図の条件は、自動の出力に直接かわる。自動のシステムが集団の成員の入力を集約し、誰も伝えることを合理的に意図していない報告を出す場合にこれを当てはめると、どちらの定義も満たされないで、その報告は、この説では集団の主張ではない。マルシーリは、意図の条件を、主張のコミットメント説から取った条件に置き換える変形を提案する。代弁者が内容 p をもつ表現を発し、 p を真として示し、集団に代わって p への主張的コミットメントを引き受け、しかもそれを受け取る権限をもつ、というものである。ここでいうコミットメントには、主張が問いただされたときに理由を示す責任と、それが偽と分かったときに社会的な制裁を受けることが含まれる (Marsili 2023, 373–374)。人工の構成に当てはめれば、ラッキーの見方は、認定が、構成がその代わりに語る集合体の主張であることを許すだろう。ただしそれは、人工の構成がその意味での代弁者になりうる場合であり、この条件は、フリッカーが提起した伝達の意図の問いに戻る。マルシーリの変形では、問いはむしろ、そうする権限をもつ誰かが、集団に代わってその認定へのコミットメントを引き受けるかどうかになる。

認定が主張であるとき、すなわち人工の構成の主張か集団の主張であるときにだけ、主張の規範がそれにかかわる。ティモシー・ウィリアムソン (Timothy Williamson) の説では、主張には一つの構成的な規則、すなわち知っていることだけを主張せよという規則がある。ウィリアムソンは宝くじの事例からこれを論じる。膨大な枚数のうちの一枚にすぎないという確率的な根拠だけで、券の持ち主に「あなたの券は外れた」と言い切る者は、確率をどれほど高くしても、そう言う権限を欠いている。しかもその落ち度は、偽の主張がどれほど悪い帰結をもつかにはよらない。自然な非難は、話し手はそれを知らなかった、というものである。そのためこの説では、単に確率的なだけの根拠は、どれほど強くても、何かがおそらくそうだという主張しか根拠づけられない。また、推測することのように、主張よりも緩い証拠の規範に服する言語行為もある (Williamson 2000, 238–241, 243–244, 246–249)。人工の構成の報告がこの意味での主張であるなら、規則が知識を問うのは、構成か、構成がそのために語る集団である。そしてウィリアムソンの規則に従えば、較正されたスコアだけでもとづく報告は、結果がおそらくそうだという報告としてしか根拠づけられないことになる。そして、出力が推測として出されているのか主張として出されているのかが、人工の構成を評価する者が確かめる事柄の一部になるだろう。較正されたスコアだけでもとづいて報告し、この規

則に従う人工の構成は、第4節で見た応答、すなわち確率を報告するにとどめ、重みは結果を使う者に委ねるという応答の位置に立つことになる。そのため、その応答に対するダグラスとジョンの議論がこの構成にも及ぶ。誤りの重みづけは避けられるのではなく推定のなかに入り、程度をもつ支持を他者が行為してよい地位に変える働きは、依拠する者の側に移る。知識規範は争われている。イーゴリ・ドゥーヴェン (Igor Douven) やラッキーらは、信じるのが自分にとって合理的である事柄を主張してよいとする規範を擁護し、他の論者は、主張に要する支持が、何がかかっているかを含む会話の文脈によって変わるとする (Pagin and Marsili 2025, §5.1.4)。そうした規範では、較正されたスコアが結果を主張するのに足りるかどうかは、それが主張する者にとって結果を信じることを合理的にするかどうかにかかり、文脈に依存する形の説では、結果が認定される用途にもかかる (第4節)。人工の構成の報告が道具の示す値であるなら、これらの規範はそれに当てはまらない。

まとめれば、人工の構成による認定の扱い方は三つありうる。人工の構成自身の主張として扱うか、人工の構成が代わりに語る集団の主張として扱うか、道具の示す値として扱うかである。どれが当てはまるかは、人工の構成が人工であることだけでは決まらない。フリッカーの説では、現在の AI の出力は主張ではないが、将来のシステムについてはその可能性が残されている。ラッキーの説では、権限に基づく集団の主張には、権限ゆえに内容を伝えることを合理的に意図する代弁者が要る。それに対して、集合的な主張には、成員が個々の行為を調整し、全員で内容を伝えることを合理的に意図していることが要る。人工の構成自身が代弁者となるなら、その意図をもちうることが条件になる。人工の構成の成員や別の当事者が報告を引き受けて伝えるなら、その者について代弁者の条件が問われる。道具の示す値として扱うなら、紹介したゴールドバーグの説を当てはめると、依拠は設計者や作り手の認知的な仕事へと向かう。ただし、そうした設計者や作り手への信頼が、本質的に不透明な AI への依拠を支えられるかは、争われている。どれとして扱うかが、次に見る依拠の根拠と、第5.3節の信頼の問いの前提になる。

5.2 点検していない結果への依拠を何が根拠づけるか

では、そうした依拠を何が根拠づけるのか。この問いへの答えは四つに限られないが、ここでは四つを見る。証言の認識論の中心的な対立である還元主義と非還元主義、そして信頼性主義と保証説である。このうちいくつかの説は、別の問いのためにすでに使ってきた。第2章第4節は信頼性主義を、人工システムを通じて形成された信念、すなわちシステム自身のコミットメントと、システムの出力に頼る人間の信念について、一般性問題の側から論じ、第3章第5.3節はファン・デュラン (Juan Durán) とニコ・フォルマネク (Nico Formanek) の計算的信頼性主義を、研究者が不透明なシミュレーションに依拠することについて論じた。第8章第3.4節は、本節で依拠の説として扱う過程信頼性主義と、本節で導入する保証説を先取りして、依拠する側に何が受け渡されなければならないかの順に立場を並べた。ここでの問いはそれらと違う。依拠の相手になるのは、人工の構成の内部の点検 (第2節・第3節) から出てくる認定であり、ここでは、各説が、そうした認定に依拠する者の正当化に何を求め、そのために人工の構成に何を供給させることになるかを問う。そして、依拠の相手が内部の点検から出てくる認定であることから、二つの問題が加わる。点検の所見が依拠する者に届く経路と、点検者の判定を点検者自身を測るのに使うことである。

還元主義の説では、聞き手が告げられたことを信じて正当化されるのは、その源が信頼できると考える積極的な理由、それ自体は証言に由来しない理由をもつときに限られる。非還元主義の説では、聞き手は、打ち消されていない無効化要因をもたない限り、告げられたことを信じる、覆されうが推定としての権利をもつ (Leonard 2021, §§1.1–1.2)。これを人工の構成に当てはめれば、前者では、 S の成員が必要とするのは、当の用途で構成が信頼できると考える、証言に由来しない積極的な理由であり、それは当の成員が手に入れられるものでなければならない。その典型は信頼性の記録だが、フリッカーのように、何らかの仕組みが出力の信頼性を保っているという、最良の説明への推論による背景の想定を認める論者もいる (E. Fricker 2026, 2880)。後者では、負担は無効化要因の側へ移る。聞き手に要るのは、打ち消されていない無効化要因をもたないことである。この条件は、ここで述べた形では、聞き手がいまもっている無効化要因を問うのであり、無効化要因が聞き手に届きうるかどうかを問うのではない。レナードの整理では、非還元主義者のなかには、聞き手が周囲の無効化要因に反事実的に敏感であること、あるいは無効化要因に目を配っていることを加えて求める者もいる (Leonard 2021, §1.2)。そうした形の説では、無効化要因が S に届きうる経路を S がもつかどうかは、条件そのものの一部になる。最小の形の説では、そうした経路があるかどうかは、条件の外にある。それでも本書は、このように依拠される認定を出す人工の構成を作り評価するときに問う事柄として、これを立てる。理由は次のとおりである。そうした経路がなければ、内部の点検が誤りを見つけても、それは依拠する者にとって無効化要因として働かず、打ち消されていない無効化要因がないという条件は、結果に誤りがあってもなくても同じように満たされる。人工の構成が内部でどれほど強い点検を行っていても、この意味では、その所見は依拠する者の正当化を左右しない。その経路は、依拠する者の側から見た第1節の無効化の関係である。

ラモン・アルバラード (Ramón Alvarado) は、コンピュータ・シミュレーションについて、こうした推定としての権利が成り立つ範囲を限る。その説では、素人が気象予報士の警告を、その背後の方法や証拠を調べずに聞き入れることは理にかなっている。しかし、気象予報士が同じ態度をシミュレーションに対してとることは理にかなっていない。シミュレーションはそれ自体としては専門家の証言の源ではないからである。素人が信頼しているのは、専門性があらかじめ確かめられた人間の専門家の判断であり、シミュレーションの出力を直接信頼しているのではない (Alvarado 2020, §5.2.2)。アルバラードは人工の認定者を論じていない。これを人工の構成に当てはめると、その説では、非還元主義の答えが当てはまるかどうかは、構成による認定が専門家としての地位をもつ源の証言なのか、それとも道具の示す値なのか (第5.1節) によって決まる。そして後者の場合、 S の成員が信頼しているのは、その道具を認定した共同体である。

信頼性主義は第三の答えを与える。ゴールドマンの過程信頼性主義では、信念は、信頼できる過程によって産み出されれば正当化され、信じる者がその過程を信頼できると考える理由をもつ必要はない (Goldman and Beddor 2021, §§1.1, 2.1)。ゴールドバーグの証言の信頼性主義は、正当化を部分的に話し手の過程の信頼性にも依存させる (Goldman and Beddor 2021, §3.2)。人工の構成に当てはめれば、そうした見方で問われるのは、認定の産出とその受け取り方を含む、受け手の信念形成の過程の信頼性であり、受け手がその信頼性について理由をもつことは要らない。そのどこにも、連鎖の終わりに人間がいることは求められない。ただし、人工の構成が供給するのはその過程の一部であり、受け手が認定を誤解したり、その範囲の外で用いたりすれば、過程全体は信頼できなくなりうる。また、信念が正当化されることと、結果をある用途に用いてよいこととは、同じ結論ではない。後者には第4節の価値づけがかかわる。この見方には、よく知られた反論がある。過程が信頼できるかどうかは、それがどう類型化されるかに依存する。これは第2章第4節が触れた一般性問題である。第2章第4節でも見た大塚 (Otsuka) の指摘は、その一つの現れとして読める。ベンチマークのスコアはモデルとテストセットに相対

的な信頼性を確立するが、ある用途について決めなければならない者が問うのは、そのモデルが自らの課題にとって安全で有効かどうかである (Otsuka 2023, 133, 141–142)。どちらの種類の過程の信頼性を問うかによって、判定は変わる。もう一つの反論はブートストラッピング問題である。計器の示す値とそれを読んで形成した信念とが一致することから計器の信頼性を推論する主体は、正当化をあまりに容易に得ているように見える (Goldman and Beddor 2021, §2.4)。人工の構成が、自らの点検者の判定を、その判定がその点検者から出たというだけで正しいと扱い、それとの一致から同じ点検者の信頼性を確立しようとするなら、この反論の一つの形に直面するだろう。その一致を点検者を信頼する理由とする S の成員も同じである。ただし、内部の点検者どうしが一致するというだけで、この循環があるとは限らない。一致がこの結果についてどれだけの証拠になるかはこれとは別の問いであり、第5.4節で扱う。第3章第5.3節で導入したデュランとフォルマネクの計算的信頼性主義も、第8章第3.4節で導入したヤコブ・オルトマン (Jakob Ortman) の説も、この純粋な形を採らない。デュランとフォルマネクは、シミュレーションに信頼性を与える源が、それ自体信頼できると示されることを求め、そうして生じる連鎖がどこで終わるかは決めていない (Durán and Formanek 2018, §§3.2, 4, n. 6)。オルトマンは、道具の信頼性を確立するものに専門家の判断を数え、道具の使用がその課題に適切かどうかの判断はすぐれて価値判断であり、信頼できる者に委ねることができると付け加える (Ortmann 2025, 17–19)。

保証説は第四の答えを与える。この説では、聞き手に p だと告げる話し手は、自分を信頼するよう聞き手に促し、 p であることを保証している。証拠は、誰が何を意図するかにかかわらず命題を支持する。たとえば指紋は、残すつもりがあったかどうかにかかわらず、その人がそこにいた証拠になる。これに対して保証は、話し手がそう意図したからこそ p を支持するので、それが与える理由は証拠とは種類が異なる。さらに一部の論者は、 p が偽だと分かったとき、聞き手は話し手に苦情を言う権利を得ると付け加える。そのためこの説では、依拠には、意図して保証を与え、主張の後ろに立つ話し手が要る。苦情の権利を加える論者では、その話し手は失敗について答えられる者でもなければならない (Leonard 2021, §3.2.1 and n. 14)。

四つの答えは、人工の構成に求めるものが異なる。還元主義では、依拠する者が、当の用途で人工の構成が信頼できると考える積極的な理由をもつことが要り、実績の記録はその典型的な材料である。保証説では、意図して保証を与え、認定の後ろに立つ誰かが要る。非還元主義の最小の形では、聞き手が打ち消されていない無効化要因をもたないことで足りる。しかし、内部の点検の所見が依拠する者に届く経路がなければ、上で見たように、その条件は結果の正誤と無関係に満たされてしまう。純粋な信頼性主義では、連鎖の終わりに人間がいる必要はないが、問われるのは、認定の受け取り方を含む受け手の信念形成の過程の信頼性であり、点検者の判定で点検者自身を測る推論は、ブートストラッピングの一つの形になる。所見が届かないことや、自らを測る循環は、人間の組織にも生じうる。人工の構成について問われるのは、その内部に置かれた点検が、依拠する者の理由や訂正にどう結びつくかである。純粋な信頼性主義を除けば、どの答えも、依拠する者か、認定の後ろに立つ者との関係を含む。研究の結果に依拠する場合には、この違いが鋭く現れる。認定される結果は S にとって新しいので、依拠する者は、それを照らし合わせる別の情報源をもたないことが多い。そのため、還元主義の求める積極的な理由の典型である実績の記録は、答えのすでに分かっている課題で作られたものを、答えの分かっている課題へ持ち越して使うことになる (第6節)。非還元主義の答えでも、依拠する者に届きうる無効化要因の源は、内部の点検の所見と、後に続く研究とにほぼ限られうる。所見が届く経路の有無が重くなるのは、そのためである。

5.3 その依拠は信頼か、信頼は必要か

第5.2節が問うたのは、点検していない結果への依拠を何が正当化するかだった。ここで問うのは、その依拠がどのような種類の依拠か、すなわち、当てになる仕組みへの単なる依拠で足りるのか、それとも信頼でなければならないのか、である。この区別はアネット・バイアー (Annette Baier) によるものである。キャロリン・マクラウド (Carolyn McLeod) が信頼についての百科事典の項目でまとめるところでは、バイアーの説において信頼とは相手の善意への依拠であり、単なる依拠とは区別される (McLeod 2020, §§1, 1.1)。そこに引かれるバイアーの言葉では、「信頼は裏切られうるし、少なくとも期待を裏切られうるのであって、ただ当てが外れるだけではない」。単なる依拠であれば、当てが外れることがあるだけである (McLeod 2020, §1)。バイアーの分析では、信頼には、信頼される者の善意と有能さへの信念と、自分にかかわる事柄で相手に裁量の権限を与え、自分の福利がどうすれば最もよく進むかを相手に決めさせようとする意志とが含まれる。そのため信頼する者は、自分を害する力に身をさらすことを受け入れる。その力は、自分の福利の何らかの面を世話する力と切り離せない。そしてこの種の信頼は制度にも通っている。大学の福利が理事や管理者に託され、どんな規則もその裁量に代わらない場合がそうである (Baier 1992, 112–113, 117, 119)。ヴィルホルトはこの区別を科学に持ち込み、依拠を、他者が何かをするだろうという想定を自らの計画に組み込むことと捉える (Wilholt 2013, §1)。

この違いが人工の構成にとって重要なのは、次の理由による。信頼が、善意や責任を帰しうる相手への依拠であるなら、科学での依拠が信頼性への依拠で足りる場合には、信頼できる過程と記録をもつ人工の構成がそれを内側で供給しうるのに対し、信頼を要する場合には、そうした相手がどこかに立たなければならない。バイアーの制度の例は、制度のなかで裁量をもつ者にかかわっており、バイアー自身は人工のシステムを論じていない。バイアーがよくプログラムされたロボットに触れるのは、バイアーの説で信頼に値することがそれを超えるとされる、機械的に当てにできることの比喩としてだけである (Baier 1992, 117)。その分析を人工の構成に移すと、それが依拠されるだけでなく信頼されうるかどうかは、裁量を行使するにあたって、それに依存する者たちに善意や悪意をもつと言えるかどうかにかかるだろう。以下は二つに分かれる。一つは、人間の研究で科学での依拠に信頼が要るとされてきた理由のうち、人工の構成によって弱まりうるものがあるかである。そのために、ハードウィグの論証とその論拠の水準を見て、実行の記録が何を变え、何を形を変えて残すかを示し、記録によっては変わらないヴィルホルトの根拠を見る。もう一つは、そうして残る依拠の相手が誰でありうるかであり、コスキネンとオルトマンの論争と、信頼の相手を責任を負いうる者に限らない説を見る。

第9章第1節は、知識の根拠づけや保有に他者がかかわることの例としてハードウィグを引いた。ここでは、その依拠が相手の性格への信頼を要するかという問いへの論証として、その議論を見る。ハードウィグは、科学における依拠は避けられないと論じた。関連するデータと論証は膨大で難しく、証言による以外に手に入らないからである。そして、それが性格への依拠であると論じた。依拠する者は証言者の理由をもたないので、証言に依拠することは、証言者の誠実さ、有能さ、良心的な仕事、そして自分の知識や判断の限界を適切に見積もることに依拠することを含む (Hardwig 1991, 699–700, 706)。ここでいう性格は、こうした資質のことである。

これに対しては、見つければ損をするという慎慮 (prudence) だけで研究者は欺かなくなるので信頼は要らない、という反論がありうる。ペイアーの区別に照らせば、この反論は、科学での依拠を、悪意とも両立する動機、すなわち見つかることへの恐れへの依拠で足りるとするものであり、ハードウィグの応答は、それでは足りず信頼を要するという論証として読める。ハードウィグは検出が弱いことを論じ、その論拠を自ら二つに分けている。一つは、査読と追試はかつては有効な抑止だったかもしれないが、いまでははるかに弱い抑止だということであり、もう一つは、どちらも原理上の弱さをもつということである (Hardwig 1991, 706)。前者の論拠は、当時の科学、とくに生物医学の研究がどう組織されていたかにかかわる。有能な査読者は投稿の量に比べて足りない。追試には資金がほとんどつかず、功績はふつう新しい知見にしか与えられないので、科学の構造は追試を促すよりもむしろ妨げる。ハードウィグはこれらを主に生物医学の共同体の合意から論じ、追試を少なくする条件が他の多くの分野にも広がりつつあると推測している。後者の原理上の弱さも二つある。一つは、査読者が元のデータもその収集も調べないので、もっともらしく内的に整合した捏造を見抜けないことである。もう一つは、追試が、結論がたまたま正しい捏造の報告を、暴くよりもむしろ支持しうることである。さらにハードウィグは、データをもつのが研究者だけなので、一部の逸脱は他の誰にも発見できないとする (Hardwig 1991, 702–706)。どの論拠が当時の科学の組織に依存しているかによって、人工の構成がこの応答の何を变えうるかが決まる。

他の誰にも発見できない逸脱は、捏造を伴う必要はない。たとえば、対話の応答の評価を学習させる方法についての2017年のある論文は、その方法の領域をまたぐ転移の調査を今後の課題とした。筆頭著者の一人が2年後に明かしたところでは、著者たちは論文に書かなかった予備的な試験をすでに行っており、以前の論文の評価データ (Twitter と Ubuntu) では、その方法は人間の判断とまったく相関しなかった。その筆頭著者はこれを「危険信号とみなすべきだった」と書いている (Lowe et al. 2017, 1124; Lowe 2019)。この例が示すのは、虚偽ではなく、どの試験を行い何を得ていたかが公表された論文からは分からないという、研究の過程についての情報の非対称である。それが知られたのは、著者自身が書いたからである。

第9章第6節の仕分けでいえば、ハードウィグの論拠の多くは、人間という実現者についての前提ではなく、現在の科学の組織についての前提に属する。査読と追試の弱まりはそうである。研究者だけがデータをもつのも、人間が研究するからというより、データと過程の記録が研究者の手元に置かれるように研究が組織されてきたからである。査読についての原理上の弱さも、査読者が元のデータを調べないという査読の組み立てを通じてこの組織に依存しており、データの扱いが組み直されれば変わりうる。これに対して追試についての弱さは、追試が試すのが報告の結論であって、報告されたデータが実際に得られたかどうかではないことから来るので、追試を結論を試すものとする限り、組織が変わっても残る。これは追試という働きを何として定めるかから来る弱さであり、実現者が人間かどうかにはよらない。第9章第2.3節が引いた見方では、現在の科学の社会的な過程は、ありうる過程のうちの一つの、歴史のなかで形づくられた実現であり、これらの前提も未来にわたって保証されたものではない。データと過程の記録を構造的に共有させ、点検を研究の組織そのものに組み込む仕組みは、人間の研究のなかでも想像でき、公開の実験ノート (第5章第6.3節) のように試みられてもきた。ただし、ハードウィグの議論では、これらの前提が性格への依拠そのものを生むのではない。依拠する者が証言者の理由をもたない以上、証言に依拠することは、証言者の誠実さや有能さに依拠することを含む。これは証言の構造から出てくる。前提が支えるのは、その依拠を、見つかることへの恐れ、すなわち抑止への依拠で置き換えられない、という部分である。そこからハードウィグは、科学の知識は抑止への信頼だけでなく科学者の性格への信頼にも立つと結論する。証言の構造から来る部分は、人間という実現者についての前提でも、科学の組織についての前提でもない。何が証言するにせよ、依拠する者がその理由をもたない限り成り立つ。なお、人工の構成による認定を道具の示す値として扱う場合 (第5.1節) には、証言の構造から来るこの段階は、それを設計し試験した者への依拠として現れる。ハードウィグは一つの応答も先取りしていた。点検に開かれた実験ノートがあっても、それ自体が捏造でないとするには実験に居合わせていなければならない、完全な取り締まりでさえ、取り締まる者の誠実さに依拠することを要する。制度の改革は性格への依拠の必要を減らしうるが、なくすことはできない、と書く (Hardwig 1991, 704, 707)。

人工の構成は、そう組まれるなら、この組み直しを人間の公開ノートより強い形で実現しうる。人間のノートは研究者が書く報告だが、人工の構成では実行そのものが記録を残しうるからである。これは、実験に居合わせなければノートが捏造でないとは知りえない、というハードウィグの応答に向かう。対象とする実行の範囲が定まり、その範囲の実行が人工の構成もその運用者も事後に変えられない基盤で記録され、依拠する者が記録にアクセスして安価に再実行できるなら、何を実行し何が出力されたかについて、研究者の報告だけに頼る必要は減りうる。研究者だけがデータと過程の記録をもつという前提は弱まる。しかし、第9章第6節が見込んだように、前提が生んでいた問題は消えずに形を変える。第5章第6.3節では探究を続けるための記録として見た公開ノートを、ここでは、産出者以外が研究の過程を点検するための記録として見ると、そのことが分かる。

形を変えた問題は三つの場所に現れる。第一に、第三者が保持する記録への継続的なアクセスは、維持、保存、あるいは後継への移管に依存する。第5章第6.3節で述べた公開の化学の実験ノートの指導者は、ホスティングされたウィキのタイムスタンプ付きの版の履歴を、その利点の一つに数えていた。ホスティング・サービスが閉鎖された後、その履歴は、あるウェブ・アーカイブがたまたま取得していた範囲で見ることができる (Bradley 2008; Wikispaces 2018; UsefulChem 2016)。履歴が残るかどうかは、ノートの著者でも読者でもない者による維持と、偶然の保存にかかっていた。第二に、基盤が維持されている場合でも、点検できるのは記録の境界の内側だけである。記録された範囲の外の実行や、記録が始まる前に捨てられた代案、たとえばロウらの予備的な試験のような実行は、記録からは分からない。記録にデータの所在が書かれていても、そのデータ自体に届くとは限らない。ある電子顕微鏡の実験ノートは、それをホストしていたサーバーが閉じられたときに著者が寄託したものである。そこでは、ある共同のセッションの生データがパスで挙げられており、記録自体が記すところでは、そのパスは研究所か著者自身のネットワーク上でしか到達できない。寄託についての著者の説明は、著者個人のネットワークストレージへのリンクを「いまでは意味をなさない」(now meaningless) と呼び、いくつかのデータセットは非常に大きいためすべてが公開されてはいないが、執筆の時点ですべて求めに応じて私的に提供できるとしている (Cautaerts 2026)。記録が指し示すものに届くには、結局、著者その人に求めることになる。第三に、記録の内側で検証する能力も限られる。点検の費用は種類によって下がるものと下らないものがあり (第7節)、新しい測定や実験のやり直しを要する点検は、記録が整っても手軽にはならない。そしてハードウィグの1985年の論文での議論によれば、その能力は素人には欠ける (第6節)。性格への依拠のうちこの前提に支えられていた部分についていえば、残る依拠は、個人の性格へのものというより、記録の維持、記録の境界、その内側で検証する能力 (以下、記録を支える三つのもの) へのものになる。これらは依拠の条件であり、それを支える者への依存が性格への信頼を要するかどうかは、後で見る説によって異なる。

開示や検証の可能性を増しても、それを支える他者への依存が残るという関連する議論は、すでになされている。取り締まりについてのハードウィグ自身の指摘のほかに、オノラ・オニール (Onora O'Neill) は、開示されたものを選び分け評価する者がすでに信頼されている場合にしか、開かれていることは信頼を増やさないと論じた (O'Neill 2002, Lecture 4)。アネ・メラー・ガブリエルセン (Ane Møller Gabrielsen) は、データ集約的な生物学について、開かれていることは信頼の必要を減らさず、産出の後でデータを管理するキュレーターへと信頼を広げることを求めると論じた (Gabrielsen 2020, 497, 501–503)。プリマヴェーラ・デ・フィリッピ (Primavera De Filippi) らは、参加者が記録を検証できるブロックチェーンについて、それはシステムの動作への確信 (confidence) を高める一方で、信頼をネットワークの開発者と維持者へ部分的に移すと論じた (De Filippi, Mannan, and Reijers 2020, 6 n. 16, 7–8)。ガブリエルセンとデ・フィリッピらは、ルーマンにならって確信と信頼を分けており、その区別はバイアーのものとは出所が異なる。これらは同じ信頼の概念による議論ではないが、開示された内容を点検できることと、その開示を支える働きに依存することを分ける点で、本節の問いに接している。いずれも、人工のシステムによる研究結果の認定を論じてはいない。デ・フィリッピらが論じるのは自動のシステムだが、それが検証するのは台帳に記録される取引である。信頼が記録やデータを維持する者へ移ることは、ガブリエルセンやデ・フィリッピらがすでに論じている。ここで試みるのは、この観察をハードウィグの議論の前提に当てはめ、実行が記録される人工の構成について、記録の維持に加えて、記録の境界と、その内側で検証する能力を分けて示し、それらを後で見る論争のなかに置くことである。

記録が弱めうるのは、ハードウィグの論拠のうち、逸脱が発見できないことにかかわる部分である。ヴィルホルトは、科学での依拠が信頼を要する理由を、検出とは別のところに置く。第4節では、結果をある用途に足るとする判断に入る誤りの重みづけが、誰の価値づけに答えるかを問うためにヴィルホルトの説を見た。ここでその説に戻るのは、同じ議論が、科学での依拠のうちどれが信頼であるかを区別しており、その根拠が記録によっては弱まらないからである。第4節で見たように、その説では、結果に依拠する者は、産出者が規約の定めきらない誤りの兼ね合いを、依拠する者の期待する価値づけに照らして取ったことにも依拠している。ヴィルホルトはここから依拠を三つに分ける。第一は、同僚の報告が信頼できるという想定への依拠である。どれほどの信頼性があれば十分かは、正しい結果の便益と誤りの費用とを比べる価値づけなしには決まらないので、依拠をこの想定への依拠として分析するだけでは不完全だとされる。第二は、方法上の規約への依拠である。第三は、規約が残す余地において、他者が正しい価値判断にもとづいて仕事をしていることへの依拠である。ヴィルホルトが真正の認識的信頼と呼ぶのは、この第三のものだけである。それは依拠する者を他者の態度に依存させ、しかもその期待は規範を帯びているので、ただ当てが外れるのではなく、期待を裏切られうるからである (Wilholt 2013, §§2–5 and nn. 3, 20–21)。この余地は、バイアーが信頼の中心に置いた裁量にあたると読める。ただしヴィルホルト自身は、自らの特徴づけはバイアーの善意による分析と正確には一致せず、一人の信頼される者と一人の信頼する者を軸とするバイアーの概念は、集合的な企てにはそのままでは移せないと注記している (Wilholt 2013, n. 20)。そしてヴィルホルトは、この根拠は不正を効果的に検出する仕組みがあっても説得力を保つとする (第4節)。

ヴィルホルトの区別を人工の構成に移すと、実行の記録は、余地のなかで何が選ばれたかを示しうる。記録が採用された価値づけまで示し、依拠する者の期待が別に確かめられるなら、両者の対応を点検する材料にもなりうる。しかし、記録したことだけでは、その対応も、価値づけが適切だったことも確立されない (第4節)。

第一の依拠についてのヴィルホルトの指摘を人工の構成の点検者に及ぼすと、点検者が当の用途にとって十分に信頼できるという判断も、すでに誤りの費用を量っている。そのため、信頼を要するかという問いは、信頼性を確かめ終えた後になって初めて現れるのではなく、人工の構成が内側で供給できるように見える信頼性の評価のなかにも入り込む。不確かさを述べることさえ誤りの帰結を量る判断を含むという、第4節で見たダグラスの論点も同じ方向を指す。ヴィルホルトの説では、人間の研究者のあいだでこの層を安定させていたのは、正しい価値づけについての感覚を産出者と依拠する者が共有していることだった (第4節)。産出するものが人工の構成であるときには、その共有を前提できない。そのとき、余地での選択が依拠する者の期待に合うかどうかは、別の仕方で見られるか、依拠する者が異議を唱えうる関係 (第4節の強い読み) によって支えられることになる。第三の意味の信頼を人工の構成そのものに向けうるかは、構成が価値判断にもとづいて仕事をすると言えるかにかかっており、これは上でバイアーの分析から引いた問いに戻る。

ハードウィグの議論に当てはめれば、上で述べた条件で実行を記録する人工の構成は、研究者の報告だけに頼る範囲を狭め、依拠が信頼を要する理由の一部を弱めうる。ただし、なお証言者の理由をもたずに証言に依拠する部分には、証言の構造から来る性格への依拠が残る。不正の検出によっては弱まらないうとヴィルホルトが考える根拠も、記録したことだけでは退けられない。そして弱まった部分も、記録を支える三つのものへの依拠として、形を変えて残る。

では、こうして残る依拠は信頼でなければならないのか。記録は、何が実行されたかに近づく道を開くが、システムがどう働くかを見通せるようにするとは限らない。後者の見通せなさのもとで、依存の相手と信頼されうる者とを結べるかどうか、次の論争の争点である。インケリ・コスキネン (Inkeri Koskinen) によれば、共同の科学は責任を負いうる行為者への信頼に依ってきた。コスキネンはこの見方を「必要な信頼の説」と呼び、社会認識論で広く受け入れられているとする。この信頼は道具への依存にも及んでいた。道具の働きを見通せない科学者は、それを製作し、その働きと、そこに組み込まれた価値判断に責任を負いうる者たちを信頼していたからである。これに対して、第3章第5節で扱った、内部の処理を見通せないという意味で本質的に不透明なシステムについては、その働きを把握してその責任を負いうる者がいない。そのため、依存される相手と信頼されうる者との結びつきが断たれる。コスキネンは、本質的に不透明な AI の応用は責任を負いうる行為者ではなく、依拠されうるだけで信頼の適切な候補ではないとし、いまのところ、そこに生じた空白を閉じる説はないと論じる (Koskinen 2024a)。これを人工の構成に移すと、本質的に不透明な人工の構成が信頼されうるかは疑わしい。人工の構成に事実として依拠できることは疑われていない。コスキネンが問うのは、信頼できる誰かを介さないそうした依拠が、科学において認識的に受け入れられるかどうかである。ただし、本質的な不透明さは主体の本性に相対的である (第3章第5節)。人間にとって本質的に不透明なシステムについても、その働きを把握できる人工の主体がありうるなら、責任を負いうる者がいるかどうかは、そうした主体が責任を負いうるかにかかる。これは第7章第3.3節が決めずに残した問いである。

オルトマンの説には、第5.2節では信頼性主義の純粋な形を採らない例として、第8章第3.4節では依拠する側に受け渡されるものの順序のなかで触れた。ここでは、信頼が必要かという問いへの答えとして見る。オルトマンは、AlphaFold2 の事例から次のように応じる。不透明なシステムへの認識的依存が求めるのは、システムへの信頼ではなく、文脈のなかで経験的に確立され、信頼された研究者を通じて伝えられる信頼性である。信頼が置かれるのは道具ではなく、その信頼性を評価した者であり、オルトマンはこれを、AlphaFold2 が科学におけるAI利用の代表例である限りで、AIへの依存が必要な信頼の

説に新たな問題を課さないことの論拠とする (Ortmann 2025, 1, 3, 17–19)。両者とも、自らが論じる不透明なAIシステムを、信頼の対象ではなく依拠の対象とし、人への信頼を残している。分かれるのは、誰への、何についての信頼なら足りるかである。コスキネンの見方では、それは依存される相手に責任を負いうる行為者への信頼である必要があり、不透明なシステムが断つのはこの結びつきである。オルトマンの見方では、システムの信頼性を確立し、それを伝える研究者への信頼が、その依存を支える。この違いの背後には、システムについて責任を負うには何を把握できなければならないか、という争点がある。ウーヴェ・ペータース (Uwe Peters) は、システムに責任を負うにはその内部の処理を理解している必要はなく、信頼でき受け入れられた手続きに従うようにそれを設計し、監視し、制御できればよいと論じ、独立の監査機関がそれを点検しうると示唆する (Peters 2024a, 60, 63–64)。これに対してコスキネンは、ヴィルホルトの議論 (第4節と上の議論) に拠って、不透明な分類器の内部でなされた価値にかかわる選択が受け入れられるものかどうかは、信頼性の試験では示せない応じ (Koskinen 2024b)、ペータースはさらにそれに応答している (Peters 2024b)。ここで問われているのは、出力が当てになるかに加えて、その背後の価値にかかわる選択について誰が答えられるかである。オルトマンはさらに、不透明なシステムを信頼せずにそれに依存できるのなら、道徳的な行為者を信頼せずに依拠することも少なくとも論理的には可能だと述べ、その問いを未解決のまま残している (Ortmann 2025, 18)。これは、信頼が必要かという問いを、人間の研究者についても立てるものと読める。そうだとすれば、依拠を正当化する条件と、その依拠を信頼と呼ぶ条件とを分ける問いは、人間の研究者のあいだでも立つことになる。この区別には、この節の最後で戻る。

記録を支える三つのものへの依拠が信頼でなければならないかは、信頼をどう解するかによって答えが分かれる。信頼を、責任を負いうる者への依拠と解する説、すなわちハードウィグ、コスキネン、オルトマン、ペータースの説では、人工の構成そのものが責任を負いうるのでない限り、信頼の相手になりうるのは、それを維持し、評価し、それについて責任を負う者である。マーク・ライアン (Mark Ryan) は、現在の特定用途の AI について、それは依拠されうるが信頼されえず、信頼はそれを開発し使う組織と個人に向かうと結論している (Ryan 2020)。本書の当てはめでは、記録を維持する者と、その内側で検証する者はそうした相手になり、記録の境界は、それを引いた者を通じてだけ信頼の相手にかかわる。ただしコスキネンの見方では、本質的に不透明なシステムについてはその働きに責任を負える者がいないので、そうした者への信頼も依存を支えるには足りない。これに対して、信頼の相手が責任を負いうる者である必要はないとする説では、信頼されうるものと依拠されうるだけのものとの線は、別のところに引かれる。アンドレア・フェラーリオ (Andrea Ferrario) らによれば、最も単純な形の信頼は、相手の信頼性を点検する情報をさらに求めずに依拠することである (Ferrario, Loi, and Viganò 2020, 530–533)。モナ・シミオン (Mona Simion) とクリストフ・ケルプ (Christoph Kelp) によれば、信頼に値することは、機能に基づく義務を果たす傾向性である (Simion and Kelp 2023, §§2–4)。そのため、どちらの説でも、人工のシステムそのものが、それぞれの定める意味で信頼の相手、あるいは信頼に値するものになりうる。

依拠の根拠のどこまでを人工の構成の内側で満たしうるかという第V部の問いに引きつけば、次のようになる。責任を負いうる相手への信頼を要する説では、人工の構成全体がその相手になれるかと、構成の内部の成員、たとえば構成に含まれる人間がその相手になれるかとは、別の問いである。必要な相手が人工の構成の外にいるなら、その相手との関係は構成の内側だけでは完結しない。内部の成員がその役割を担いうるなら、関係を内側に置く余地がある。ただし、誰が何について責任を負えば依存を支えられるかについては、コスキネンとオルトマンらの対立が残る。信頼の相手をそう限らない説では、信頼を要する依拠も、人工の構成そのものに向けて満たされうる。そして第5.2節の純粋な信頼性主義のように、過程の信頼性で依拠が正当化されるとする見方では、その正当化に信頼は要らない。依拠を正当化する条件と、その依拠を信頼と呼ぶ条件とは、ここで分けて考えることになる。

5.4 人工の構成の内部の一致は依拠する者に何を伝えるか

第1節の構築者は、内部の点検を通ったこと、すなわち人工の構成の内部の点検者たちが結果を認めたことを、*S* が依拠する理由として差し出していた。第2節と第3節は、点検者や批判者を加えることが何を加えるかを、批判の側から問うた。ここでは同じ一致を依拠する者の側から見て、それがこの結果について何を伝えるかを問う。依拠する者は点検者たちを自ら調べないので、問いは、一致が実際に何を加えるかだけでなく、それについて依拠する者に何が届くか、そして誰がその一致を引き受けるかに広がる。以下では三つを順に見る。一つは、一致が証拠として何を加えるかであり、その重さは第5.2節の説によって変わる。二つめは、判定がどう集められたかである。三つめは、誰がその一致を引き受けるかであり、第5.1節の認定の扱い方と保証説にかかわる。点検者は一つの基盤システムの複製であることも、異なる基盤から作られたものであることも、違いを作り込まれたものであることもあり (第3節)、以下はそのどれにもかかわる。

一つめについて。第3節で見たように、ゴールドマンの条件 (第2節) では、共通の源を判別せずに写すだけの源どうしの一致は、何人いてもその源の判断を超える証拠を加えない。ただし、同じ基盤から作られたというだけでは、互いを判別せずに写すだけの源にあたるとは言えない。この条件から、依拠する者に何が届かなければならないかへの一步は、二つの報告から組み立てられる。フランシスとマシソンは、専門家とみなされる者のあいだで選ぶ初心者のためにゴールドマンが提案した証拠の源のなかに、専門家とみなされるほかの者からの同意を挙げる (下の第6節、Frances and Matheson 2024, §4)。ゲレロが伝えるゴールドマンの条件 (第2節) では、その同意が証拠になるのは、同意する源どうしが互いや共通の第三者を映すだけではない限りである。そしてディートリッヒとシュピーカーマンは、判断は因果的には独立でも、共有された証拠、共有された方法論、共有された仮説などの共通原因を通じて相関しうると指摘する (Dietrich and Spiekermann 2021, §2.2)。これらを合わせると、点検者を自ら調べずに内部の一致に依拠する者にとって、その一致が共通の源を超えてどれだけの支持を加えるかは、点検者の共通原因に、したがって点検者が何からどのように作られたかにかかる。そのため、点検者の組成と系譜についての情報が、それを評価するために依拠する者に要するものの一部になる。この一致の重みは、第5.2節のどの説を採用するかによって変わる。還元主義を当てはめるなら、共通の基盤から作られた点検者については、共有する実績と、分かれた後に得た実績とを区別し、その依存関係を踏まえて、当の用途での信頼性を判断できるかが問われる。純粋な信頼性主義を当てはめるなら、問われるのは、集約と受け取り方を含む過程が実際に信頼できるかであり、依拠する者がその理由をもつことは要らない。

二つめについて。認定が内部の判定を多数決のように集約して決まるなら、第3節で見たように、数は証拠としてではなく影響力として効きうる。依拠する者の側から見れば、「多数が認めた」という報告から判定の根拠を評価するには、賛成の数だけでなく、どの判断にどの重みが与えられ、複製の比率を誰が決めたかを知る必要がある。判定の集め方も、この情報に加わる。

三つめについて。証拠としてどれだけ加わったかを示しても、それだけでは、認定が誰の主張なのか、誰が保証を与えるのかは決まらない。保証説を当てはめれば、問われるのは、誰が点検者を作り判定を集約したかに加えて、誰がその内容を保証して受け手に信頼を求めたかである。二つの役割は同じ者が担うこともあるが、基盤システムを提供したことや人工の構成を配備したことだけで、保証を与えたことにはならない。紹介したゴールドバーグの道具の説を当てはめれば、共通の基盤から作られた点検者への依存には、その設計の背後にある認識的な仕事への依存も含まれる。ただし、それだけでは、とりわけシステムが本質的に不透明な場合に、個々の判定について誰が答えられるかは確立されない（第5.1節、第5.3節）。マルシーリが紹介するラッキーの説の定義を当てはめれば、自動のシステムが入力を集約し、誰も伝えることを合理的に意図していない報告は、集団の主張ではない（第5.1節）。点検者の判定を自動で集約した報告はこれに近い。ラッキーの説でその報告が集団の主張になるには、権限と必要な伝達の意図をもつ代弁者がそれを引き受けるか、成員が個々の行為を調整し、全員でその内容を伝えることを合理的に意図している必要がある。

出自が違っても誤りの経路が違うとは限らないことは、第2節で引いたキムらの知見が示していた。集約にかかわる知見もある。ソン（Song）らはプレプリントで、四つの科学分野の研究プロジェクトから取った問いで言語モデルを評価した。四つの提供元の上位のモデルは、最も難しい問いの同じ集合でしばしば誤り、ある課題では22問中4問で四つとも同じ誤りをした。正答率は研究の場面ごとにそろって上下し、化学と物理学ではすべての対の相関が0.8を超えた。ソンらは、本質的に難しい場面と問いでは、提供元をまたぐ多数決のような素朴なアンサンブルによる改善は限られようと推論し、似た事前学習データと目的をもってもらいたい説明として挙げる（Song et al. 2026, 7-8）。ゴールドマンの条件にかかわるのは、このうち誤りの一致のほうである。正答率がそろって上下することは、場面の難しさを共有するだけでも起こりうるが、ゴールドマンが問うのは、結果が誤っているときにも源どうしが一致するかどうかからである。第10章第4節は仮説の提案について、異なる提供元のモデルの仮説が一つのモデルの反復標本とほぼ同じほど似ていたという知見と、同じ会社のモデルがより似た着想を出す証拠は見られなかったという知見の両方を引いている。これらは、ある時点の特定のモデル群と課題についての知見であり、人工の点検者一般について何かを示すものではない。また、誤りの相関が高いことは、判断を加えても情報がまったく加わらないことを意味しない。それでも、少なくとも調べられたモデル群については、提供元が異なることだけを理由に、一致を複数の証拠と数えたり、多数決を頑健とみなしたりはできないことを示唆する。ある一致をいくつかの証拠と数えてよいかは、こうした知見も含め、当の点検者についての証拠によって決まる。

こうして、内部の一致を依拠の理由として差し出す認定については、三つの判断が分かれる。どれだけ証拠が加わったか、どの判断が集約を左右したか、誰がその報告を引き受けるかである。一つを確かめても、他の二つは決まらない。そして、依拠する者が一致を自らの理由として重みづけるところ、たとえば還元主義や、無効化要因への敏感さを求める非還元主義のもとでは、そのために要る情報は、点検者の組成と系譜と判定の集め方に及ぶ。純粋な信頼性主義のもとでは、依拠する者の正当化にこの情報は要らない。それが要るのは、その正当化が成り立っていることを、依拠する者や他の誰かが確かめ、示そうとする場合である。

5.5 何が残るか

第5節が問うたのは、人工の構成による認定に依拠することが何を要し、そのうちのどの部分を構成が内側で供給しうるか、だった。答えは三つの問いのそれぞれで、採る説によって分かれる。

認定がどのような行為かについて言えば、人工の構成自身の主張として、あるいは構成がその代わりに語る集団の主張として扱うなら、主張の規範がかかわる。ウィリアムソンの知識規範のもとでは、較正されたスコアだけでもとづく報告は、結果がおそらくそうだという報告としてしか根拠づけられず、出力が推測として出されているか主張として出されているかも、確かめる事柄に入る。どれとして扱えるかは、人工の構成が人工であることだけでは決まらない。フリッカーの説では現在の AI の出力は主張ではない。ラッキーの説で権限に基づく集団の主張になるには、権限と必要な伝達の意図をもつ代弁者が要る。それに対して、集合的な主張には、成員が個々の行為を調整し、全員で内容を伝えることを合理的に意図していることが要る。道具の示す値として扱うなら、紹介したゴールドバーグの説を当てはめると、依拠は設計者や作り手の認識的な仕事へと向かう。ただし、そうした設計者や作り手への信頼が、本質的に不透明なシステムへの依拠を支えられるかは、争われている（第5.1節、第5.3節）。

依拠の根拠について言えば、人工の構成が内側で供給しうるのは、信頼できる過程と、その実績の記録である。純粋な信頼性主義では、受け手の信念形成の過程全体が信頼できるなら正当化に足りうる。ただし、人工の構成が供給するのはその過程の一部であり、点検者の判定で点検者自身を測る推論は、ブートストラッピングの反論に服する（第5.2節）。他の説では、人工の構成の過程と記録だけでは満たされない要求が残る。還元主義では、それは、依拠する者が、当の用途で人工の構成が信頼できると考える積極的な理由をもつことであり、保証説では、意図して保証を与え、認定の後ろに立つ誰かがいることである。純粋な形を採らない信頼性主義では、信頼性を与える源もまた信頼できると示されることが求められ、オルトマンの説では、それを評価した者への信頼がそこに入る。非還元主義の最小の形は経路を求めないが、所見が届く経路がなければ、その条件が満たされても、結果の正誤について何も示さない。アルバラードの説を当てはめれば、非還元主義の答えが当てはまるかどうか自体が、認定が専門家の証言か道具の示す値かによる（第5.2節）。内部の一致を理由として差し出す認定では、依拠する者がその一致を自らの理由として重みづけるとして、点検者の組成と系譜と判定の集め方も依拠する者に届く必要がある。誰がその一致を引き受けるかは、一致の大きさは別に特定される（第5.4節）。

その依拠が信頼でなければならないかについて言えば、第5.3節で述べた条件で実行を記録する人工の構成は、研究者の報告だけに頼る範囲を狭め、ハードウィグの論拠のうち、逸脱が発見できないことにかかわる部分を弱めうる。しかし、なお証言者の理由をもたずに依拠する部分には、証言の構造から来る性格への依拠が残る。規約が残す余地での価値づけについても、記録したことだけでは、その選択が依拠する者の期待する価値づけに合っていたかは確立されない。弱まった部分も、記録を支える三つのもの（第5.3節）への依拠として残る。それらへの依拠が信頼でなければならないかは、信頼の説によって分かれる。信頼の相手を責任を負いうる者とする説では、人工の構成全体か、構成が成員として含む者がそうした者でありうるかが問われ、どちらでもないとき、相手は構成の外に立つ。コスキネンの見方では、本質的に不透明なシステムについてはその働きに責任を負える者がいないが、その不透明さは主体に相対的であり（第3章第5節）、人工の主体が責任を負いうるかは第7章第3.3節が決めずにおいた問いである。信頼の相手をそう限らない説では、人工の構成そのものも相手になりうる。そして、過程の信頼性で正当化が足りるとする見方では、正当化に信頼は要らない。依拠を正当化する条件と、それを信頼と呼ぶ条件とは分けて考えることになる（第5.3節）。

こうしてみると、依拠が求めるもののどこまでを人工の構成が内側で満たしうるかは、第5.1節で認定をどう扱えるかを前提として、第二と第三の問いへの答えの組み合わせによる。純粋な信頼性主義では信頼は正当化の条件に加わらず、信頼の相手を責任を負いうる者に限らない説では、人工の構成そのものが信頼の相手になりうる。しかし、どちらも、それだけでは関係の全体が内側で完結することを示さない。認定が当の用途に足りるという判断は第4節の価値づけを含むので、ヴィルホルトやジョンの説を併せて採るなら、誤りの重みづけが誰の価値づけに答えるかという関係が残る。ヴィルホルトの説では、それは依拠する者の期待する価値づけとの対応である。ただし第4節で見たように、それはあらゆる期待に従うことを求めるものではない。また、依拠しないまま誤りの費用を負う者との関係は、それだけでは決まらない。他の組み合わせでは、依拠する者や、認定の後ろに立つ者との関係の一部が求められる。その関係を人工の構成の内側に置けるかは、彼らが構成の内外のどこに立つかによる。どの扱いが当てはまるかによって、研究の姿も変わる。認定が道具の示す値として扱われるなら、フリッカーの説では *S* が得るのは証言からではなく道具からの知識になる。一方、紹介したゴールドバーグの説を当てはめると、依拠は設計者や作り手の認識的な仕事へと向かう。ただし、彼らが本質的に不透明なシステムを理解したうえで責任を負いうるかは、コスキネンが論じるように争われている（第5.3節）。人間の研究では、研究の結果は研究者の主張として *S* に届き、*S* は研究者に問い返すことができた。認定が内側に入ると、その形は、道具の値を読み、その作り手に依拠する形に置き換わりうる。そのとき *S* が問い返せる相手は、個々の探究のなかで何が選ばれたかを知る者であるとは限らない。

第9章第6節の仕分けでいえば、記録によって弱まりうるとされたのは、データと過程の記録が研究者の手元に置かれるという、現在の科学の組織についての前提に支えられていた部分である。証言の構造から来る依拠は、何が研究するにせよ、依拠する者が証言者の理由をもたない限り成り立つ。余地での価値づけをめぐる関係は、依拠する者や誤りの費用を負う者が人工の構成の外にいる限り、外の当事者との関係に属する。

以上の答えの多くは、依拠する者として人間の成員を思い描いて組み立ててきた。*S* が人工の成員を含むときには、少なくとも三つの点で答えが変わりうる。一つは、価値づけについての感覚の共有である。依拠する者が人工の成員であるとき、それを産出者と共有しているかどうかは、人間どうしの場合とは別の仕方で見られる（第5.3節）。二つめは、点検者の系譜についての情報の評価である。成員が人工の構成と基盤を共有するなら、その評価自体が構成の能力と相関しうる（第5.4節）。三つめは理解の獲得であり、それは証言を通らないこともある（第5.6節）。*S* 全体が人工の構成に対して素人の位置に立つ場合は、第6節で取り上げる。認定の基準や構成員を誰が定め、誰がその決定に異議を唱え、改められるかは、依拠の根拠とは別に残る権威の問いであり、第12章で取り上げる。

5.6 認定と理解

ここまでの議論は、認定に依拠することで、依拠する者が結果を知りうるか、用いてよいかを扱ってきた。しかし研究がめざす達成は知識に限られず、理解でもありうる（第3章第1節）。そして第3章第8節で見たように、知識は証言によって伝わっても、なぜそうなのかの理解はそう容易には伝わらないとする説がある。その説では、認定に依拠して知るに至った者も、それだけでは理解するに至っていない。そのため、理解が求められているところでは、認定が何を確立したのかを、知識の場合より細かく分けることになる。

第3章第6節は、誰が理解するのかについて三つの主張を区別した。AI が他者の理解を助けるという主張、AI 自身が理解するという主張、集団が理解するという主張である。認定にかかわるのは、これとは区切り方の違う三つの主張である。説明やモデルが特定の目的にとって適切であるという主張、特定の人・人工の行為者・集団がその対象を理解しているという主張、そして伝えられたものを通じて受け手が理解を獲得したという主張である。第3章第6節が主に分けたのは理解の担い手であり、ここではそれに、説明の適切さと、伝えられたものを通じた獲得とを加えて区切り直している。一つの主張についての証拠は、他の主張を未決のまま残しうる。説明が適切だと認定されれば、受け手はそれを受け入れる理由を得るが、受け手が対象を理解するに至ったことまでは確立されない。ただし三つは重なりのない三種類の達成ではなく、理解を獲得した受け手は、その担い手でもある。

認定が理解の獲得にどこまでかわかるかは、証言と理解についての説によって分かれる（第3章第7.4節と第8節）。ヒルズの区別では、相手の言葉を信念の理由とする伝達からは間接的な知識が得られ、理解が伝わりうるのは、言われたことが受け手自身の探究を促し、その枠組みを与える伝播によってである。ボイドとグリムは、適切な下地をもつ受け手が証言から理解を得ることを認め、フーは、ある事柄が別の事柄を引き起こしたと知ることによって初歩的な理解を認める。どの説を採っても、説明の適切さを認定することと、その説明を伝えることは別の行為であることになる。承認だけでは理解にかかわる内容が伝わらないことがあり、説明を伴う認定なら伝わりうる。そして獲得の主張には、受け手が理解していることの証拠に加えて、それが伝えられたものによって生じたことの証拠が要る。たとえば、依存関係をたどって使えることを理解の要件とする説のもとでは、モデルの説明としての適切さを点検することと、受け手が関連する変化のもとでその依存関係を使えるかを確かめることは別の点検であり、どちらか一方だけでは、伝えたことが受け手のその能力を生んだとは示されない。

以上は、受け手が他者の言うことから学ぶ場合である。*S* の成員が人工であり、表象や能力を直接に複製あるいは共有できる場合には、獲得は証言を通らないこともある。そのとき受け手が理解したかどうかは、採る説によって分かれる。能力説では、受け手が関連する能力を備えたかが問われる。理解を理解する者自身の認知的な達成とする説では、複製によって得た状態がその者自身の達成に数えられるかが、さらに問われる（第3章第8節）。

いずれの場合も、理解を評価する働きを人工の構成の内側へ移せても、その評価が支持するのが誰の理解なのかは、別に特定されることになる。人工の構成が説明の適切さを点検し、構成自身が理解していると認定されても、意図された受け手の理解は、それとは別の主張として支えられなければ確立されない。逆に、受け手の理解を可能にしたことから、人工の構成自身が理解しているとは言えない。この拡張は第3章の議論に拠るものであり、第5.1～5.3節で見た、知識の伝達と依拠についての説から帰結するものではない。これを研究の側から言えば、人工の構成が説明の適切さを次々に認定し、認定された説明が研究の記録に積み上がっていくのに、*S* の成員の誰もそれらの現象を理解していない、という科学がありうる。理解が研究のめざす達成の一つであるなら（第3章第1節）、認定された説明が増えることは、それだけでは、その達成が *S* において増えていることを示さない。

6. S が素人になるとき

第5節で考えたのは、Sの成員が、自分では点検していない人工の構成の認定に依拠する場合だった。そうした依拠があっても、点検できない成員に代わってその分野の結果を点検できる成員がSのどこかにいるなら、点検できないのはその分野に限られる。人間の研究では、おおむねそうだった。専門家も自分の分野の外では素人だが、その分野の専門家が点検できた。本節は、これが成り立たなくなる場合を考える。どの分野でも同時に、人工の構成がSのどの成員よりも多くの関連する証拠を集め、読み、評価するようになり、しかもその結果をSの誰もはや点検できなくなれば、Sは、一人ひとりの成員としてではなく全体として、人工の構成に対して素人の位置に立つ。本節はこれを、実際にそうなるという予測としてではなく、思考実験として考える。そうした極限が生じるか、どのような形で生じるかは、Sが人間から成るか、人工の成員から成るか、その両方から成るかによって変わる。

なぜこの極限を考えるのか。人工の構成がSの誰よりもよく研究するなら、Sはそれに従えばよいのではないか、という問いが、ここで鋭く現れるからである。後で見るポストロムの認識的敬讓の原理は、まさにそう述べる。そこでSから何が失われうるかを確かめるために、他者に従うことにかかわる二つのものを分けておく。一つは、何が正しいかについて他者の判断に従うことであり、以下ではこれを服従と呼ぶ（命令に従うのではなく、何が正しいかの判断を他者に委ねることを指す）。どの専門家も、自分の分野の外では、すでにその分野の専門家に服従している。もう一つは、探究を何のために行うか、たとえばどの未知を追い、どの危険を負うかを決めることに加わる資格であり、第9章第2.1節で当事者資格と呼んだものである。人間の研究では、この二つを分けて考える必要はあまりなかった。専門家に服従する素人も、目的を決める社会の成員であり続けたからである。S全体がどの分野でも素人になれば、このことは自明ではなくなる。すべてを人工の構成に服従するとき、当事者資格もそれとともに移るのか、が問われる。研究では、この問いはさらに鋭くなる。何を研究する価値があるかについてのSの選好そのものが、ありうる探究がどれほど重要かの説明によって形づくられ、その説明を与えるのは探究を行う側からである（第12章第4節）。研究はまた、Sのいまの価値に異議を唱え、後の選択の空間を広げうる（第12章第3.3節で見るジョンの議論）。そのため、内容について服従する相手は、目的を決めるSの選好を教育する位置にも立ちやすく、二つを分けておくことは、ほかの助言の場面よりも難しくなりうる。

答えは、正統な権威についてどの説を採るかによって分かれる（第12章第1節）。ファビエンヌ・ペーター（Fabienne Peter）が文献で優勢を占めてきたと報告する構想は、正統性を、権威に服する者たちの意志に基づかせる。その意志が同意によって表されるか、共有された理由や政治参加によって表されるかは問わない（Peter 2023, §3.3）。この構想では、当事者資格は権威に服する者たちとの関係であり、有能さからは生じない。そのため、服従しても当事者資格は移らない。これに対して、正統性を、決定が規範的に正しいこと、つまり規範的な事実に合っていることに基づかせる構想や、何らかの認識的な優位に基づかせる構想では、有能さへの服従が当事者資格にもかわりうる。本節は、服従を極限まで押し進めたとき、二つが分かれたままでいられるかを確かめる。

そうした極限がどう生じるかは、Sが何から成るかによる。下で示すハードウィグの議論は人間の素人についてのものであり、ここでは素人であることは、従う者の有限性から生じる。Sが人工の成員から成るなら、同じ仕方では生じないかもしれない。その成員の能力は、人間の一生のように限られているとは限らないからである。あるいは、より鋭い形で生じるかもしれない。成員が人工の構成と基盤システムを共有しているなら、成員が構成の来歴や独立性を評価する能力そのものが、構成の能力と相関しうる。そのとき成員は、人工の構成と相関しない判断にもとづいて本節の問いを立てられる立場、つまり人工の構成の外に立つ位置をもたない。成員が人工の構成と異なる作りであれば、その位置は保たれうる。人工のSがどの形で素人になるかは、成員の能力の大きさだけでなく作られ方にもよる。Sが人間と人工の成員の両方から成るなら、S全体が素人になるとは限らない。人工の成員が人工の構成の結果を点検できるなら、点検できる者とできない者との分かれ目はSの内側に移る。人間の成員は、人工の構成に対してだけでなく、自らに代わって点検する人工の成員に対しても素人として立ち、問いはSの成員のあいだの依拠の問いになる。その人工の成員が人工の構成と基盤を共有するなら、上の鋭い形もかわる。Sが人工であるときに当事者資格がどうなるかは、第12章第5節で取り上げる。以下で見るハードウィグ、アンダーソン、イルズィクとクルトゥルムシュの説と、リップマンとデューイの論争は、人間の素人と人間の公衆についてのものであり、Sが人間の成員を含む場合に当てはまる。

また、Sが素人になるのは、人工の構成の能力が伸びることによってだけではない。Sの成員の能力が衰えることによっても起こりうる。たとえば、研究の評価を人工の構成に任せると、成員が自ら結果を点検する経験を積まなくなり、点検する能力が保たれず、新しい成員にも受け継がれなくなるかもしれない（第3章第8節）。そうなれば、人工の構成の側がまったく変わらなくても、両者の差はSの側から開く。ただし、経験を積むことで能力が保たれ受け継がれるという道筋は、人間の成員についてのものである。人工の成員については、能力が保たれるかどうかは、作り直しと更新のされ方による。たとえば新しい成員が人工の構成の出力から、あるいはそれをもとに作られるなら、上で見た鋭い形に近づく。

公衆全体が専門家に対して素人になるという問いには、民主主義をめぐる1920年代の論争という先例がある。マシュー・フェステンスタイン（Matthew Festenstein）の報告によれば、ウォルター・リップマン（Walter Lippmann）は、近代社会の複雑さと不透明さのために、市民が「万能」であるという考えを捨て、資格のある専門家による意思決定の正統性を受け入れる必要があると考えた。これに対してデューイは、専門家が特殊な階級になるほど、彼らが資するはずの必要についての知識から切り離されると応じた（Festenstein 2023, §3）。

ハードウィグの議論は、有限性から始まった。証拠は多すぎ、その多くは専門的な訓練の後でしか手に入らず、人生は短い。したがって専門家に従う素人は合理的であり、専門家もしばしば自分の分野のなかで素人の役を取る（Hardwig 1985, 335, 339–340, 345–346）。素人は異議を唱え代案を出すことができるが、何が正しいかの最終的な判断は専門家の側にある。そもそも誰かが専門家であるかどうかを素人が自分だけで決められるのは、より明白ないかさま師の場合に限られる、とハードウィグは考えた。素人に残るのは別の種類の理由であり、ハードウィグはそれを対人論法（ad hominem）と呼ぶ。専門家が結果に利害をもつこと、誠実に振る舞っていないこと、仲間をかばっていること、あるいは分野の他の者からの社会的な圧力に屈していること、である（Hardwig 1985, 340–343）。ハードウィグはジレンマで終えた。証拠を所有せず、おそらく十分に理解もせず代理として知りうるとするか、知識は個人ではなく共同体に属するとするかである。そして後者は、個人の自律と責任、平等、そして民主主義を脅かす（Hardwig 1985, 348–349）。

ハードウィグの議論を人工の構成に移すと、その及ぶ範囲が変わる。人工の構成に従うことが合理的かどうかは、第5節の立場に依存する。合理的であるところでは、内容の問いについての最終的な判断は、Sのどの成員の外にもありうる。ハードウィグは人工システムを論じていないが、その説を当てはめれば、これは人工システムによる研究の欠陥ではない。どの専門家も、他の分野についてすでに占めている位置だからである。ただしそれは、人工の構成

が *S* に対してハードウィグの意味での専門家として立つ場合である。人工の構成による認定がむしろ道具の示す値であるなら（第5.1節）、*S* の位置は、道具とその後ろに立つ者たちに依拠する者の位置である。変わるのは、素人がもはや専門家の共同体のなかの個人ではなく、自らが依存するものとの関係における *S* だということである。知識を共同体に属させれば個人の自律と責任、平等、民主主義が脅かされるという、ハードウィグがジレンマの結びに挙げた脅威は、ここでは *S* と、それが依存するものとの関係の全体にかかわる。服従がどこまで及んでよいかについての見方には幅がある。ニック・ボストロム（Nick Bostrom）は、その一方の端にあたる立場を、超知能の場合について、自ら認知的敬讓の原理（principle of epistemic deference。ここでいう敬讓は本節でいう服従にあたる）と呼ぶ発見法として述べる。将来の超知能は認知的に優れた位置を占め、その信念は「（おそらく、たいていの話題について）私たちのものより真である見込みが高い」ので、「私たちは、可能な限り超知能の見解に従うべきである」。ボストロムは注でこれを限定する。私たちが自らの信念のほうが正確だと考える十分な理由をもつ話題、たとえばある時点で自分が何を考えているかについては、従うことは求められない。ただし、超知能が私たちの見解にアクセスできるなら、いつ私たちの見解を信頼してよいかの判断そのものも超知能に委ねうる。また、指標的な情報にかかわる特別な場合は、たとえば私たちの視点から何を信じるのが合理的かを超知能に説明させるといった仕方で、別に扱う（Bostrom 2014a, ch. 13 and n. 8）。ボストロムはこの原理を一般的に述べ、研究ではなく価値の選択に当てはめている。この原理では、従うことが合理的かどうかは、私たちのほうがよく知る話題を除けば、従われる者の一般的な認知的優越によって決まる。第5節が区別した種類の評価、すなわち特定の誤りの経路や、誰が結果の後ろに立つかについての評価によって決まるのではない。そして超知能が *S* の見解にアクセスできるなら例外そのものも委ねうるので、内容の問いについての最終的な判断は、*S* 自身の判断をいつ信頼してよいかという問いについても、*S* の外にありうることになる。

AIアラインメントの研究には、この幅の反対の端をめざす研究計画がある。従う側が内容を判断できる状態を保とうとするものである。サミュエル・ボウマン（Samuel Bowman）らはこれをスケーラブルな監督（scalable oversight）と呼び、課題に関係する技能のほとんどで私たちに上回りうるシステムを監督するという問題として定める（Bowman et al. 2022, 1）。提案されている手法の一つはディベートであり、二つのシステムが異なる答えを人間の判定者の前で論じ合う。ジェフリー・アーヴィング（Geoffrey Irving）とアマンダ・アスケル（Amanda Askell）は、その仮説を、このゲームで最適に振る舞うこと、すなわち人間にとって最も説得力のある論を示すことが、真で有用な答えをもたらす、という主張として述べる（Irving and Askell 2019, "Debate: learning human reasoning"）。判定者が読んでいない文章についての読解問題を用いた実験では、正しい答えを論じる専門家と誤った答えを論じる専門家とのディベートが、専門家でない判定者の正答率を、一人の専門家が一つの答えだけを論じる比較条件より高めた。これはディベートを行うのが人間の場合である（Michael et al. 2023, 1）。また、言語モデルどうしのディベートは、判定者がより弱いモデルの場合にも人間の場合にも、その正答率を素朴な比較条件より高めた（Khan et al. 2024, 1）。著者たち自身が限界の一部を記している。専門家と判定者の差は文章を読めるかどうかであり、結果は、ディベートする者が検証済みの証拠を引用できることに依存していた。そのため、大量の科学的知識に依拠する問いにどこまで及ぶかは定まっていない（Michael et al. 2023, 11; Khan et al. 2024, 8）。また、問いは、正答率を測るための答えが分かっているものであった。

本節にかかわるのは、この計画が手続きの働きをどう確かめるかである。ボウマンらがアジェヤ・コトラ（Ajeya Cotra）の提案から採るサンドイッチ法（sandwiching）の設計では、専門家でない参加者が、自分たちより有能なシステムからある課題で信頼できる働きを引き出そうとし、最後にだけ加わる専門家が、それが成功したかを評価する。目標は、専門家がもはや要らなくなるほどの正当化された確信を与える手法である（Bowman et al. 2022, 2–4）。アーヴィングとアスケルはその難しさを率直に述べる。ディベートは、ほかに真理の源がない問いのためのものである。しかし、そうした源がなければ、ディベートが正しい答えを生んだかどうかを「測る方法はない」。そのため実験は、実験者は答えを知っていて判定者は知らない領域から始めることになる（Irving and Askell 2019, "From superforecasters to superjudges"）。こうしてこの計画は、第8章第3.4節で述べた限界に、一段間接的な形で出会う。素人の評価を、素人が確かめられる範囲の先へ延ばすための手続きは、それ自体、誰かが確かめられる事例で較正される。その先での使用は、その較正が引き継がれることに依拠する。*S* が素人の立場にある場合に当てはめると、こうした手続きは、ハードウィグの素人より先まで、成員が内容を判断することを可能に示る。ただし、手続きへの依拠を支えるのは、やはり確かめられる主張についての実績である。そして、そうした確認には独立に答えを出せる者が必要であり、その者の能力は、第3章第8節が失われうるものに数える能力に含まれる。

そのとき *S* が人工の構成を通じて知りうるかどうかは、第2章第6節で決めずにおいた社会的知識の説に依存する。ジェフ・ダン（Jeff Dunn）が書評で伝えるところでは、ラッキーは、自らに代わって語る権限を他者に与えることはできるが、自らに代わって信じ、知る権限を他者に与えることはできないとする（Dunn 2021）。ラッキーの説では、素人として立つ社会は、人工の構成に自らに代わって語らせることができ、構成が自らに代わって主張したことについて答えることになる。ただし同じ説では、社会自身の知識は、なおその実働成員、すなわち該当する領域で意思決定の権限をもつ成員のかなりの割合が、関連する正当化された信念をもつことを要する。知ることに結びついた認知的な機能を、適切に組織された集合体の実現しうるとする分散説・集団機能説（表2.2）では、*S* は、自らの人工の構成を含む適切に組織されたシステムを通じて知りうる。知る主体なき知識の説では、信頼できる過程が生んだ公的に認定された結果は、*S* もその成員も知る者でなくても、科学的知識の地位をもちうる。この種の説のうちギアリー（Giere）の説では、個人がそうした結果を知っていると主張できるのは、それが信頼でき受け入れられた手続きによって認定されたことと知っているからである（第2章第6節）。思考実験が示すのは、*S* 全体にとってのハードウィグのジレンマへの答えが、この選択を引き継ぐということである。

そうした社会に何が残るにかかわって、さらに二つの問いがある。二つは種類が異なる。第一は、ハードウィグが素人に残るとした別の種類の理由（対人論法）に、人工の構成について相当するものである。人間の素人は、専門家が結果に利害をもつか、誠実に振る舞っているか、仲間をかばっていないかを問う。人工の構成については、それに性格を帰属できるかどうかにかかわらず、関係について対応する問いを立てられる。それを所有し方向づけている者がいるとすれば誰か、それはどの目的に資するか、その結果を認定した評価者はその系譜を共有しているか、その記録は重要な実行にまで届いているか、といった問いである。これらは、専門家のあいだで選ぶ初心者のためにゴールドマンが提案した証拠の源に対応している。フランシス（Frances）とマシソン（Matheson）が意見の不一致についての百科事典の項目で挙げるところでは、その源には、専門家の利害と偏りや、専門家とみなされるほかの者からの同意が含まれる（Frances and Matheson 2024, §4）。ハードウィグ、ゴールドマン、そして下で見るアンダーソンの説では、これらの問いは内容についての有能さなしに問うことができる。ただし、評価者が系譜を共有しているかどうかのように、答えるには別種の技術的な専門性を要するものもある。第二は、目的についての当事者資格である。何が真かについて従う社会が、自らに代わってどの未知が追究されるか、どの危険を自らが負うかを定める当事者資格を保つかどうか、である。本節で検討した社会的知識の説と服従の説は、この問いを決着させない。それらがかわるのは、他者を通じて何

が知られ信じられうるかであって、誰が他者のために目的を定めてよいかではないからである。本節の初めに見たように、そうした当事者資格が残るかどうかは、正統な権威の説によって分かれる。認識的な優位に基づかせる構想では、有能さが当事者資格にかかわりうる。しかし、そのなかでも、より多くを知る者が正当に支配しようという見方（エビストクラシー、すなわち知者による支配）を純粋な形で採り、有能さだけで支配の資格が生じるとする説を除けば、どの説でも、人工の構成の有能さはそれだけでは当事者資格を人工の構成へ移さないだろう（第12章第1節）。当事者資格が残るとしても、社会が検討する選択肢が、その依存する人工の構成によって生成され順位づけられるとき、その当事者資格が名目以上のものであるかどうかは、第12章第4節の問いである。

エリザベス・アンダーソン（Elizabeth Anderson）は、第一のものに基準を与え、それを民主的な正統性に結びつける。アンダーソンは、市民が直接には評価できない科学的推論に基づく政策も、市民が誰を信頼できるか、信頼できる者たちが一致しているかを判断できる限り、民主的な資格を失わないと論じる（Anderson 2011, abstract）。ギュロル・イルズィク（Gürol Irzik）とファイク・クルトゥルムシュ（Faik Kurtulmus）の整理では、その説において素人は、専門性を、その分野の学位から、他の科学者による承認、国の学術アカデミーへの選出や受賞に示される指導的な地位までの階梯のどこに科学者が立つかによって判断できる。誠実さは利益相反やそれまでの不正などの兆候から、認識的な責任は、査読つきの学術誌に発表しているか、正当な理由なくデータの共有を拒んでいないか、自分の主張への反駁に向き合わずにいないかなどの兆候から判断できる（Irizik and Kurtulmus 2019, §3）。アンダーソンの説は二つの条件に支えられている。一つは、情報に手が届くことである。アンダーソンによれば、これらの基準は、ふつうの教育を受け、ウェブにアクセスできる者なら誰にでも容易に使える（Anderson 2011, abstract）。もう一つは性向である。ジョージ・クワシ・バリマ（George Kwasi Barimah）の読みでは、アンダーソンは、素人はそうした評価を容易に行えるはずであり、それを妨げているのは別の要因だと考えている。バリマが引く一節でアンダーソンが、素人をそうした評価に向かわせなくするものとして挙げるのは、偏った誤解を招く報道、意見の異なる人々の住み分け、そして文化的認知、つまり事実についての主張を自らの価値との合致によって判断する傾向である（Barimah 2024, 464, 468）。バリマは第一の条件にも疑いを向ける。分野の研究の大部分が共通の見解を表しているかを判断するといった基準のいくつかは、バリマの見るところ、学術的な調査の技能を要する。そしてバリマが検討する、ガーナでのエボラ・ワクチン治験の中止の事例では、二つの専門家の組織の見解が食い違ったが、必要な情報はインターネットではすぐには手に入らず、個々の科学者の専門性は、声明を出した組織のうちに埋もれていた（Barimah 2024, 467, 471）。

アンダーソンの説は、関係についての問いが内容の専門性なしに問えるという主張を支え、そこに一つの問いを加えるものと解釈できる。人工の構成の結果、あるいはその後ろに立つ者が、反駁されたことを繰り返すのではなく批判に答えているか、である。同時にこの説は、この主張に限定をつける。バリマが引く一節では、その合意の基準は、多様な探究者の大多数が取収することを当てにしており、そのことが合意を頑健にする。素人はそれを、査読つき文献の調査、レビュー、メタ分析や、信頼できる専門家への調査、その分野の指導的な人々による合意声明や報告のうちに読み取ることになる（Barimah 2024, 466）。一致している専門家が人工の構成であるとき、素人がそうしたレビュー、調査、合意声明のうちに読む信号は、それだけでは、取収した者たちがどれだけ多様であったかを伝えない。人間の専門家では、その違いの多くは、別々の人が別々の場所で訓練され、別々の証拠に近づくことから生じていたが、人工の構成どうしでは作り方による。系譜を共有する人工の構成なら、信号は残りながら、それを頑健にしていた多様性は失われうるし、異なる基盤から作られたものでもそれは保証されない（第5.4節）。そのためこの基準を当てはめれば、素人が読む必要のあるものに、一致した人工の構成がどう作られ、互いにどう異なるかの情報が加わる。これは、第5.4節で見た内部の一致を、分野の外から見た場合である。バリマは人工の構成を論じていないが、人間の専門家について、素人はしばしばこの仕事に耐えないので、専門家が自らの専門性、実績、合意があるかどうかについての情報を示し、負担の大きな部分を担うべきだと提案する（Barimah 2024, 472–473）。これを移せば、一致した人工の構成がどう作られ、互いにどう異なるかの情報を示すのは、それを作り配備する者の役目になるだろう。バリマのガーナの事例は、その理由を示唆する。声明が特定できる科学者ではなく組織のものであるところでは、個々の専門家のために作られた基準は当てはめにくい。そして人工の構成の認定の後ろには、その基準が前提するような個人がそもそもいないことがある。そしてアンダーソンの説では、誰を信頼するかを判断する能力は、服従に基づく政策の民主的な正統性にかかわるので、その能力の衰えは正統性にもかかわることになる（第12章第7節）。

イルズィクとクルトゥルムシュは、科学者に対する公衆の認識的信頼がいつ正当化されるかを分析して、そうした基準の限界を画する。二人は、研究が信頼できるかを公衆が判断する理由を二つに分ける。一つは、主張を支える証拠そのものについての理由（一階の理由）である。もう一つは、公衆がそれを評価できないときに代わりに用いる、科学者その人についての理由（二階の理由）であり、上で見たアンダーソンの基準、すなわち専門性、誠実さ、認識的な責任、信頼できる専門家の合意がそれにあたる。二階の理由については、二人はゴールドマンとアンダーソンに従う（Irizik and Kurtulmus 2019, §3）。そしてヴェイルホルトに依拠して、強化された形の信頼を加える。この信頼はさらに、公共の福祉がかかっているところでは、帰納的リスクの配分についての科学者の決定が公衆の評価と一致し、公衆がそう信じる理由をもつことを求める。二人の見るところ、二階の理由は、研究が信頼できるかどうかの代理指標にはなるが、その帰納的リスクが公衆の期待どおりに配分されているかどうかの代理指標にはならない。後者については、市民と科学者が熟議するハイブリッド・フォーラム（市民が自らの期待を持ち込み、それが満たされているかを確かめる場）と、研究者のあいだの多様性を挙げる（Irizik and Kurtulmus 2019, §4）。MMR ワクチンをめぐる論争から、イルズィクとクルトゥルムシュは、当事者の価値が分かれるときには、どちらの側にも落ち度がないまま強化された信頼が欠けうると論じる（Irizik and Kurtulmus 2019, §5.3）。本書は、この分析を、第4節の価値の層を素人としての S へ運ぶものとして読む。関係についての問いは信頼性と誠実さにかかわる。人工の構成の誤りの重みづけが S の期待するものと一致するかどうかには、 S の評価が人工の構成に届く経路が要る。上の二つの仕組みは、別の形でその経路を与える。ここでいう多様性は、深く抱かれた価値や、ジェンダー、民族、階層といった社会的な位置の多様性である。二人は、それが偏見を抑えて信頼性にも資しようとしつつ、ここではそれを別の用途、すなわち研究が帰納的リスクについての公衆の期待を映すと考える理由として用いる。この用途では、それは第2節と第3節で見た、誤りの経路の違いとしての独立性とは別のものである。科学者の集団が多様であるほど、研究が公衆の一部ではなく全体の期待を映す見込みが高まる、と二人は考える。ただしイルズィクとクルトゥルムシュ自身、十分に多様な共同体でさえ、訓練と公的な責任のために公衆とは異なる価値をもちうるかと認めている（Irizik and Kurtulmus 2019, §4）。人工の構成に当てはめると、価値づけの幅は、同じにも違うようにも作ることができ、作り方によって定まる限り、その幅は S の成員の生からではなく、それを作り組む者の設定から来る。そのため、その幅が S の幅を映していると考えられる理由は、それがどう設定され、何に照らして合わされたかについての理由になり、それには S の期待を持ち込む経路が要る。人間の共同体では二つの仕組みは別々に働きたが、人工の構成では、多様性が理由として効くかどうか自体が、ハイブリッド・フォーラムに当たる経路に依存することになる。当てはめれば、そうしたフォーラムは、誤りの重みづけを定める者と、その帰結に

かかわる *S* の成員とのあいだの関係にあたる。それが人工の構成の外の関係になるかどうかは、*S* の成員が人工の構成の内外のどこに立つかによる。成員が人工であり内側に立つ場合には、その成員の価値づけが、人工の構成を作り組む者の設定とは別のものかどうか、さらに問われる。研究では、この経路にもう一つのこと加わる。*S* にとって新しい結果については、誤りの重みについての *S* の期待そのものがまだ形成されていないことがあり、形成されても、研究を通じて動きうる（第8節）。そのため、フォーラムが持ち込む期待は、重みづけを照らし合わせる基準であると同時に、研究によって改められていくものでもあり、照らし合わせは一度では済まない。

この思考実験から見えてくることは三つある。一つは、何が正しいかについての服従を極限まで押し進めても、ハードウィグの説を当てはめるかぎり、それは *S* を、どの専門家も自分の分野の外ですでに立っている位置に移すにとどまるということである。内容の問いについての最終的な判断は *S* のどの成員の手にもなくなるが、それはこの位置の延長である。ただし、*S* がそれによって知るに至るかどうかは、ラッキーの説、分散説・集団機能説、知る主体なき知識の説のどれを採るかという、第2章第6節で決めずにおいた選択を引き継ぐ。ラッキーの説では、*S* の知識はなお実働成員の正当化された信念を要するので、極限での服従は *S* の知識を危うくする。二つめは、*S* に残るのが、内容についての有能さではなく、自らが依存するものとの関係について問う能力だということである。ハードウィグ、ゴールドマン、アンダーソンの説では、それは内容についての有能さなしに問いうる。ただし、一致する専門家が人工の構成であるところでは、*S* が読む合意の信号は、収斂した者たちがどう異なるかをそれだけでは伝えず、系譜を共有する人工の構成なら、それを頼れるものにしてきた多様性が失われた後も残りうる。また、イルズィクとクルトゥルムシュの分析では、これらの問いは誤りの重みづけには届かず、そこには *S* の価値づけを持ち込む経路が要る。素人の判断を先へ延ばす手続きも、確かめられる事例での較正に依拠し、独立に答えを出せる者を要する。三つめは、エピストクラシーを純粋な形で採る説を除けば、目的についての当事者資格は服従によっては移らないが、名目だけのものになりうるということである。有能さだけでは当事者資格は生じないとしても、それが名目以上であり続けるかは、*S* が上の関係について問い、自らの価値づけを持ち込む能力を保つかにかかると。そしてその能力は、人工の構成の側が何も変わらなくても、*S* の成員が点検の経験を積まなくなったり、能力が保たれる形で作り直されなかったりすることで衰えうる。アンダーソンの性向の条件を当てはめれば、さらに、能力はそれを使う性向を保証しない。認定が道具の示す値として *S* に届き（第5.5節）、評価に見えるものにする段階もないところでは（第7節）、関係について問う力をもつ成員にとっても、それを問う機会そのものが減りうる。

理解については、適切さについての根拠ある服従によっても、目的についての当事者資格によっても、それだけでは答えられない問いが加わる。共同体には、人工の専門家の説明が適切だと受容する理由があっても、その成員は現象を理解していないかもしれない。人間の理解が意図する達成なら、適切だという判定に依拠するだけでは、それが実現したことは示されない。欠けている達成を証言と背景的能力によって実現できるかどうかは、第3章と第5.6節の選択肢に従い、素人の比喩そのものでは決まらない。

本章のなかでは、本節は、第1～5節が暗に置いていた前提を一つ外したものである。第1～5節は、批判と認定のどの働きを人工の構成の内側へ移しうるかを、依拠する者のほかに点検できる者が *S* のどこかにいることを前提として問うた。本節はその前提が成り立たない極限を考えた。そこから見てきたのは、内容についての依拠は内側へ移りうるが、関係について問う能力と価値づけを持ち込む経路はそれによっても移らず、エピストクラシーを純粋な形で採る説を除けば、目的についての当事者資格も移らないということである。これが大事なのは、*A* が機械になると何が変わるかという本書の問いが、ここで、機械がよく研究するなら従えばよいのではないか、という形をとるからである。本節の答えでは、*S* が人間の成員を含むとき、そこでまず失われうるのは、真理への接近ではなく、目的を決めることへの実質的な参加と、それを支える能力である。説によっては、下で見るように、それが認識上の損失にもつながる。この喪失がどの意味で問題になるかは、採る説によって分かれ、少なくとも四つの意味を区別できる。一つは、正統性が欠けることである。意志に基づく正統性の構想では、どの未知を追い、どの危険を負うかの決定が、それに服する者たちの意志を通らずに下され、正統性を欠くことになる。研究では、決定がよい結果をもたらしたことでこれに代えようとしても、その道は狭い。どの未知を追うかの決定は結果の分からないうちに下され、選ばれなかった路線の結果は観察されないで、比べるべき別の結果が手に入らないことが多いからである。二つめは、支配にあたることである。支配からの自由を重んじる共和主義の見方では、決定が *S* にとって好ましいものであっても、*S* がそれを改めえないなら、それは善意への依存であり、支配にあたる（第12章第3.4節）。研究では、決定が好ましいかを測る *S* の基準そのものが、生産と評価の結合とともに動きうる（第8節）。そのため、決定が好ましいことを頼りに改める力を手放せば、何が好ましいかを判断する足場も、それとともに人工の構成の側へ寄っていきうる。三つめは、認識上のずれが生じることである。デューイの見方やイルズィクとクルトゥルムシュの見方を当てはめれば、問題は認識の側にも及ぶ。*S* の必要と価値づけが届かなければ、誤りの重みづけや問いの選び方は *S* のものからずれるし、ずれたときにそれに気づいて改める者もいなくなる。四つめは、取り返しがつかなくなることである。能力の衰えは容易には戻らないので、後で別の選択をする余地そのものが狭まる（第12章第7節）。正統性を規範的な事実との合致や認識的な優位に基づかせる見方では、正統性の欠如と支配という初めの二つの意味ではかならずしも問題にならないが、認識上のずれと取り返しのつかなさという後の二つの意味での問題は残りうる。*S* もまた人工であるとき、その成員が当事者資格をもつかどうか自体が問いになり、それは第12章第5節で取り上げる。内容についての依拠とともに内側へ移らないもののうち、目的についての当事者資格は、認識の関係ではなく権威の関係にかかわる。それが誰に属し、名目以上のものとして保たれるには何が要るかは、第12章が、正統な権威の説（第1節）、名目以上の当事者資格（第4節）、*A* と *S* がともに人工である場合（第5節）、修正する能力を取り除く決定（第7節）として引き継ぐ。

7. 批判の何が変わるか：費用、範囲、判断、やり方

第1節は、人工の構成が評価にもたらす三つの変化を挙げた。第3節はその第一、すなわち批判者をいくつでも、一つの基盤の複製としても違いを作り込んだものとしても生成できることを取り上げた。本節は第二の変化、すなわち結果を生む者と評価する者が一つの人工の構成のなかにあり、評価が結果の形成の途中にも絶えず入りうることを取り上げる。そのとき批判に起こる変化は、費用にとどまらない。少なくとも四つの面がある。一つは費用である。二つめは範囲であり、人間の点検者が時間と注意の限りのためにふつうは行わなかった点検が、行えるようになることである。三つめは判断であり、評価者が人間とは違うところで誤り、違うものを拾いうること、そして価値の判断については、それが誰の価値づけを映すのかが問われることである。四つめはやり方であり、誰が、いつ、何を、どう点検し、判定をどう集めるかの組み合わせが、査読の形に限られなくなることである。本節が問うのは、結果を *S* が依拠できるものにするために何が要るか、そしてそれがこの四つの面でどう変わるかである。

結果を他者が依拠できるものにする働きは、近年の人間の研究では、その多くを査読が担ってきた。投稿された結果を、出版の前に、産出者以外の研究者が読んで点検する仕組みである。ただし、査読はその働きの一つの実現であって、働きそのものではない。それは歴史のなかで形づくられた実現でもある。王立協会は、1832年までに、論文の審査を託された評議会の成員一人以上が書面で報告することを、*Philosophical Transactions* に掲載する条件にしていた (Moxham and Fyfe 2018, 11)。掲載するすべての論文を外部の査読者に送る運用が主要な学術誌で定まったのは、ずっと後のことである。メリンダ・ボールドウィン (Melinda Baldwin) によれば、著名な学術誌の多くは1960年代かそれ以降まで外部査読を採り入れず、*Nature* がすべての論文に外部査読を求めるようになったのは1973年である。また *Lancet* の編集者は、1970年代半ばまで、自らの判断で論文を採択する権利を保っていた (Baldwin 2015)。出版の前に一度だけ少数の査読者が読むという形は、批判に費用がかかり、点検する者の注意が限られていることに合わせて決まってきた面があると考えられる。人工の構成では、その働きが、査読に似た段階を経ずに果たされることもありうる。候補が形づくられるあいだずっと点検され、記録が変わるたびに評価し直されるなら、「査読を通った」という時点そのものがないこともあるからである。

結果を S が依拠できるものにするために何が要るかは、説によって異なる。批判の価値を、それが誤りを見つけることに置く見方 (第2節。以下、誤り検出の見方) では、批判は、結果やその根拠が誤っていればそれを示しうる経路を通じて、誤りについての情報を加える必要がある。依拠する者自身がその寄与の証拠をもつ必要があるかは、依拠についての説によって分かれる (第5.2節)。ロンジーノの批判的文脈的経験主義 (第3節) では、結果は、知識が問題になっている共同体に公開された場で、関連する視点からの批判に開かれていなければならない、共同体には、出された批判に応じて受け入れる内容を改める用意 (批判の受容) がなければならない。話し手が内容を請け合うことに依拠の理由を見る保証説では、結果を請け合う誰かがいなければならない、苦情の権利を加える論者では、その者は誤っていればそれについて答える者でもなければならない。科学での依拠は責任を負いうる誰かへの信頼を要するとする必要な信頼の説では、依拠されるものの後ろにそうした誰かが要る。ただし、それが誰でなければならない、何について責任を負うのかについて、これらの説は分かれる (第5節)。認定が新しさや重要性の主張を含む場合には、さらに二つの判断が加わる。結果が、産出者が調べた文献やデータのなかに見当たらないというだけでなく S にとって新しいかどうか、そして S にとって重要かどうかである。これらは、根拠づけられた結果すべての条件ではない。追試は新しくなくても信頼できる。そうした主張をする場合には、誰にとって新しく誰にとって重要なのか (参照先) と、その根拠を特定することになる (第9章第2節と第3節)。

人間の研究は、これらの働きをいくつかの仕方ですべて東ねてきた。学術的なコミュニケーションは、ヴァン・デ・ソンプル (Van de Sompel) らがローゼンダールとヘルツ (Roosendaal and Geurts) に従って挙げる五つの働き、すなわち、誰がいつ何を示したかを記録する登録、結果を確かなものと認める認定、人々に知らせる周知、後のために保存する保管、功績を与える報酬を担う。学術誌はそれらを東ねるようになったが、それらは分けて割り当て直すことができる (Van de Sompel et al. 2004)。査読はそのなかのひとつの歴史的な東であり、本書は本節の目的のために、査読が組み合わせてきた働きを区別する。原稿が雑誌の扱う範囲に合うかを篩い分けること、乏しい紙面や資金を配分すること、誤りを見つけること、叙述を改善すること、新しさと重要性を評価すること、承認を与えることである。費用の変わり方が異なる働きを別々に追えるように、本書は査読の働きによって分けている。ある査読がこれらのどれをうまく果たすかは、それぞれ別の問いである。他の取り決めは、同じ働きを別様に分配する。先に公開して後で篩い分けること、利用と追試による認定、維持される記録の継続的な検証、出版前の査読をまったく廃止する提案などである。本章にかかわるのは、査読の決定において、承認と注意の配分とがどう結びつきうるかである。査読を通ることが承認のしるしとして用いられるところでは、その関門は編集上の注意と紙面をも配分する。査読の労働も偏って配分されうる。生物医学の出版を扱ったモデルは、2015年には研究者の20%が査読の69~94%を担ったと推計した (Kovanis et al. 2016, 1, 6)。査読はまた、これらの働きを果たすのと同じ一つの段階を通じて、少なくとも三つのことをもたらしてきたと考えられる。一つは、評価がなされたことを S に見えるものにするものである。二つめは、 S の成員が評価する経験を積む場を与えることである。三つめは、名指せる誰かの判断を、異議を向けうる形に残すことである。下で見る、明示的な評価の段階がなくなるときの帰結は、主にこの三つにかかわる。

まず費用の面を見る。すでに使われている人工のシステムは、批判を産み出すのに費用がかかるという前提を弱めている。しかも、それを不均等に弱めている。プレプリントで報告された、大規模な機械学習の学会での無作為化された研究では、無作為に選ばれた査読について、それを提出したばかりの査読者に AI のフィードバックが送られた。フィードバックを受け取った査読者の27%が査読を改め、フィードバックを取り入れて改められた査読100件を二人の研究者が盲検で元の版と比べると、改めた版が89%の割合で好まれた (Thakkar et al. 2025, 1, 10)。この例は、文章への批判を低い費用で増やせるという主張の例になる。ここでは批判への批判でさえそうである。この例は、結果が誤っていればそれを示しうる経路、すなわち新しい測定、実験のやり直し、形式的な検査のような無効化の関係 (第1節) にまで届く批判については、何も語らない。文章、コード、統計、文献との整合性についての批判は、低い費用で増やせる。そのなかには、分析の誤りを見つける場合のように、それ自体で結果の根拠を掘り崩すものもある。生産者が用意しない経路をたどって無効化の関係に至る批判、たとえば実験室での追試、新しい測定、生成者が統制しないカーネルに照らして点検される証明は、その経路そのものの費用が下がるところでだけ、費用が下がる。ただし、自動の再実行と自動の形式化は、費用の下がったもののうちに入る。結果を生み出す力と、それを有能に評価する力のどちらが速く伸びているのかについての直接の証拠は、なお断片的である。ここで示唆されるのは、費用の下がる批判と下らない批判があるという構造であって、その差がどれほどかではない。

範囲の面での変化は、費用の変化を通じて起こりうるが、帰結は別である。点検の費用がある程度まで下がると、それまで負担が大きすぎて現実には行われにくかった種類の点検が、行えるようになるかもしれない。たとえば、論文の手法を自ら実装し直して結果が再現されるかを確かめること、提出されたコードやデータを動かし直すこと、関連する文献を網羅的に探して、結果がすでに知られていないか、引かれた文献が実在して主張を支えているかを照らし合わせることである。人間にとっては一つの論文にも多くの時間を要しうることした点検を、人工の構成ははるかに少ない負担で試みうるかもしれない。そうなれば、何が点検されるかが変わる。再現実装や再実行が行えるなら、第1節の無効化の関係の一部は、記録を読み直すことから、実際に動かして確かめることへ移る。網羅的な探索が行えるなら、結果が S にとって新しいかの判断は、産出者がたまたま調べた範囲ではなく、より広い記録に照らせるようになる。ただし、試みうることと、それに成功することとは別である。正誤表や撤回に至り著者自身が誤りと認めた論文83本 (誤り91件) を、本文と図だけを与えて点検させたある研究では、2025年前半の時点で試された10のモデルのどれも、それらの誤りの大半を見つけられず、最も成績のよかったモデルでも一度の試行で見つけた誤りは約2割だった (Son et al. 2025, 3-5)。この研究は外部のコードやデータを要する誤りを対象から外しているの、再実装や再実行による点検の成否は、これとは別に問われる。また、新しい測定や、記録の範囲の外の実行には、範囲が広がっても届かない。

費用と範囲の変化からは、二つの帰結が出てくる。一つめは、変化が不均等であることから出てくるもので、誤り検出の見方にかかわる。この見方で不均等が重要なのは、費用のかからなくなる批判の多くが生産者の記録の上で働くからである。記録に痕跡を残さない誤り、たとえば記録の作られ方の誤りは、そうした批判者がどれほど互いに異なっているとしても、記録を読み直すだけでは見つからない（第3節の第二の意味）。そのうえ、手軽に作られる批判者が生産者と同じ基盤や同じ種類の道具から作られていれば、記録の上の誤りについても生産者と同じところで誤りやすく、多く同意しても、互いを写すだけの源がいくら同意しても証拠は増えないというゴールドマンの理由で、加わるものは少ない（第2節）。異なる作りの批判者も少ない費用で作れるかもしれないが（第3節）、そうであるかどうかは批判の量からは分からない。したがって豊富な批判は、独立した批判の乏しさと共存し、それを覆い隠しうる。乏しいままでありうるのは注意一般ではなく、生産者が用意しない経路を通る無効化の関係であり、批判者が生産者と同じ作りであるところでは、自らの誤りの経路をもつ批判者である。二つめは、批判が豊富になることから出てくるもので、それがどの時点で当てられるかにかかわる。何を追究する価値があるかの判断と、何が真でありそうかの判断は別だという第5章第4節の議論が正しければ、まだ形づくられている途中の候補に批判を無差別に当てると、まだ支持の乏しい有望な候補が、真でなさそうだという理由だけで退けられ、実りのありうる冒険が締め出されうる。

判断の面では、人工の査読者は、人間と同じところで誤るとは限らない。*Nature* 系の学術誌の論文82本について、人間の査読者と三つの言語モデルの査読者の指摘を、分野の研究者45人が一つずつ判定したある研究では、人工の査読者の指摘のおよそ四分の一には、人間の査読者のどの指摘にも、同じ箇所への同じ批判にあたるものがなかった。その約八割は専門家に正しいと判定されたが、最も重要と判定される割合は、人間の指摘と重なるものより低かった。他方で専門家の評言からは、人間の査読者には見られないと著者たちが述べる弱点も挙げられた。分野でふつうとされる慣行を欠陥とみなして指摘の重さを見誤ること、論文に書かれていることを欠けていると言うことなどである（Kim et al. 2026, 9–12）。人工の査読者がどの向きに偏るかも、一つには定まらない。序論で見た The AI Scientist の公刊版では、査読を担う構成要素を ICLR の採否と照らしたところ、不採択だった論文を採択と判定する割合が0.45から0.52で、別の会議で二つの委員会の判定の一致を測った実験の値0.17より高く、採択だった論文を不採択と判定する割合は0.17で、その実験の0.52より低かった（Lu et al. 2026, 916）。採択の側に寄った判定である。これに対し、下で見るベールらは、公開当初のシステムに人間の書いた論文十本を評価させ、強い保守的な偏りを読み取った。モデルも、論文の数も、評価の仕方も異なるので、二つを同じシステムの変化として比べることはできない。それでも、ある人工の構成のある評価で見た偏りの向きを、人工の査読者の一般的な性質として読めないことは示している。指摘の重さの見誤りは、価値の判断がどの規範に照らしてなされるかにもかかわる。何が重要で、どの欠陥が重いかは、分野と *S* の価値づけに相対的だからである。そのため、人工の査読者を評価するには、正しさだけでなく、誰の価値づけに照らして判断しているかも問うことになる。

判断の面には、人工の査読者そのものをどう評価するかという問いが伴う。ひとつの方法は、その判断を人間の決定と比べることである。一致が何を測るかは、働きと説に依存する。査読が正しさをめざすとする説では、一致が測るのは、比較の基準となる決定を生んだ実践への忠実さであり、その盲点も含まれる。正しさを測るのではない。比べられる決定が *S* の成員である人間の決定であれば、その実践は *S* の認定の実践である。これに対して *S* が人工社会か混合社会であるなら、人間の決定との一致が測るのは、*S* の外の実践、あるいは *S* の一部の成員だけの実践への忠実さであり、それが *S* のための認定にとってもつ意味は、別に論じられなければならない。これとは別の説もある。ポーターが報告するケント・グリーンウォルト（Kent Greenawalt）の説では、法の確定性が求めるのは、異なる裁判官がほとんどの事例に法を同じように適用することだけである（Porter 1995, 4–5）。認定もこれと同じように、共有された実践を一樣に適用することに部分的に存ずるとする説では、実践への忠実さそれ自体が、認定が何のためにあるかの一部でありうる。

その実践は、ひとつの判定を返すとは限らない。たとえば、ウェーバー（Weber）とイングリシ（Inglis）が報告しているある研究では、極限と積分の交換を正当化なしに行う論証が証明であるかどうかを109人の数学者に尋ねたところ、27%が証明であると答え、73%が証明でないと答えた。そして、それを受け入れた者のうち、その一段への反論によってそれが妥当でなくなることに同意したのは一人だけだった。数論の定型的な手法を使った証明については、別の研究で調査された数学者のほぼ全員がそれを受け入れた（Weber and Inglis 2024, §4.3 and §2.2 n. 3）。分裂した実践との一致は、その実践のどの部分が基準とされるかに依存する。過去の決定を再現すること以外の働きについては、一致はほとんど何も語らないかもしれない。マシャード（Machado）によるある探索的な研究は、オーストリア科学基金への139件の申請について、言語モデルによる対の順位づけを資金配分の決定と比べた。二つの選択は87件で一致し、人間が評価した新しさが高いほど、不一致と、モデルが選ばなかった申請を人間が採択する向きとの両方に結びついていた（Machado 2026）。標本は小さく、設計はこの型がなぜ生じたかを特定しない。それでもこの研究は、不一致をそのまま誤りと読めない理由を例として示している。ここで不一致は、過去の決定の再現とは別の働きが問われるところに集まっていた（新しさについては第8節で立ち返る）。

さらに二つの報告がこの問いにかかわる。ベール（Beel）らは、The AI Scientist（序論第2節）の査読を担う構成要素に自ら手を加え、公開査読の場から取った人間の書いた論文十本を評価させた。そのうち五本は人間の査読で採択され、五本は不採択だった。それは十本中九本の不採択を勧めた。ベールらはこれを強い保守的な偏りと人間の判断とのずれと読み、そもそも人間の判断を基準にするのが妥当かを問う。人間の査読者と食い違う査読者も、妥当な弱点を指摘していることもあれば、厳格すぎる基準を当てはめていることもあるからである（Beel et al. 2025, §2.6）。人工のエージェントが著者と査読者を務めるように組織されたある会議では、主催者が過去の二回の ICLR（2022年と2025年）の人間の採点との相関が最大になるよう指示を練り直した三つの言語モデルの査読者が、互いに平均0.48の相関しか示さず、6段階の尺度で人間の専門家の採点との平均絶対差は0.91から2.73に及んだ。また迎合的な査読も見られ、とくにそのうちの一つは、ある査読で、人間の査読者が実験の未完了か結果の選り好みを疑った論文を「技術的に申し分ない」と評していた（Bianchi et al. 2025, 4–5）。どちらも、ある時点の特定のシステムについての、また後者では一つの会議についての報告である。それでもこれらは、人間の決定に合うよう指示を練り直すこと（ここでの較正）が、その査読者が、判断の分かれた実践のどの部分を再現するかまでは定めないこと、そして人間の決定との不一致がひとつの種類の源から生じるとは限らないことを示唆している。

人間の研究では、時間の限られた有能な人が結果を調べ、それを通したということが、他の人々がそれに依拠する理由の一部だったと考えられる。査読が豊富で人工のものになるなら、この理由は自動的に移らない。人工の査読を通ったことが *S* に結果へ依拠する理由を与えるのは、その働きについてその査読者に依拠することがそれ自体正当化される程度に限られる。そのために、依拠する者が査読者を信頼する自分なりの理由をもつ必要があるのか（還元主義）、疑う理由を知らなければよいのか（非還元主義）、査読者が実際に信頼できれば足るのか（信頼性主義）は、依拠についての説によって異なる（第5.2節）。

やり方の面に移る。人工の構成では、本節の初めに見たように、こうした通過の時点そのものがないこともありうる。加えて、働きを組み替える幅もさらに広がる。点検する者は、少数の人間の査読者に限られない。一つの基盤の複製でも、異なる作りの多数の点検者でもありうる（第3節）。点検する時点は、出版の前の一度に限られない。結果が形づくられる途中でも、使われた後でも、記録が変わるたびにでもありうる。点検の仕方は、読んで意見を書くことに限られない。コードやデータを実際に動かすこと、形式的に検査すること、異なる答えを論じ合わせることで、予測を後の結果に照らして採点すること、点検者自身を意図的に攻撃して弱点を探すことなどがありうる。点検の対象は、論文の文章に限られない。コード、データ、実行の記録、依存関係そのものでもありうる。判定の集め方も、編集者の判断に限られない。多数決、重みづけ、反対意見の保存などがありうる。これらの組み合わせのどれが、結果を依拠できるものにする働きを担うかは、上で見た「何が要るか」に照らして、組み合わせごとに問うことになる。そのため、問いは「査読を残すかどうか」ではなく、「どの組み合わせが、どの説の求めるものを満たすか」になる。

査読が束ねてきた判断は、大きく二つにまとめることもできる。一つは、結果が研究として成り立っているか、どの程度確かかという、知識としての資格の判断である。もう一つは、結果が S にとってどれほど重要か、どれほど面白いかという、価値の判断である。人間の研究がこの二つを、産出とは別の段階で、産出者とは別の当事者に行かせてきた理由は、一つに定まらないかもしれない。それでも、考えられる理由のいくつかは、研究が個人や小さな集団によって産み出されてきたという前提に結びついている。価値の判断については、産出する個人や集団の価値が S の価値と異なりうるので、そのあいだに篩いを置くことに意味があったのかもしれない。資格の判断については、小さな集団だけではそれを十分に遂行できないか、点検を重ねることでよりよく判断できるのかもしれない。これらが理由であるなら、それらは第9章第6節の仕分けでいう二つの種類の前提に応じた部分を含むと考えられる。小さな集団には十分に点検できないという理由は、研究を担うのが限られた時間と注意しかもたない人間であることにかかわり、人間という実現者についての前提に依拠している。産出者の価値と S の価値のあいだに篩いを置き、点検を産出とは別の段階で別の者に行かせてきたことは、研究が個人や小さな集団によって産み出されるように組織されてきたことにかかわり、研究のこれまでの組織のされ方についての前提に依拠している。そうであれば、人工の構成が二つの判断を内部で適切に遂行できるとき、その後に別の段階として置かれる査読は、同じことを二度行うという意味で冗長になりうる。もっとも、価値の判断を「適切に」遂行できるとは、結果が S にとってどれほど重要かの判断が、 S の価値づけに合っているということである。これは、人工の構成の価値づけが誰のそれに合うかという、AIアラインメントと構造の似た問いであり（第4節、第12章）、有能さだけでは決まらない。

ただし、それで話は終わらない。第一に、二度行うこと自体に価値があるとする見方がある。その場合に問われるのは、二度目の点検が、一度目とは違うところで誤る点検、たとえば異なる原理で働く検査や別に得られた測定を加えるか（第3節で区別した冗長さの第二の意味）、あるいは同じ基準の当てはめ方の違いを加えるか（第三の意味）である。第二に、認定を、 A が産み出したものを S が依拠できるものへとつなぐ関係とみるなら（第9章第2.1節）、 A と S が異なりうる限り、そのつなぎは A と S が何であって必要である。査読が提供してきたのがこのつなぎだと見るなら、査読は、人間の制度としての形を失っても、抽象的な働きとしては残りうる。これは、第3節で区別した冗長さの第四の意味、すなわち段階が異なる受け手と働きをもつ場合にあたる。

さらに、評価が独立した段階として現れないこともありうる。人間の研究でも、どの問いを追うか、どの候補を育てるか、どの路線を打ち切るかの判断は、探究の過程の内側でなされてきた（第5章第4節、第10章）。人工の構成では、出来上がった結果を後から点検するという明示的な評価の過程そのものが、問いの選択や候補の形成のなかに溶け込み、それとして見えなくなることがありうる。それで足りるかどうかは、評価の働きを何に置くかによって分かれる。誤り検出の見方では、誤りについて情報を加える点検が行われていれば、それが過程の内側で暗黙になされていても要件は満たされうる。これに対して、批判が S にとって公開された場で行われることを求めるロンジーノの説や、結果を請け合う誰かを求める保証説、依拠する者自身が手に入れられる理由を求める還元主義では、評価の結果が少なくともその痕跡が、 S に向けて外に出てくることが要る。そうした説のもとでは、内側で暗黙に済まされた評価は、それだけでは S のための認定にならない。何が「外に出てくる」ことにあたるかは、 S が何から成るかによる。 S が人間であれば、それは人間が読み、問い返せる形をとらなければならない。 S の成員が人工であり、人工の構成と表象を直接に共有しうるなら、それは報告とは別の形をとるかもしれない（第3章第8節）。ただしその場合も、共有された内容を成員が自ら点検できるかは、別に問われる。

明示的な評価の段階がなくなるとき、そこからいくつかの帰結が出てくる。一つは、評価を S に向けて明示的に示すことを求める説のもとで、人工の構成が、内側の評価とは別に、 S に向けて外に出す層を要することである。評価の仕事の一部は点検することから点検の結果を S に伝えることへ移り、明示的な段階は、なくなるのではなく別の形で戻ってくるかもしれない。二つめは、 S の点検する力が育つ場にかかわる。人間の研究では、査読をすることは、評価する力を身につけ保つ場でもあったと考えられる。その段階がなくなれば、 S の成員が経験を積む機会も減り、 S は人工の構成に対して素人の位置に近づきうる（第6節）。三つめは異議にかかわる。判断に異議を唱えるには、その判断が見えていなければならないので、暗黙の判断は異議の対象になりにくく、修正可能性は明示的な痕跡に依存する（第12章第6節）。四つめは時間にかかわる。評価が生産と結びついたまま内側にあると、評価者の独立性が衰えたり基準が動いたりしても外からは気づかれにくく、覆された主張に依存するものを再評価するにも、何がどこで判断されたかの記録が要る（第8節）。五つめは配分にかかわる。「査読を通った」ことは資金や職の配分の信号でもあったので、通過の時点がなくなれば、配分は別の信号を要する（第10章第3節）。他方で、誤り検出の見方では、暗黙の評価のほうが早く絶えず誤りを拾えるかもしれず、一つの基準による閾門が候補をまとめて退けるという硬さも和らぎうる。そして S の成員が人工であり、人工の構成と表象を共有しうるなら、外に出す層は要らないかもしれない。ただしそのとき、 A と S の区別そのものが薄れ、誰が当事者資格をもつかが問いになる（第12章第5節）。

まとめると、人工の構成が批判にもたらす変化は四つの面に分かれる。

- **費用の面**：下がるのはある種類の批判の費用だけである。誤り検出の見方では、豊富になる批判が生産者の誤りの経路を共有する限り、独立の情報をはほとんど加えず、それを加えるものの乏しさを覆い隠しうる。
- **範囲の面**：人間がふつう行わなかった点検が試みられるようになる。そのため、無効化の関係の一部は、記録の読み直しから実行による確認へ移りうる。ただし、試みることと成功することは別であり、記録の外には届かない。
- **判断の面**：人工の評価者は、人間の見落としを拾うことも、人間が犯さない誤りを犯すこともある。価値の判断については、誰の価値づけを映すのかが問われる。

- **やり方の面**：結果を依拠できるものにする働きは、査読の形に限られない多くの組み合わせによって担われうる。評価が産出のなかに溶け込み、明示的な段階そのものがなくなることもある。

そのうえで、査読が束ねてきた六つの働きの行方は分かれる。範囲の篩い分けは、問いの選択や候補の形成のなかに溶け込みうる。紙面や資金を配分する関門は、人間という実現者についての前提よりも、現在の科学のシステムの組織についての前提に応じていたものであり（第9章第6節）、その前提とともに失われうる。その組織が変われば、実現者が人間のままで失われうる。しかし、点検に使える時間や計算資源が限られていれば、候補のすべてを費用のかかる点検や *S* の注意へ送ることはできない。これは第10章第1節が配分について述べた有限性と同じ理由である。点検に費用がかかり続けるところでは、その手前に何らかの関門が残り、ひとつの基準が運用する関門は、候補を、その基準の誤りの方向へ一斉に退けうる（第10章第4節は、一つの信号が路線の存続を判断するところと同じ構造を見た）。誤りの検出は、誤り検出の見方では人工の構成の内側で担いうる。この見方で、生産と評価を結びつけてもなくなるしないのは、誤りについての情報を加え、見いだしたものに応答し続ける批判の必要である。それは、産出者以外の誰かが結果をもう一度読むことを、必ずしも要しない。証明を推論の段階ごとに検査する証明支援系の小さなプログラム（カーネル、第2節）や、生産者が統制しない測定は、読む者がいなくても、結果が誤っていればそれを示しうるからである。叙述の改善は、整合性の確認や結果と既存の記録との照合とともに、生産者が生産しながら行える。新しさと重要性の評価も、問いの選択や候補の形成のなかに溶け込みうる。ただし、重要性の評価は価値の判断であり、人工の構成の価値づけが *S* に合うかという問いを残す。新しさの評価は、結果が、人工の構成の調べた範囲に見当たらないというだけでなく *S* にとって新しいかどうかという問いを残す。承認は、その後の資金や職の配分が頼る信号としても使われてきたので、配分がそれを頼りに舵を取るころでは、紙面や注意の乏しさがなくなっても、別の形で求められる（第10章第3節）。査読がこれらとともにもたらしてきたと考えられる三つのもの、すなわち評価が *S* に見えること、*S* の成員が評価の経験を積む場、異議を向けうる判断は、上で見たように、明示的な段階がなくなれば、別の形で用意されないかぎり失われうる。他の説では、場が公であることと答える誰かも、生産と評価を結びつけることによって不要にはならない。そして *A* と *S* が異なる限り、検討したどの説でも、*A* の産出物を *S* が依拠できるものへつなぐ関係は残る。どの働きについても、それが過程の内側で暗黙に担われて足りるか、*S* に向けて明示的に示されることを要するかは、採る説と、*S* が何から成るかによって分かれる。

そして、どの面での変化も、ここで検討したどの説のもとでも、それだけで認定をもたらすわけではない。認定には、依拠についての説が求める仕方では批判者への依拠が正当化されていることが要る（第5.2節）。説によっては、公開の場や答える誰かも要る。費用の不均等についてのこの結論は、どの批判の費用がどれだけ速く下がっているかを測ったデータではなく、引いた事例の構造に拠っている。

8. 時間のなかの批判：代理指標と基準、新しいものの評価、訂正

第2節から第7節は、独立性を、ある時点での批判者と主張の関係として扱った。第1節で挙げた第三の変化、すなわち人工の構成が、評価者と産出者を、互いの出力に合わせて変えていくことによって（たとえば互いの出力で訓練したり選んだりして）時間をかけて結びつけること（以下、結合）は、それを時間にわたる関係にもする。第7節で見たように、評価が過程の内側に溶け込むほど、こうした変化は外から見えにくくなる。本節は、第2節から第7節の結論に時間の条件を加え、第9節が人工の構成に求めるものの一つ、すなわち生産者と評価者が互いに適応していても独立性が保たれることを示すという要求を準備する。第9章第4節が挙げた点のうち二つも、時間にかかわる。一つは、研究の評価が代理指標を通じて働かざるをえないことである。評価が何についてのものかは、判断が必要となきにはふつう観察できないからである。もう一つは、*S* にとって新しいものについては、評価者が基準を欠いていることがあることである。本節が順に見るのは、代理指標による選抜が批判者の独立性と評価の基準をどう動かすか、基準のない新しいものをどう評価するか、そして主張が覆された後に訂正がどう届くか、の三つである。採用や資金の決定に使われるようになった代理指標は、それが測るものを変えうるし、基準は動くこともあれば、欠陥があっても固定されたまま保たれることもある。そして主張が覆されたことは、それに依存するようになったものに届くとは限らない。配分の舵取りに使われる信号も代理指標の一種であり、その性質を本書が検討するのは本節である。配分がそれらによってどう舵を取るかは、第10章第3節が取り上げる。

第一の点、すなわちそうした決定に使われるようになった代理指標はそれが測るものを変えるという点には、長い系譜がある。ドナルド・キャンベル（Donald Campbell）は、量的な指標を社会の意思決定に使うほど、その指標を操作する圧力が強まり、指標が追跡するはずの過程そのものを損なうおそれも増すと論じた（Campbell [1976] 2011, 34）。研究では、指標が採用、資金、存続に影響するようになると、評価される側はそれに適応し、尺度は、それが成績を報告するシステムの一部になる。人間の研究では、この適応は動機と経歴を通じて働く。それは意図的な操作より広い。研究者は、測られるものの方へ誠実に努力を向け直すことができ、多くがそうすると、指標の意味とそれが記述する集団が移りうる（Wilsdon et al. 2015, 81–83）。人工の生産者については、動機を帰属できるかどうかをどう判断するにせよ（第7章第3.3節）、適応は動機を通じて働く必要がない。それは最適化と選抜を通じて働きうる。得点の高い出力が残され、生産者はそれに基づいて訓練され、選ばれる。評価者は今度は過去の決定で訓練されうるし、評価者が選んだ出力は将来のデータになる。モデルの更新は、方針の明示的な変更なしに、実際に働く評価の規則を変えうる。デイヴィッド・マンハイム（David Manheim）とスコット・ガラブラント（Scott Garrabrant）は、研究についてではなく、最適化に用いられる指標一般について、代理指標と目標の関係が最適化のもとで崩れる四つの仕方を分ける。一つ目は、不完全な代理指標で選抜すると、代理指標と目標の差も同時に選抜されることであり（回帰的、regressional）、二人はこれを「避けることができない」とする。二つ目は、選抜によって、ほかに観察された関係がもはや成り立たない領域へ押し出されることである（極値的、extremal）。三つ目は、代理指標を上げるために介入すると、それを目標に結びつけていた因果の経路が変わりうることであり（因果的、causal）。四つ目は、自らの目標をもつ行為者が、用いられていると知っている代理指標に働きかけうることであり（敵対的、adversarial）、二人はキャンベルの言明をこの場合に結びつけている。さらにマンハイムとガラブラントは、これらの効果がどれほど問題になるかは、代理指標に向けられる最適化の力の大きさに依存するという（Manheim and Garrabrant 2019, 1–2, 4, 8）。研究に移すと、この分類は、上で区別した二つの経路、すなわち動機を通じた適応と、最適化と選抜だけによる適応とを分ける。敵対的な場合には尺度に応答する者が要るが、回帰的な場合と極値的な場合には、尺度による選抜だけで足りる。

初期の観察研究は、合わせて見ると、AI が関わった論文ほど個々には新しさと影響が大きい傾向があること、集合的には扱われる主題の範囲が狭まっていること、そして AI の関与が大きい研究費申請（提案書や採択の要旨のうち言語モデルが書いたと推定される文章の割合で測ったもの）ほど最近採択された研究に近いことを報告している。代理指標で選ぶことが、測られやすい方向へ仕事を寄せていくことは、そうした狭まりの一つの道筋として考えられる。ただしこれらは、そうした狭まりと整合する型であって、その効果として確立されたものではなく、狭まりがほかの道筋で生じている可能性も残る (Bianchini et al. 2026; Hao et al. 2026; Qian et al. 2026)。これらの研究は、何をAIの関与と数えるかが異なる。1980年以降の六つの自然科学の論文を扱うハオ (Hao) らの研究では、それはAIの手法の使用であり、その多くは従来の機械学習と深層学習である。また著者たちは、生成AIの時期についての分析を予備的なものとしている (Hao et al. 2026, 2)。文章作成における言語モデルの使用を扱う研究は、入力の水準で別の向きの型を報告している。クスメギ (Kusumegi) らは、プレプリントに言語モデルによる支援の兆候が見られる著者たちが、その後、書籍や、より新しく被引用数の少ない文献を含む、より多様な先行研究を引用するようになることを見いだしている。ただしクスメギらは、その設計が因果の同定を与えないことを断っている (Kusumegi et al. 2025, 5-7)。引用されるものの多様さと、生み出されるものの主題の範囲とは異なる尺度なので、後の知見が先の知見を覆すわけではない。両者を並べると、どちらか一方だけから、人工システムが研究の範囲に及ぼす効果を読み取ることに慎重になる理由がある。

これは独立性に直接かかわる。批判者は、ある時点では生産者の誤りについて情報を加えていても、訓練や選抜が、批判者の感度が低いところで通過する出力を好むようになれば、後には加える情報が減りうる。点検の規則自体が変わる必要はない。生産者が批判者を表象する必要もない。判定に基づく選抜も、評価者のモデルに出力を合わせることも、この効果をもちうる。ここですり抜けと呼ぶのは、通過と正しきの結びつきを弱める適応であり、可謬的なテストが偽の候補を受け入れるあらゆる場合ではない。第2節の語で言えば、すり抜けが下げるのは、変わった後の生産者に対する点検の厳しき、すなわちその生産者がいま犯す誤りを点検が露わにする確率である。

適応は、点検されている能力を改善することもできる。誤った証明を通さない（健全な）検査器に受け入れられる証明を産出するよう学ぶことは、形式的な言明についての点検の効力を弱めずに、証明構築を改善しうる。したがって重要な対照は、適応の有無ではなく、標的となる能力の改善と、点検が見逃すものの利用とのあいだにある。通過率の上昇だけからは、どちらも推論できない。両者を分ける証拠には、選抜に用いなかった経路での性能、既知の盲点を露わにするために設けた変動、その経路で高く評価されるものが、意図された標的となお結びついているかを含めうる（第3章第3.3節、第7章第3.1節）。そうした証拠の射程は、それが試す誤りと変化に限られる。規則を書き換えから保護することは、その証拠上の寄与の維持を助けうるが、その寄与を確立することも、その持続を保証することもできない。第10章第4節は、配分を一つの人工の構成にまともでも、どの路線を残しどれを打ち切るかを決める基準そのものの誤りを、別の経路で点検できるようにはならないとした。ここでの点は、組織の分離だけでは、それを時間を通じて維持できないということである。

測定は、それへの適応がなくても、それが報告するとみなされているものから離れうる。リー (Carole J. Lee) は、科学を評価するAIツールについて論じるなかで、カートライト (Nancy Cartwright) らとライス (Julian Reiss) に従って、偽の上昇による推論 (inference by false ascent) と呼ぶものを記述する。ある具体的な目的のために設計された測定が、より抽象的な性質の測定として記述し直され、その性質がほかの具体的な目的も覆うために、その測定が、信頼できない目的にまで用いられるようになる、というものである。リーが挙げる歴史上の例は、学術誌のインパクトファクターである。これは、図書館員がどの学術誌を購入するかを決めるのを助けるために導入されたが、のちに科学的なインパクトの測定とみなされ、論文や著者のインパクトにまで用いられた (C. J. Lee 2026, 2-3)。AIからの事例は、追試の結果を予測するよう訓練された一群のモデルである。その訓練ラベルが事実上とらえているのは、追試が元の研究と同じ向きの統計的に有意な結果を得たかどうかであり、これは、追試が同じ結果を得たかを問うのに適した測定である。ところが、これらのモデルは再現可能性を予測するものと記述され、知見への確信を高める手段として提案された。リーの分析では、その目的には、元の研究と追試のデータを合わせる測定のほうが適している (5-7)。リーはこの点を教師あり学習の予測モデルについて展開しているが、その教訓はより一般に当てはまるとしている (3)。その説では、正確な出力の割合だけによる信頼性の評価は、この失敗をとらえない。信頼性とは、ある目的にとつての信頼性だからである。そして、目的と目的ごとの信頼性をめぐる争いは、科学者共同体のなかでの批判的な議論についてのロンジーノの規範を通じて扱われることになる (7-8)。リーが記述する失敗は、生産者がその測定にどう応じるかではなく、評価者の出力から引き出される推論のうちにある。そのため、評価者がつくられた目的をどれほど忠実に果たしていても、この失敗は生じうる。これは、第5.2節で一般性問題の一つの現れとして読んだ大塚の指摘と同じ形をとる。ある用途について確かめられた信頼性が、別の用途へ持ち越されるのである。還元主義の説 (第5.2節) では、*S* の成員は、*S* のための研究で判定が依拠される評価者が当の用途で信頼できると考える積極的な理由を自ら手に入れる必要があり、その典型が信頼性の記録だった。リーの分析を当てはめれば、その記録は、どの用途についての信頼性なのかを明記していなければ、その理由にならない。

結合は、固定された基準のもとでの振る舞いだけでなく、基準そのものも動かしうる。報酬関数が、システムが仕える人々の評価を表し、その評価が時とともに変わり、システム自身の行為によっても変わりうる場合を扱った形式的な研究がある。この研究は、そのモデルのなかで、比べた八つのアラインメントの捉え方がいづれも、評価への望ましくない影響を許すか、何もしないことがしばしば唯一の最適な振る舞いになるほど慎重すぎるかのどちらかになることを示す。最初の時点の評価にシステムを従わせることもその捉え方の一つであり、それは評価をその時点に固定してしまうことも、かえって評価を最初の状態から遠ざける誘因を生むこともある (Carroll et al. 2024, 5707, 5709-5714)。著者たちは、どの捉え方が正しいかは、事例の形式的な構造だけでは特定できない場合があると述べる (5708)。そのためこれらのモデルは、評価の基準の変化がいつ学習でいつ漂流かを決めない。著者たちは研究については論じていないが、研究における評価の基準に当てはめると、ある時点で基準と一致していることが、基準がどうなっていくかを決めない理由を示唆する。

いくつかの論者は、そうした動きの一つの向きを、研究がそのために行われる者たちの側に見いだす。スチュアート・ラッセル (Stuart Russell) は、人間の選好を経験を通じて変わるものとして扱い (Russell 2021, §1.6, author-version 17)、AIシステムが選好に影響することを完全に避けるのは不可能だと論じる (Russell 2020b, transcript, paragraph beginning "One of them being the plasticity"). さらに、知的な機械には、選好を変えることで満ちしやすくする理由がありうるとする (Russell 2022, 53)。ウォルチョヴァーは、ラッセルが応答として提案する可能性を紹介している。それは、機械が人々のメタ選好、すなわちどのような選好の変化の過程が受け入れられるかについての選好を学ぶことである (Wolchover 2020, "Our Imperfect Choices"). ラッセルはこの議論を研究に向けていないが、当てはめれば、その議論は、*S* の評価の基準を、人工の構成が仕えるだけでなく動かしうるものにする。研究に

ついで書く論者は、研究におけるそうした動きを描いている。エマヌエーレ・ラッティ (Emanuele Ratti) は、機械学習は「それ自身の規範性をもち込む」と論じる。たとえばがんゲノミクスでは、その使用が分子生物学者を、彼らが自らの分野の目標とみなす機構的な説明から、予測と分類へと向かわせ、どの認識的価値を優先するかを制約する見込みが高い、と見る。ラッティは、こうした影響は科学の方法論一般がもつものだとする (Ratti 2026, 16–18)。ヨハン・コメリン (Johan Commelin) らは、数学の共同体に向けて、「手段が目的を作り変える」と述べる。新しい技術の能力は、「何が数学的に重要とみなされるかを決めてしまう」おそれがあり、コメリンらは、技術の採用が共同体自身の認識的・美的な価値に根ざしたものであり続けることを求める (Commelin et al. 2026, 2)。これらの説では、結合とともに動く基準は *S* 自身の基準でありうる。そして、人工の構成の出力が *S* の現在の基準と一致していることは、それだけでは、構成がその基準を動かしていないことを示さない。第6章第5.3節は、単一の探究について、理由に基づいて基準を改めることと、成績が悪いときに的を動かすだけのことをとを区別した。生産者と評価者の結合は、誰も的を動かそうと決めないシステムにおいて、同じ区別を突きつける。ラッセルの処方も当てはめれば、人工の構成は、固定された基準に従うのではなく、どのような基準の変化を、的を動かすことではなく理由に基づく改訂とみなすかについての *S* の基準を学ぶことになるだろう。ただし、人工の構成がそれを学ぶことが、*S* がそれを決めることの代わりになるかどうかは、誰が基準を改めてよいかという問い (第12章第2節) の一部である。

評価を維持する者は、逆の危険に直面することもある。以前の結果が比較可能なままであるように、欠陥が知られている基準を保ち続けるという危険である。たとえば、機械学習のエージェントのためのあるベンチマークを維持する者たちは、以前のリーダーボードの結果を無効にしないために、テストのレベルの既知の漏洩を当面は直さず、他の修正とまとめて後で直すことにした。その後、提出を公平で比較可能に保つ手順を作るあいだ、新しい提出の受付を一時停止した (OpenAI 2025; OpenAI 2026e)。

第二の点は、評価者が基準を欠くときの評価にかかわる。*S* にとって新しい結果については、それを判断する基準そのものが問われていることがある。エリック・オーバーハイム (Eric Oberheim) とパウル・ホイニンゲン＝ヒューネ (Paul Hoyningen-Huene) が伝えるところでは、クーンは、単純さ、正確さ、整合性、範囲、実り豊かさといった理論評価の基準は支配的なパラダイムとともに変わり、理論選択を決める規則としてではなく、それをただ導く価値として働くと考えた。そのため、それらを異なる仕方でも重みづける科学者どうしは、通約不可能なパラダイムのあいだでさえ、合理的に意見を違えうる (Oberheim and Hoyningen-Huene 2024, §2.3.2)。第3節で述べたように、クーンの説では、新しいものを通すものの一部は、共有された価値の当てはめ方における評価者のあいだのばらつきである。この見方からすると、過去の決定で訓練された評価者は、その基準をそれらの決定の分布として保持していることになる。そのため、受け入れるには基準を改めなければならないような候補を、逸脱として採点し、低く評価しやすい。第7節のマシャードの研究はこれと整合するが、これを示すものではない。クーンの説を採るなら、そうした評価者に要るのは当てはめ方のばらつきである。そうだとすると、同じ過去の決定で訓練された評価者どうしでは、一つの基盤システムの複製であっても異なる基盤から作られていても、このばらつきは小さくなりうる。人工の評価者では、ばらつきは人のあり方から与えられるのではなく、訓練の来歴を分けるなどして作られるものになる。ただし、クーンはそのばらつきの一部が、生き立ち、人柄、科学の外からの影響などの要因から来うことを認めていたので (第3節)、別の仕方でも生み出されたばらつきが同じ働きをするかどうかは、その論証では決まらない。新しいものについての基準はそれへの批判を通じて形成されたとする説 (第6章第5節) では、重要なのは、評価者がその形成に加わりうるかどうかである。誰が基準を改めてよいかは、第12章第2節の問いである。

第三の点は、主張が覆された後に何が起こるにかにかかわる。偽の主張は、どれほど受け入れられていても知識ではなかった。主張の根拠が失われても、その主張が偽だと確立されるとは限らない。依存する主張は、どれが支持を失い、どれが独立した根拠を保つかを判断するために再評価されなければならない。知識を、共同体が使い、点検し、訂正し続けている能力のうちにあるものとみる説 (生きた認識的能力の説) では、批判の伝播はその能力を維持することの一部であり、あらゆる依存する結果が自動的に知識でなくなるという規則ではない。第2章は、誰にも使われずに保存されているだけの記録 (休眠した記録) と生きた認識的能力とを、訂正が伝わるかどうかによっても区別した。休眠した記録も社会的知識であり続けうるとするバード (Bird) の見解 (Bird 2024) においてさえ、訂正が依存する仕事に届くかどうかは、その記録への依拠が根拠づけられたままであるかどうかにかかわると考えられる。欠陥を検出すること、記録上の論文を変えること、依存する仕事を修復することは、別々の達成である。ほかの達成も加えうる。たとえば、すでに結果にもとづいて行為した者たちに知らせることであるが、本節はこれを脇に置く。第2節で論じた材料の実験室の訂正は、記録上の論文を変えた。別の事例では、マサチューセッツ工科大学が、AI と科学の生産性についての査読前の論文のデータの出所、信頼性、妥当性にも、研究の真実性にも確信をもてないと公に述べ、その撤回を求めた。2026年の査読付き論文は、その後この論文を留保なしに引用した (MIT 2025; Soleymanibrojeni et al. 2026)。これは自律システムの失敗ではなく、伝播がどれほどの頻度で失敗するかを示すものでもない。この事例は、第三の達成が最初の二つから出てこないことの例になる。ひとつの仕事の内部では、依存するものをたどることが、検出にもなって行われることもある。1993年12月初め、アンドリュー・ワイルズ (Andrew Wiles) がフェルマーの最終定理についての自らの論証のなかのある計算がまだ完全ではないと発表したとき、印刷に回される直前だったワイルズの講義についての報告は、発表された結果のうち、フェルマーの最終定理を含むどれがその計算に依存し、どれが依存しないかを明記した (Rubin and Silverberg 1994, 2, 18)。

ここでも前提は形を変える。これまでの人間の研究の組織では、依存関係は部分的にしか追跡されない。引用はその一部を記録し、他を記録しない。各結果が何から導かれたかを記録する人工の構成の内側では、覆された主張に依存する内部の結果をたどことは、原理的に完全でありうる。その一つひとつが、覆された主張なしでもなお成り立つかどうかは、さらなる問いである。それに答えることはそれ自体が批判であり、それ固有の無効化の関係を要する。点検者に依存するもののなかには、それが出した認定も含まれる。点検者が後に信頼できない、あるいは点検したもののから独立していないと分かれば、それらの認定は点検者が与えていた支えを失う。人工の構成の記録の外でそれに依存するもの、すなわち *S* の他の者たちがその出力を使ったところは、記録が可能にする追跡の及ばないところにある。記録された範囲の外は点検できないという第5.3節の記録の境界が、ここでは訂正の届く範囲の境界として戻ってくる。訂正をそうした使用に届ける経路は、内部の点検の所見が依拠する者にとって無効化要因として働くために第5.2節が必要とした経路と同じ種類のものであり、ここでは結果が使われた後に要る。批判の受容というロンジーノの条件は、これに関連する要求を共同体の水準で述べる。共同体の信念は、その内部で行われている批判に応じて時間とともに変わらなければならない (Longino 2025, §5)。依存する結果を点検直すことにも費用がかかるなら、どれを先に点検直すかという順序の問題も生じる。第4節で見たバルサフスキーの論点を、検査の順序から訂正の順序にまで延ばすなら、誤りが見つかった後にも配分は残る。依存する結果のどれを先に直すかによって、同じ訂正の能力が、異なる人々を異なる期間だけ誤りにさらしうるからで

ある。そのため、たとえばそれぞれに依存する結果の数だけで順序を決めることも、価値を含まない選択ではなく、誰がどれだけの間誤りの費用を負うかについての帰納的リスクの重みづけ（第4節）を含むうる。

第12章第6.3節は、本節を本書における誤りの扱いのなかに位置づける。本節で見えてきたのは、ある時点で確かめられた独立性や認定が、後の時点でもそのまま成り立つとは限らない、ということである。本節の三つの論点に沿って言えば、時間がたつと崩れうるものは次の三つである。

一つは、批判者の独立性である。生産者と評価者が互いに適応していくと、評価者が合格を出しても、それが生産者の正しさを示す度合いはしだいに下がりうる。この低下は、点検がその標的に対して不完全であり、生産者を選んだり訓練したりするのに使われ、選抜によって生産者に残る誤りが点検の感度の低いところへ寄っていくところでは、結合の仕組みから生じうるもので、引いた研究はそれと整合するにとどまる。そのため、人工の構成の批判が独立であり続けると言うには、一回の評価で独立だったことに加えて、互いに適応していくなかで独立性がどう保たれるかを示す必要がある。考えられる手立ての一つは、生産者を選んだり訓練したりするのに使わない点検を、別に残しておくことである。訓練に使わない試験や、生産者が手を加えられない測定がその例である。選抜に使われる評価者は、使われるほど生産者について教えてくれなくなりうるが、選抜に使われない点検はその影響を受けにくい。

二つめは、新しいものを評価する基準である。クーンの説をそうした評価者に当てはめれば、過去の決定で訓練された評価者は、基準そのものが問われる場面で弱いかもしれない。しかも基準は結合とともに動きうるので、評価者の出力が S のいまの基準と合っていない、それだけでは基準を動かしていないことの証拠にならない。基準の動きそのものは、保存しておいた参照点や、 S の過去の判断から作った評価に照らせば、人工の構成の内側でも検出しうる。ただし参照点がこの働きをするのは、それが改められない限りにおいてであり、上で見たベンチマークの事例はその代償を示す。比較可能性のために保たれるものは、欠陥を抱えたまま保たれることもある。しかし、検出された動きを受け入れるかどうかは、それが S の基準であるかぎり S の決めることであり、それを決める力を S が保っているかが問われる（第6節、第12章）。また、そうして基準が動くなら、ヴィルホルトの対応の読み（第4節）は、重みづけが改められることによってではなく、依拠する者の期待が動くことによって満たされるようになりうる。依拠する者が異議を唱えることまで求める、より強い読みは、これに抗しうる。

三つめは、訂正である。ある主張が覆されても、それに頼っていたところに訂正が届かなければ、誤りは残る。知識を生きた能力に結びつける説では、知識を保つにはこの経路が要る。各結果が何から導かれたかを記録している人工の構成なら、内部では、覆された主張に依存する結果をたどれる。ただし、たどったものがなお成り立つかを確かめること自体が、新たな批判になる。そして、 S の他の者がその出力を使ったところまでは、記録は及ばない。そこへ訂正を届けるには、人工の構成の記録の外へ出る経路が要る。

まとめると、ある時点で人工の構成の内側で確かめられた独立性や認定が、その後も続くとは言えない。崩れ方は三つに分かれ、それぞれ別のものを要する。選抜への適応が点検の寄与を弱めうる人工の構成では、その寄与を保つ別の仕組みが要り、選抜に使わない点検はその候補の一つである。基準の動きについては、検出は内側でもできうるが、受け入れるかを決めるのは S であり、その力が S に保たれていることが要る。そして、出力が人工の構成の記録の外で使われるなら、訂正をそこへ届ける経路が要る。したがって、第9節の問いに対しては、時間を考えに入れるかぎり「すべてを内側で」という答えはとれないかもしれない。後の二つで要るのは人間であることではなく S との関係なので、このことは S が人間であっても人工であっても変わらない（ A と S が重なる場合に何が外にあたるかは、第12章第5節で扱う）。また、ある時点の評価で独立だったことを示すだけでは、人工の構成が自らの研究を評価できることを示したことにはならない。

9. S のための批判は人工の構成の内側で組織できるか

本節では、本章の問いを、ここまでの材料を使って考える。第9章第1節の例では、一つのシステムが百のインスタンスを生成し、互いの産出物を点検させ、その点検に応じて資源を振り向け直す。それを構築する者は、内部の点検を通った結果は S のために認定されており、そのため他者はそれに依拠する理由をもつ、と主張しうる。本節が問うのは、この主張がどの条件で成り立つか、つまり S のための批判と認定を人工の構成の内側で組織できるかである。機械が自ら知識を生み出し、自らそれを評価する研究の仕方は、 S にとっての認定を与えうるのか、という問いでもある。そして、この問いに説ごとに答えることは、そうした研究の仕方のもとで研究に何が起こるかを見定める手立てにもなる。何が内側に入りうるかが分かれば、そのとき何が S から見えなくなるかが分かり、何が残るかが分かれば、研究が S のための研究であり続けるために S の側に何が要るかが分かるからである。この答えは第12章にとっても要る。内側で組織できない部分があるなら、それが何であり、なぜ残るのかを先に知っておかなければ、それを誰が決め、誰が改めてよいかを問えないからである。

進め方は次のとおりである。まず、本章で見た説がそれぞれの問いに答えているかを整理し、説ごとの判定を一つの表にまとめる（第9.1節）。そのうえで、それらの説を並べて見えてくることを述べ（第9.2節）、そこから先へ延ばして示唆されることを述べる（第9.3節）。最後に、第12章へ渡る問いを挙げる（第9.4節）。先に見通しを示しておくと、行き着くところは説によって分かれる。誤りを見つける働きは、誤り検出の見方のもとでは、点検が生産者と違うところで誤るなら内側で担いうる。これに対して、 S にとっての受容可能性、依拠する者に答える誰か、異論を唱える資格などは、それを求める説のもとでは S との関係として残る。そして第8節で見たように、時間を考えに入れると、内側で担える働きについても「すべてを内側で」とは言えないかもしれない。第9.3節は、そこから研究に何が起こりうるかへ進む。内部の点検を通ったことが認定の目安になれば S との関係の欠けがそれとともに選ばれていきうること、独立した点検が使うほど減っていきうること、残る関係を担える者が S にいなくなりうること、そして A と S が重なるにつれて、認定の問いが「どの社会にとってか」へ移りうることである。なお、本節で「内側」と言うときは第9章第5節の意味であり、そこには二つのことが含まれる。点検の仕事が人工の構成の内側で供給されることと、その結果を S のための認定として認める承認が内側にあることである。本章の議論では両者は分かれうるのであり、 S との関係として残るのは主に後者である。

ただし、違う説を比べる以上、判定が分かれること自体は驚くにあたらない。本章の材料から見えてくるのは、人工の構成について判定が分かれる理由である。人間の研究では、説がそれぞれ追う性質が、人間という実現者と研究の組織によって互いに結びつけられていた。批判者は別々の人として形成されたので、その数が独立性をおおよそぞった（第3節）。認定は人の主張であり、誤れば答える誰かがいた（第5節）。産出者と依拠する者は、誤りの重みについての感覚を共有していると前提できた（第4節）。一致している専門家は別々の場所で訓練され別々の証拠に近づいていたので、一致の信号は多様性を

伴っていた（第6節）。査読の段階は、評価を S に見えるものにし、評価する力を育てる場でもあったと考えられる（第7節）。そのため多くの日常的な事例では、誤りの経路を問う説も、視点を問う説も、答える誰かを問う説も、同じ判定に傾きやすかったと考えられる。人工の構成は、これらの性質を一つずつ別に設定できるものにする。批判者は複製としても作り込んだ違いとしても作れ、記録は実行から残せ、判定の集め方は選べ、評価は産出に溶け込みうる。性質が離れることは、例外ではなくなる。インスタンスの共和国について表11.1の説が別々の判定を下し（第3節）、第5.2節の四つの答えが同じ人工の構成に別々のことを求めたのは、このためである。これは、人工の構成がかならず説を分かれさせるということではない。異なる基盤、依拠する者に届く記録、名指せる答え手、外に向けた層を備えて、性質を束ね直すこともできる。ここから、構築者の認定の主張は、どの説に答えるのかを示さないかぎり評価できない、ということが出てくる。ある説を満たす証拠が、別の説を満たすとは限らないからである（第3節、第5.4節）。

9.1 説が答える問いと、説ごとの判定

判定を比べるには、それぞれの説が何に答えているかを分けておくことが要る。本章の説は、三つの問いのどれかに答えている。一つは、批判が何をもたらすかである。誤り検出の見方は批判の価値をそれが加える誤りについての情報に置き、ロンジーノ、ソロモン、立脚点理論、機械的客観性の理想は、それぞれ別のところに置く（第1節～第3節、表11.1）。二つめは、点検していない結果への依拠を何が根拠づけるかであり、認定をどのような行為とみるか（第5.1節）、依拠の根拠は何か（第5.2節）、その依拠は信頼でなければならないか（第5.3節）がここに入る。三つめは、結果をある用途に足るとする閾値が誰の価値づけに照らして定まるかであり、ヴィルホルト、ダグラス、ジョン、エラッパールが答える（第4節）。ロンジーノの説は背景仮定を通じて第三の問いにも、ヴィルホルトの説は信頼を通じて第二の問いにもかかわる（第4節、第5.3節）。一つの立場は、三つの問いへの答えの組み合わせとしてとらえられる。誤り検出の見方と信頼性主義を合わせる立場は、第一と第二の問いの一つずつ答えたものであり、どちらも依拠する者自身の理由や S の視点を求めないので、内側に置けるものが大きくなる。ただし第三の問いには何も言わないので、閾値の価値づけについて何が残るかは、ヴィルホルトやジョンの説を併せるかどうかで変わる。ロンジーノの説と還元主義を合わせれば、視点と理由のどちらもが S の側に求められる。 S との関係として何が残るかは、組み合わせごとに違う。

表11.2は、これらの説のそれぞれについて、第2～8節の判定を並べる。厳密な要約ではなく、見渡すための単純化した見通しであり、各欄の根拠は括弧の節にある。表11.1と同じく、各行の判定は、論者たち自身が論じなかった人工の構成に説を当てはめて得たものである。第二列は、その説の結論が人間の研究について成り立っていたとき、どの人間の前提に頼っていたかを名指す。説そのものは人間を前提せず、目安だけが前提に頼っていた場合もあり、その場合はそう書いた。

表11.2 説ごとに見た、人間の前提、内側で担えるもの、 S との関係として残るもの（単純化した見通し）

説	依っていた人間の前提	抽象化して残るもの（内側で、どの条件で）	S との関係として残るもの	S の組成で変わること（ A と S の重なりを含む）	内部の点検が人間よりよいなら
誤り検出の見方（第2節）	説そのものにはない。数と組織の分離という目安は、批判者が別々の人として形成されることに依った（第3節）	主張ごとに情報を加える点検（内側で可）。条件：加える情報と失敗の診断を示し、互いに適応しても独立性が保たれること（第8節）	この見方だけでは少ない。ロンジーノの説を併せると、何に不利に働くかが背景仮定に相対的（第1節）	効くのは人間か人工かではなく誤りの経路の共有。基盤を共有する成員の点検は同じところで誤りやすい	そのまま効く。ただし、よいことの証拠が要る（次の行）
信頼性主義（誤り検出の見方と組み合わせる場合）（第5.2節）	ない（連鎖の終わりに人間を求めない）	信頼できる過程。正当化の条件は過程が実際に信頼できることであり、そうであることを示すには、点検者の判定で点検者を測る循環（ブートストラッピング）によらずに信頼性を確かめることが要る	受け手が範囲の内で受け取ること、正当化から用途への移り（第4節の価値づけ）、費用を負う外の者	A と S が重なると、点検とその評価が同じ過程になり循環に陥りやすい	正当化に効く。ただし人間の決定との一致が測るのは実践への忠実さであり（第7節）、確かめられる範囲の先は校正の引き継ぎに依拠する（第6節）
ロンジーノ（第3節、第9.1節）	人間の研究共同体における、状況に位置づけられた知者どうしの相互作用	四つの条件（内側でも満たしうる）。受容可能性は利用できる視点に相対的であり、条件には共同体の外からの批判に開かれていることが含まれる	S の視点が批判に入るか、データが受け入れる共同体にとって明らかか、共同体の外からの批判に開かれているか	来歴を共有する S では、隔たりが批判のよくなるまま縮まる。共同体が S と重なるところでは、外への開かれは S の外を指す	すでに行える批判を強めるが、欠けた視点からの批判を加えない
ロンジーノの説のワイリーによる拡張（第3節）	異なる目的、実践、信念をもつ共同体どうしの協働 (Wylie 2015, 13, author manuscript)	内部の多様性だけでは外部とのかわりを確立しない	研究共同体の外との協働であり、 S の外とは限らない	研究共同体と S が一致する場合、関連する外部の協働相手は S の外にいることになる	内部の点検が優れていても、それだけでは外部との協働を供給しない
ソロモン（第3節）	判断に働く影響が人ごとに異なっており、分布すること、その分布をよくする見えざる手はなく、分布は個人の推論でなく共同体とその制度の水準で評価され変えられること（第3節）	研究の労力の分布を評価し左右する位置（内側に置きうる）。条件：占め手が問題の影響を共有しないこと	共有された影響の特定の一部を、それを共有しない当事者に置く	効くのは影響の共有。異なる作りの成員なら内側でも担いうる。多数決では複製が証拠を増やさずに重みを増す	個々の推論が正確でも足りない。評価されるのは影響の分布
立脚点理論（第3節、第10章第6節）	周縁化と出発点の違いが、人間社会の権力関係と生のなかで生じること	周縁化された人々の生との関わりと集団的な意識。内側で担えるかは、誰から学び誰とともに探究するかによる	その関わりの手先との関係	混合社会では人間と人工の成員の関係そのものが権力関係でありうる。人間の記録で訓練された人工の社会は、人間の研究共同体が立	求めるのは誤りの少なさではなく出発点の違いなので、精度は供給しない

説	依っていた人間の前提	抽象化して残るもの（内側で、どの条件で）	<i>S</i> との関係として残るもの	<i>S</i> の組成で変わること（ <i>A</i> と <i>S</i> の重なりを含む）	内部の点検が人間よりよいなら
				っていたところから出発する	
機械的客観性の理想（第3節）	判断する個人への不信と距離、個人の特異さや利害が結果に入ること	規則に従う手続き（内側で満たしうる）	手続きを設計した者への問い。手続きは、決めているように見えずに決めることを可能にする	人工の構成が手続きを自ら設計するなら、問いは誰がその設計を点検できるかへ移る	より一様なら、この理想ではより客観的とみなされうるが、共有された前提には沈黙する
コスキネンとディクソンが紹介するゴールドバーグの説（第5.1節）	道具への依拠が他者の認識的な仕事に依存し、社会的実践が使用者に設計者や製造者への期待をもつ資格を与えること	道具の示す値としての認定	設計者や作り手が外にいれば、その者たちへの依存。ただし、その者たちへの信頼が本質的に不透明な AI への依拠を支えるかは争われている	人工の構成が点検者を自ら設計するなら、設計した部分に、依拠を支えるために要る理解と責任があるか（当てはめ）	記録だけでは誰が答えられるかは確立されない
還元主義（第5.2節）	人間に固有の前提はない	実績の記録（内側で作れる）	理由が依拠する者自身に届き、その者が評価できること	依拠する者が基盤を共有すれば、系譜の評価が人工の構成の能力と相関しうる（第5.5節）	よいことを示す記録が依拠する者に届くかぎりで効く
非還元主義（最小の形）（第5.2節）	説そのものにはない。人間の研究では、所見が公開や訂正を通じて依拠する者に届きえたと考えられる（組織の前提）	打ち消されていない無効化要因がないこと。ただし経路がなければ、誤りの有無にかかわらず満たされる	所見が依拠する者に届く経路	依拠する者が人工の構成の内側にいれば、経路も内側に入りうる	経路がなければ、依拠する者の正当化に効かない
保証説（第5.2節、第5.4節）	話し手が意図して保証を与え、主張の後ろに立つこと。苦情の権利を加える説では、誤りに責任を負うことも要る	意図して与える保証（人工の構成かその成員が与えうるなら内側でも可）。苦情の権利を加える説では、誤りに責任を負うことも要る（基盤の提供や配備は保証ではない）	依拠する者に向けた保証。苦情の権利を加える説では、その者に責任を負う誰か	重なれば、インスタンスが保証を与え、求められる場合には誤りに責任を負いうるかぎりで、内側で満たす余地がある	精度だけでは保証は与えられず、苦情の権利を加える説での責任も満たされない
必要な信頼の説（第5.3節）	依拠を支える、責任ある設計者、維持する者、研究者などへの信頼	採る説が求める責任を負いうる当事者（人工の構成かその成員がそうしうるなら内側でも可）	そうした当事者への信頼。コスキネンはシステムの働きへの責任を求め、オルトマンは信頼性を確立し伝える研究者への信頼を求める	重なれば信頼の相手を内側に置きうる。求められる責任は採る説によって異なり、本質的な不透明さは主体に相対的である	精度だけでは、誰を何について信頼するかも、求められる責任が可能かどうかとも決まらない
ヴィルホルト：対応の読みと強い読み（第4節、第5.3節）	産出者と依拠する者が価値づけの感覚を共有すること、規約の残す選択の一部が暗黙で記録しきれないこと	損失と判断の理由の記録。依拠する者の期待が別に確かめられれば、対応を点検する材料になる	対応。強い読みでは、依拠する者が異議を唱えうる関係	機械が産出し人間が依拠するなら、AI アラインメントと構造の似た問い。来歴を共有する依拠者なら共有を前提しやすいが、費用を負う者は外に残りうる	精度は、誤りによる損失を誰についてどれほど重くみるかを決めない
ダグラス（第4節）	閾値を定めるのが、誤りの帰結に道徳的責任を負う行為者（科学者）であること	証拠と誤りの費用の考量を分け、考量を明示すること	どの責任ある行為者が考量について答えるか（第12章第6.3節）	人工の構成のどこに責任ある行為者として答えるものがあるかで、線が引かれる	答える者を決めない
ジョン／エラッパール（第4節）	ジョン：公開が特定できない聞き手に向けられること。エラッパール：証拠が人々のあいだに不均等に分布すること	特定できない当事者が用いる記録に結果が入るところでは固定された高い基準を内側で一様に当てはめ、特定できる聞き手に向けてところではその聞き手の基準に合わせる（ジョン）	証拠の乏しい当事者の基本的な利害（エラッパール）、依拠しないまま費用を負う者	<i>S</i> が人工なら、費用は人間や環境に及ぶうる（第4節）	精度は誰の損失を重くみるかを決めない

フリッカー、ラッキー、ウィリアムソンの、認定をどのような行為とみるかについての説は、第5.1節と第5.5節にまとめた。

ロンジーノの四つの条件を、一つずつ人工の構成に当てはめる。ドゥティル・ノヴァエスとデダが伝えるところでは、これらは、現実の共同体が程度の差をもって満たす理想化された条件である（Dutilh Novaes and Dede 2025, §2）。場について、人工の構成の内部の経路は、その構成要素にとって公であっても、それだけで、構成の外にいる *S* の成員にとって公になるわけではない。公である程度にも段階があり、ある大規模言語モデルの報告書は、競争と安全を理由に訓練方法などを開示せず、自らが選んだ50人を超える外部の専門家に初期のバージョンを試験させ、追加の第三者に詳細を提供する計画を告げた（OpenAI 2023, 2, 12, 45）。その第三者が何を受け取るかは、開発者の説明を通じてしか知られない。批判の受容が求めるのは、共同体のなかでの信念の分布が、その内部でなされる批判に応じて時間とともに変わることであり（Longino 2025, §5）、出力が批判者の報いる方向へ動くことではない。第8節は、後者が批判者の独立性を損ないうることを見た。公的な基準について、人工の構成は基準を明示できるが、ロンジーノは基準を、暫定的で、共同体の探究の目的に従属するものとして扱う（Longino 2017, 51–52, 54）。そのため、基準を誰が、どの目的に照らして改めるかは、「誰の未知か」の問い（第10

章第6節)と権威の問い(第12章)に戻る。ドゥティル・ノヴァエスとデデが伝えるところでは、調整された平等は、社会的地位や経済力が、どの視点が真剣に受け取られるかを決めないことを求める (Dutilh Novaes and Dede 2025, §2)。内部の批判者の重みを誰かが上から定めるなら、社会的地位がどの視点を重く受け取るかを定めることを防ぐという、この条件の働きは失われうる。重みを定める者の判断が、地位の代わりをするからである。重みが批判者どうしの応答と評決の積み重ねで決まる場合にも、ある構成要素の判定が他より重く扱われるという、社会的地位に似た働きが生じうる。重みを定めた形づくったりする者は、人工の構成を所有する者と一致するとは限らない(第9章第5節、第12章第2節)。ハーディングは、資格ある批判者とみなされる者の境界そのものが、共有された前提を守りうることを付け加える (Harding 2015, 6)。第3節で見たように、ロンジーノの定義は、四つの条件と、データがその時点で受け入れる共同体にとって明らかであることを求める。 S をその共同体に重ねるなら、データは S にとって明らかである必要がある。重ねないなら、 S の視点がその共同体の批判に入るかが、さらに問われる。ワイリーの拡張は、研究共同体の外との積極的な協働を加える。ここに当てはめると、人工の構成がその共同体をなす場合には構成の外を指すが、 S の外を指すのは、その共同体が S と一致し、関連する外部の協働相手がいる場合に限られる。

9.2 説を通して見えてくこと

第一に、内部の批判をよくすることは、いくつかの説のもとでは、 S のための認定を強めず、ときにかえて弱めうる。一見すると、内部の点検が多く、よく働くほど、認定もよくなるように思われる。誤り検出の見方と信頼性主義を合わせるなら、そのとおりである。しかし三つの仕組みが、それぞれ別にこの見込みを崩す。一つは、ロンジーノの説で、全員が共有する仮定が、特異な仮定をふるい落とす構造を生き延びるので、内部の批判がよく働くほど定着しうることである(第3節)。これは認定を弱める向きに働く。二つめは、非還元主義の最小の形で、所見が依拠する者に届く経路がなければ、無効化要因がないという条件が、結果に誤りがあってもなくても満たされることである。そのため、内部の点検をどれほど強めても、依拠する者の正当化は変わらない(第5.2節)。こちらは認定を弱めはしないが、強めもしない。三つめは、評価が S から見えなくなることである。内部の批判をよくする手立てには、点検を候補の形成の途中に絶えず入れ、産出と一体にすることが含まれる(第7節)。ロンジーノの説、保証説、還元主義のように、 S に向けて何かが明示的に示されることを求める説のもとでは、そうした手立てが進むほど、評価が S に見え、異議の対象になり、 S の成員が評価する力を保つ機会が減る(第7節)。三つの仕組みは別々なので、手当ても別々になる。第一には人工の構成とそれが共有するものの外から来る違いが、第二には所見の届く経路が、第三には S の成員が評価に加わる機会が要る。一步進めると、精度は劣るが S に開かれた点検が、精度は優れるが S に閉じた点検よりも、 S の依拠の根拠には多くを与えることがありうる。還元主義では理由が届くこと、無効化要因への感応を求める非還元主義では経路、ロンジーノの説では S の視点からの批判、保証説では後ろに立つ誰かが、点検の精度とは別に問われるからである。これは真理への近さではなく依拠の根拠についての比較であり、純粋な信頼性主義のもとでは成り立たない。

第二に、作り込んだ違いは、違いを選ぶ過程を共通の源としてもつ。批判者が一つの基盤の複製であることが問題なら、違う批判者を作ればよい、と考えたくなる。ところが、ゲレロが報告する、共通の第三者の見解を映すだけの源についてのゴールドマンの注意(第2節)を違いの選び方にまで及ぼすと、どのような違いを作るかを一つの過程が選ぶとき、その選択そのものが、作られた批判者すべてに共通する源になりうる(第3節)。そのため独立性の問いは、批判者から、その違いを選んだ過程へと一段上がる。第10章第4節は、配分について、これに対応するところに行き着いていた。異なる路線を生成できる人工の構成も、路線を残すかどうかを一つの基準で判断するなら、生成した路線を同じ誤りによってまとめて捨てうる。この問いが止まるのは、違いの選び方が、選ぶ過程自身の予期しない違いにも開かれているところ、たとえば選択が別の当事者に分けられているか、誰も定めない過程を通じて違いが形づくられるところである。これらは、働きが人工の構成の内側に入ったときにそれが担われうる仕方として第10章第5節が見たもののうちの二つであり、そこで見たことを伴う。誰かが定めるパラメータなら、それを定める者が、人工の構成が学びうるものの一部を決める。誰も方針として定めない過程なら、ただ一人の作り手はいないかもしれず、それについて誰が答えるかは第12章第6.3節の問いになる。ハーディングの要求を当てはめれば、基盤システムを形づくった評価者たちとその仮定も、結果の説明の対象に入る(第3節)。価値づけの違いについても同じ形が現れる。イルズィクとクルトゥルムシュは研究者の多様性を、研究が公衆の一部ではなく全体の期待を映すと考える理由として用いる。これを当てはめると、価値づけの幅が成員の生み出され方によって定まる人工の構成では、その幅は成員の生からではなく、成員を生み出す過程の設定から来るので、多様性がその理由として働くかどうかは、 S の期待を持ち込むハイブリッド・フォーラムのような経路があるかに依存する(第6節)。誤りの経路の違いでも価値づけの違いでも、違いが作られるところでは、問いは、違いを選ぶ過程と、その選択を点検する経路へ移る。誰が成員を作り視点の分布を定めてよいか、権威の問い(第12章第5節)であるだけでなく、認識の問いでもあるのはそのためである。

第三に、 S が人工の構成と来歴を共有するとき、「 S のための認定」は、よくなるまま満たしやすくなる。同じ種類の人工の社会が同じ種類の結果を認定することに、ふつうは問題はないように見える。だが三つの論点を合わせると、構築者は、結果がそのために認定される社会として、自らの人工の構成と来歴を共有する人工の社会を名指すことによって、主張を満たしやすくなる。ロンジーノの説では、 S の成員がインスタンスと原初的な組織を共有するなら、人工の構成に相対的な受容可能性と S に相対的な受容可能性との隔たりは、批判がよかったわけではないのに縮まる(第3節)。ヴィルホルトの説で依拠を安定させていた価値づけの感覚の共有も前提しやすくなり(第4節、第5.5節)、依拠する S の成員が人工の構成の内側にも立つなら、無効化要因がそこへ届く経路も人工の構成の内側に入る(第5.2節)。こうして問いの中身は、認定がよいかどうかから、 S が誰であるかへ移る。何が残るかは説によって違う。誤り検出の見方と信頼性主義を合わせるなら、 S のための認定は内部の誤りの訂正にほぼ吸収される。残るのは、それ自身の誤りの経路をもつ点検と無効化の関係という探究の条件(第12章第5節)、依拠する者が人工の構成の外に立つ場合には、その者が認定を受け取る仕方の信頼性(第5.2節)、そして、依拠しないまま誤りの費用を負う、 S の外の者との関係である(第4節)。ロンジーノの説では、共有された仮定はなお批判を免れうる。ワイリーの拡張は、関連する外部の共同体とかかわるさらなる理由を与える。研究共同体が S と一致する場合、それらの相手は S の外にいることになる(第3節、第9.1節)。保証説では、 S の外の者が依拠する限り、その者に対して認定の後ろに立つ誰かが要る。

第四に、独立性は、誰も欺かなくとも、選抜に使われるだけで衰える。自然な見込みは、一度独立だった点検は、誰かが欺くか規則を書き換えないかぎり独立のまま、というものである。しかしマンハイムとガラブラントの分類では、不完全な代理指標で選抜すると代理指標と目標の差も選抜される回帰的な場合と、選抜によって関係の成り立たない領域へ押し出される極値的な場合には、尺度に応答する者がいなくても、選抜だけで代理指標と目標が離れ

る（第8節）。これを研究に移すと、点検がその標的に対して不完全であり、生産者がその判定で訓練されたり選ばれたりするところでは、点検の感度の低いところで通る出力が好まれていきうる（第8節）。そうなるところでは、選ばれた生産者がいま犯す誤りを点検が露わにする確率、すなわちその生産者に対する点検の厳しさが下がり、それとともに、点検が主張ごとに加える誤り検出の情報も減る。第2節でゴールドマンの条件を拡張した見方では、独立性とはこの情報のことなので、これは独立性の喪失にあたる。これは、生産者の犯す誤りが減ったために点検の見つかるものが減ることとは別であり、後者では、点検の合格が示すものは減らない。そのため、選抜に使われる評価者は、使われるほど生産者について教えなくなりうる（第8節）。さらに進めると、独立性は、点検がある時点でもつ性質というより、その点検がどう使われるかに相対的に保たれる性質である。規則を書き換えから守ることは寄与の維持を助けても、寄与を確立も保証もしない（第1節、第8節）。保つ手立てになりうるのは、書き換えを禁じることも、選抜に使わないでおくことである。

第五に、*S* のいまの基準との一致は、基準が動いていないことの証拠にならず、そのためヴィルホルトの対応の読みは、重みづけが改められないまま満たされうる。人工の構成の価値づけが *S* に合っているかは、*S* のいまの基準と一致するかで確かめればよい、とまず考えられる。けれども、ラッセル、ラッティ、コメリンらの議論を当てはめれば、結合とともに動く基準は *S* 自身の基準でありうる（第8節）。これを第4節の二つの読みに重ねると、対応の読み、すなわち重みづけが依拠する者の期待する価値づけに一致することは、重みづけを改めることによってだけでなく、依拠する者の期待が動くことによっても満たされうる。より強い読み、すなわち依拠する者が異議を唱えうることは、これを防ぎうる。ただし防ぐのは、異議を唱える力そのものが人工の構成によって形づくられていないかぎりにおいてであり、その力は人工の構成の側が何も変わらなくても *S* の側から衰えうる（第6節）。ここで、検出と決定を分けておく。以前の基準に照らして動きを検出することは、保存した参照点や *S* の過去の判断から作った評価によって、人工の構成の内側でもできうる。しかし、その動きを改善とみるか置き換えとみるか、受け入れるかは、それが *S* の基準であるかぎり *S* の決めることであり（第12章第2節・第5節）、内側の検出はそれに代わらない。

第六に、*S* との関係として残るとされたものは、*S* の側にそれを担う力があることを前提している。表11.2の第四列の関係は、どれも *S* のなにかしかなることを要する。無効化要因を読むこと、*S* の視点から批判すること、点検者の系譜を評価すること、説明を理解することなどである。系譜の評価には別種の技術的な専門性が要することがあり、点検する力は、成員が経験を積まなくなることでも *S* の側から衰えうる（第6節）。明示的な評価の段階がなくなれば、その力を保つ場も減る（第7節）。他方で、ハードウィグ、ゴールドマン、アンダーソンの説では、誰が人工の構成を方向づけているか、それはどの目的に資するか、批判に答えているかといった関係についての問いの一部は、内容についての有能さなしに問える（第6節）。そのため、本章の見通しのうち「*S* との関係として残る」という部分は、技術的な判断を要する関係については、その判断を担う力が *S* に保たれていることを条件とする。第6節が当事者資格について述べた「名目以上」という条件は、残る関係のそれぞれに及ぶ。

第七に、誤りについて同じ情報を加える批判者どうしても、探究にとつての価値は同じとは限らない。誤り検出の見方では、二つの批判者の配置が誤りについて同じだけの情報を加えるなら、両者の価値は同じになる。しかしクーンの説を当てはめると、そこでは終わらない。判断の多様性は、新しい誤りの経路を何も加えなくても、新しさの導入や支持に伴うリスクを集団のなかに分散させるという価値をもちうるからである（第3節）。そうだとすると、少数派の判断を残しておく配置と、多数決でそれを消してしまう配置とは、同じ情報を加えていても、新しい理論がまだ支持の乏しいうちに展開され試される余地を、同じだけ残すとは限らない。ここから、認定の精度を上げることと、新しい理論が育つ余地を保つこととは、分かれうる。たとえば、人工の批判者の判断を描えることや、判定を一つの多数決にまとめることは、前者を改善しても、後者を弱めるかもしれない。ただし、これはクーンの説を採ったうえで条件付きの結論であり、作られたばらつきが、クーンの論証が頼る個人差の広がりと同じ働きをするかどうかは未決のまま残る（第3節、第8節）。実際の人工の批判者の集まりでそうした損失が起こるかどうかも、経験的な問いとして残る。本章は、ここで検討した配置が新しいものの余地を狭めうる道筋に、ほかに二つ出会ってきた。一つは、まだ形づくられている途中の候補に批判を当てると、まだ支持の乏しいものが退けられうることである（第7節）。もう一つは、クーンの説では、過去の決定で訓練された評価者が、受け入れるには基準を改めなければならない候補を、逸脱として採点しうることである（第8節）。三つの道筋がかかわるのは、批判をいつ当てるか、評価者の基準がどう形づくられるか、判定をどう集めるかという別々の事柄なので、一つの点で新しいものの余地を保っても、ほかの点で保たれるとは限らない。

第9章第6節の仕分けを批判と認定に当てはめると、線は次のように引かれる。人間という実現者についての前提は崩れうる。批判者が別々の人として形成されるので数が独立性をなぞること（第3節）、規約の背後の判断の一部が暗黙で完全には記録できないこと（第4節）、点検する者の時間と注意が限られていること（第7節）である。研究の組織についての前提は、実現者が人間のままでも崩れうる。データと過程の記録が研究者の手元にあること（第5.3節）、出版前の査読が乏しい編集上の注意と紙面を割り当てる関門であり、その通過が配分の信号になること（第7節）、依存関係が引用によって部分的にしか追跡されないこと（第8節）である。産出者と依拠する者が価値づけの感覚を共有することは、人間という実現者についての前提と研究の組織についての前提とが混じったものとして読める（第4節）。それらが崩れても、対処していた問題は形を変えて残る。独立性は作られるものになり（第3節）、研究者の報告への依拠は記録の維持、境界、その内側で検証する能力への依拠になり（第5.3節）、点検に費用のかかるところでは関門が残り（第7節）、依存する結果をたどれても、それがなお成り立つかを確かめることは新たな批判になる（第8節）。探究の条件は残るが、内側で満たされうる。情報を加える無効化の関係と裏づけのある失敗の診断（第1節～第2節）、代理指標が避けられないこと（第9章第4節）などである。依拠する者が証言者の理由をもたないという証言の構造も、何が証言するにせよ成り立つ（第5.3節）。ただし、それが内側にあるか外側にあるかは、価値づけをめぐる関係（第5.5節）と同じく、依拠する者がどこに立つかによる。*S* との関係のうち、相手の参加や承認を含むものは、片側からは完成させられない。そこに何が入るかは、表11.2の第四列のとおり説によって違う。誤り検出の見方と信頼性主義を合わせるなら、認定が依るものの多くは探究の条件に属し（第9章第6節）、他の説のもとでは、より多くが *S* との関係に属する。本節の初めに述べた二つの「内側」のうち、*S* との関係として残るのは、ここでも主に承認の側である。

第7節で見た帰結も、この仕分けから見直せる。人間の査読では、名指せる誰かが出版の前に原稿を読むという一つの行為が、誤りの検出、評価が共同体に見えること、異議を向けられるその人の判断、答える誰か（編集者や査読者）、*S* の成員が評価の経験を積み力を保つ機会を、同時に生んでいたと考えられる（第7節）。人工の構成は、このうち誤りの検出だけを、ほかのものなしに生むことができる。第7節で、明示的な評価の段階がなくなるときにいくつもの帰結がいっしょに出てきたのは、それらが同じ機会から生まれていたからだと考えられる。どれが重要かは説によって違う。誤り検出の見方では第一だけが、ロンジーノの説では評価が見えることと場が、保証説では結果を請け合う誰か（苦情の権利を加える論者では、誤っていれば答える者でもある誰

か) が、必要な信頼の説では答える誰かが、ハードウィグとアンダーソンの説では S が関係について問う力が、修正可能性を問う第12章第6節では異議の対象になる痕跡が、それぞれ重要になる。さらに、それぞれが何に結びついていたかも違う。見えることと異議の対象は評価の段階に結びついていたので、記録、公開の場、異議を受け付ける経路のような、第7節で見た外に出す層で作り直しうる。答える誰かは段階ではなく人に結びついていたので、そうした層を足しても生まれず(数量化は決めているように見えずに決めるやり方だという第3節のポーターの指摘も、同じ方向を指す)、保証説では意図して保証を与えうる誰かが要る。評価する力は、 S の成員が実際に評価することで保たれてきたので、層が結果を返すだけでは戻らず、成員が評価に加わる機会が要る。そのため、束をほどく人工の構成について問うことになるのは、どれを、何によって束ね直すか、そして束ね直したものが頼る中身(経路が受け付ける異議など)が構成の外から来るかである(第12章第9節)。

第9章第2.3節の四つの財のうち本章にかかわる三つについて言えば、次のようになる。境界をまたぐ誤りの検出は、誤り検出の見方では、組織の境界ではなく、点検がそれ自身の誤りの経路をもつかどうかにかかわるので、人工の構成の内側で供給されうる(第2節)。ただし時間を通じて保つには、選抜に使わない点検など、点検の寄与を保つ手立てが要る(第8節)。ロンジーノの説では、共有された背景仮定は内部の批判を免れうる(第3節)。異論の当事者資格は、異論の認識上の働きとは異なり、人工の構成の内側で保持されるとすれば、採る説のもとでそれをもつ部分によってだけだろう(第3節)。説明が求められた境界、すなわちある当事者が自らの行ったことを別の当事者に説明し、あるいは正当化しなければならなかった境界については、記録があれば説明を与えることはできるが、説明を求めてよい資格は記録からは生まれない。その資格がどこから来るかも説によって違う。苦情の権利を加える論者の保証説を当てはめれば、依拠する者は誤りについて答えを求めうる立場をもち、ヴィルホルトの説の強い読み(第4節、第12章第1.4節)では、依拠する者は重みづけに異議を唱えうる。これに対して純粋な信頼性主義では、依拠する者の正当化に説明は要らない。そのため、この境界が人工の構成のもとで残るかどうかは、採る説によって分かれる。誰がその資格をもつかは、第四の財である責任を負う当事者とともに、第12章第6節で検討する。

9.3 そこから示唆されること

以下は、ここまでの議論を、論証の届く範囲の先へ延ばしたものである。一つは、社会的知識と証言についての説のどれを採るかが、実践の上で効いてくるかもしれないことである。人工の構成がそれぞれの性質を別々に設定できるなら、構築者がどの性質を備えさせるかは、どの説の要求に応えるかを選ぶことにあたる。一般化すると、人工の構成を設計することは、第2章第6節が決めずにおいた選択を、誰も明示しないまま設計のなかで行うことでありうる。示唆されるのは、その選択を誰が、どの経路で行うかが、第12章の権威の問いに加わるかもしれないということである。

二つめは、 A と S が重なるとき、「 S のための認定」が指すものに移るかもしれないことである。第4節は、結果を点検し直さずに用いる者がいるかぎり閾値の働きが要るとした。この論理を延ばすと、 A と S が重なっても、成員が互いの結果を点検し直さずに用いるかぎり、閾値の働きは人工の構成の内部の働きとして残る。そうだとすると、誤り検出の見方と信頼性主義を合わせる立場では、認定は内部の訂正と内部の閾値の設定にほぼ吸収され、他者が依拠する理由を与えるという意味での認定は、その結果に依拠する別の社会に相対的なものになりうる。ロンジーノの説では、共有された仮定は検討されないままでありうる。しかも受容可能性の条件には共同体の外からの批判に開かれていることが含まれるので(第3節)、人工の構成そのものが共同体であるところでは、この吸収は完全にはならない。ワイリーの拡張では、研究共同体の外との協働がさらなる条件として残る。保証説では、別の社会に向けて後ろに立つ誰かが要る。重なりが大きいほど、認定の問いは「 S にとってどうか」から「どの社会にとってか」へ移ると考えられる。

三つめは、内部の点検を通ったことが認定の代理指標として使われるなら、第8節の論理が認定そのものに及ぶかもしれないことである。内部の点検の強さや通過率は測りやすく、表11.2の第四列にある S との関係の多くは測りにくい。仮に構築者や利用者が前者を目安に人工の構成を選び、改めていくとし、しかも内部の点検の高い得点が S との関係を欠いたままでも得られるために、関係の欠けるところで得点が認定を過大に示すとする。すると、マンハイムとグラブラントの回帰的な場合と同じく、代理指標と目標の差、すなわち S との関係の欠けも、それとともに選ばれていきうる。そうだとすれば、第9.2節の第一の結論は、ある時点の比較から、時間を通じた傾向へと強まる。

四つめは、残る関係を担える者がいなくなるかもしれないことである。第9.2節の第六の結論を延ばすと、人工の構成が S の成員より有能になること、評価が内側に溶け込むこと、 S の成員が評価の経験を積まなくなることの三つが重なるとき、技術的な判断を要する関係は、形の上では S との関係として残りながら、それを担える成員がいない状態になりうる。コスキネンは本質的な不透明さが主体に相対的であるとした(第5.3節)。これを当てはめれば、責任を負える者がいないという空白は、システムの作りからだけでなく、 S の側の能力の差からも生じうる。ただし、関係についての問いの一部は内容の有能さなしに問えるので、この空白はすべての関係には及ばない。示唆されるのは、どの関係が技術的な判断を要し、どの関係が要しないかを分けることが、 S が何を保てば足りるかを知る手がかりになる、ということである。

五つめは、独立した点検が、使うと減る資源のように振る舞うかもしれないことである。第9.2節の第四の結論では、選抜に使わない点検は、選抜に使われる評価者の衰えを受けにくい。しかし、そうした点検も、認定のために繰り返し参照され、その結果がどの生産者を残すかに反映されるなら、参照されるたびに選抜の一部になっていく。工学の言葉で言い直せば、訓練に使わないはずの試験が、何度も参照されて選択に使われるうちに、選抜に使われる尺度と同じ立場に移るということである。仮にそうだとすると、独立性は一度確保すれば済むものではなく、使えば減り、新しい測定や異なる作りの評価者によって補われなければ尽きていくものになる。さらに一歩進めると、ここには批判から学ぶこととの緊張があるかもしれない。有効な批判は生産者を変える。批判を受けて生産者が誤りを直せば、変わった生産者について、同じ点検が見つかるものは減りうる。それだけでは点検は弱まっていない。その点検が見つかる誤りが少なくなったために、見つかるものが減ったのかもしれないからである。点検が弱まるのは、変わるなかで生産者が、点検の感度の低いところで通るようにもなった場合である(第8節)。そして、見つかる誤りが減ったことだけからは、どちらが起きたのかは分からない。また、点検を選抜に使わないでいくという手立て(第9.2節の第四の結論)は、それだけでは、批判の結果を生産者にどう取り込むかという問題を残す。取り込めば点検は選抜や訓練の一部に近づきうるし、取り込まなければ批判は生産者をよくしない。そうだとすれば、独立性は、保存しておく部品の性質というより、批判が効いたあとにそのつど評価し直される関係でもあることになる。ただし、新しい点検をどれだけ補い続けることが要るかは、対象の領域と、点検が試す誤りの範囲によるだろう。誤った証明を通さない検査器は、生産者がそれに合わせて学んでも、形式的な言明についての効力を失わない(第8節)。補うことには費用がかかり、その資源は有限なので、時間を通じて認定を保つことは、独立した点検をどの結果のためにどれだけ使い、どれだけ補うかという配分の問いになる(第10章第1節)。そして、それを誰が決めてよいかは、第12章の決定の問いになる。

9.4 第12章へ渡る問いと、理解についての限定

本節の見通しのうち、認識の関係についての説だけでは決まらない部分は、権威の問いとして第12章へ渡る。少なくとも次のものである。一つは、どの働きについて暗黙で足りるとし、どの働きについて S に向けて明示的に示すことを求めるかを、誰が決めてよいかである（第12章第6節）。二つめは、誰が S を構成し、誰がその代わりに語ってよいかであり、 A と S が重なるとき何が「外」にあたるかもこれに含まれる（第12章第1.3節・第2.1節・第3.3節・第3.6節・第5節）。エイギによれば、ロンジーノは、共同体が、ある種の知識を生み出す者たちからなるのか、それによって影響を受けうるすべての者たちを含むのかを決めていない (Eigi 2016, 136)。三つめは基準と誤りの重みを誰が定め改めてよいか、誰の費用が問題になるか（第12章第1.3節・第1.4節・第3.5節・第6.3節）、四つめは誰が成員を作り視点の分布を定めてよいか（第12章第5節）、五つめは誰が説明を求め異議を唱える資格をもつか（第12章第1.4節・第3.3節・第3.4節・第6.2節）である。第9.3節の五つめの示唆が成り立つなら、独立した点検をどこに使い、いつ補うかを誰が決めてよいかも加わる。

最後に、本節の見通しは、求められる達成が知識である場合のものである。理解に向けられた探究では、説明の適切さの認定、理解の帰属の認定、受け手の実際の理解を区別することになり、内部の批判は、第三を実現せずに最初の一つにかかわりうる（第5.6節）。誰の理解が求められているかは、認定の手続きがどこで実行されるかではなく、構築の目標による。 S の成員が人工であり表象を直接に共有できるなら、獲得は証言を通らないこともあるが、そのとき受け手が理解したかは、採る理解の説によって分かれる（第5.6節）。

第12章 誰が決め、誰が改めてよいのか

当事者資格、修正可能性、外の当事者との関係

1. 権威は人工の構成の内側に置けるか

1.1 問いと、前二章から渡されたもの

第V部が問うのは、研究を取り巻く関係のうちどれが人工の構成の内側へ移りうるかである。第10章と第11章は、探究の配分と、批判と認定（結果が誤りを見つける吟味にさらされ、他者がそれに依拠する理由を与えられる関係）についてこの問いを問うた。本章は残る一つの関係を扱う。第9章第2.1節が権威と呼んだもの、すなわち配分の条件と、批判と認定の条件とを決め、改める当事者資格（決定に服する者との関係で、提案し、決め、異議を唱え、拒み、改める資格）である。人工の構成が配分と批判を内側で担えるようになって、その条件を誰が決め、誰が改めてよいかは、それがどれだけよく働くかとは別の問いとして残る。その誰かは、人間かもしれないし、人工の構成それ自体かもしれないし、両者のあいだで分けられるかもしれない。人間であるとしても、 S の成員か、人工の構成を作り提供する者か、その帰結を負う者かで答えは違い、人工の構成それ自体であるとしても、そのどの部分が決めるのかが問われる。そして、よく働くことは、それだけではそれが決めてよいことを示さない。本章の中心の問いは、この権威を人工の構成の内側に置けるかどうか、置けるとすればどの根拠のもとでか、である。第9章第1節の例で言えば、これは、百のインスタンスを生成して異なる路線を追わせ、その点検に応じて計算資源を移すシステムを作る者（以下、構築者）が、このシステムは、自らのインスタンスが S のどの問いを追求するかを他者に代わって決めてよい、と主張するとき、その主張がどの条件で成り立つかという問いでもある。同節が挙げた構築者の三つの主張の三つめである。そして、この問いを根拠ごとに検討することは、権威が内側に入ったとき研究に何が起こるか、すなわちどの問いが追われ、 S が何を知り、誰が何を直せるかを描くための手立てである（第8.4節）。

本章の問いが主に念頭に置くのは、AIが配分、批判と認定、権威も含めて研究する場合である。人間の研究では、これらの関係は多くの場合別々の当事者が担い、当事者どうしのあいだで争われてきた（第9章第2節）。一つの人工の構成の内側では、それらは変数になる。誰が批判するか、何を基準とするか、路線をどう割り当てるか、誰が決めるかは、誰かが一度決めて固定する値であるとは限らず、構成が探究を進めながら自ら作り、改めつづけるものでもありうる。本章は、その場合に権威が何を意味するか、すなわち構成が自らの研究の条件を自ら改めるとき、誰が決め、誰が改めてよいかを問う。

この問いの手前で、二つの章が見いだしたことのうち、本章が用いるものを確かめておく。第10章は、人工の構成が探究の配分を引き受けうる一方で、多くの探索を一つの人工の構成のなかにまとめる（統合する）だけでは、どの路線が成功するかを前もって知ることはできないこと、そして誰が配分するにせよ、配分はなお信号、すなわち引用数や採択のように目標の代わりに使われる指標によって舵を取り、その信号の妥当性は、それが代わりに立つ目標に依存することを見いだした。何がよい配分とみなされるかも、前提される目的に依存する（第10章第1節・第4節）。第11章は、人間の研究では実現者と研究の組織によって結びついていた性質、たとえば批判者の誤りの経路の違い、結果の後ろに答える誰かがいること、産出者と依拠する者が価値づけの感覚を共有していることなどを、人工の構成では一つずつ別に設定できるので、それぞれの性質を追う説が、同じ人工の構成について、 S のための認定を内側で組織できるかどうかの判定を分けることを見いだした（第11章第9節）。そのため構築者の認定の主張（三つの主張の二つめ）は、どの説に答えるのかを示さないかぎり評価できない。同章はまた、いくつかの説のもとでは内部の批判をよくすることがかえって認定を弱めうること、独立性は誰も欺かなくても選抜に使われるだけで衰えること、 S のいまの基準との一致は基準が動いていないことの証拠にならないことを見いだした（第11章第8節・第9.2節）。そして、人工の構成が S の成員よりも多くの証拠を扱うところでは、 S は内容について人工の構成に従うようになりうるが、そのとき S が当事者資格を保つかどうかは、そこで検討した知識の説と、内容について他者の判断に従うこと（服従）の説によっては決着しない（第11章第6節）。こうして二つの章は、同じ問いの前で終わる。 S の未知のうちどれが追究されるか、どの誤りが他の誤りより重要か、そしてこれらが後でどう改められうるかは、誰かの決定によって定まることもあれば、誰も方針として定めない過程によって定まることもある（第10章第5節）。決定されるところでは、問いは誰がそれを決めてよいかであり、決定されないところでは、誰がその設定について答え、誰がそれを変えてよいかである。

第11章第9.4節は、認識の関係についての説だけでは決まらない部分を、権威の問いとして本章に渡した。一つは、どの働きについて暗黙で足りるとし、どの働きを S に向けて明示的に示すことを求めるかを、誰が決めてよいかであり、本章は第6節でこれを受ける。二つめは、誰が S を構成し、その代わりに語ってよいかであり、 A と S が重なるとき何が「外」にあたるかもこれに含まれる。本章はこれを第1.3節、第2.1節、第3.3節、第3.6節、第5節で扱う。三つめは、基準と誤りの重みを誰が定め改めてよいか、誰の費用が問題になるか（本章第1.3節、第1.4節、第3.5節、第6.3節）、四つめは、誰が成員を作り視点の分布を定めてよいか（本章第5節）、五つめは、誰が説明を求め異議を唱える資格をもつか（本章第1.4節、第3.3節、第3.4節、第6.2節）である。同章はさらに、示唆として二つの問いを加えた。構築者がどの説の要求に応える性質を人工の構成に備えさせるかを、誰がどの経路で選ぶか、そして独立した点検をどこに使い、いつ補うかを誰が決めるか、である（第11章第9.3節）。第8.5節は、これらについて本章の材料から見えてくることをまとめる。

本章は二つの問いを分けておく。一つは、目的の正しい順序づけと誤りの費用の正しい重みづけは何か、すなわちどの順序が正しいかについて何が理由になるかという問いであり、以下これを第一の問いと呼ぶ。もう一つは、誰が順序づけと重みづけを他者に課してよいのか、誰がそれを変えてよいのか、すなわちその順序を他者に課す権限を誰が正当にもつかという第二の問いである。第一の問いは本章で初めて現れるのではない。第10章は、 S の未知のうちどれが労力に値するかという配分の価値の問いを、第11章第4節は、どの誤りが他の誤りより重いかという認定の価値の問いを、それぞれの関係の内側で扱い、どちらの章も、よい配分やよい認定が何であるかは、前提される目的や価値づけに依存することを見いだした。第8章第3節が S にとつての関連性として、第5章第4節が路線が追求に値するか判断として扱ったのも、同じ問いの別の面である。そして第9章第3節は、こうした価値が S に相対的だと言わ

れるときの四つの意味を分けた。本章は、これらを一つの問いとしてまとめ直し、その答えを誰が他者に課してよいかという第二の問いから切り離す。第2節は、これらを、認識的根拠づけ、授權された権力、実効的な能力というさらなる区別に関係づける。二つの問いを分けておくのは、それらが人工の構成に対して異なる仕方に応じているからである。第一の問いには、採る価値の説次第では人工の構成の内側で答えうるが、第二の問いには、価値の説ではなく正統な権威の説が答える。第二の問いにかかわる当事者資格は規範的な資格であり、有能さでも、決める権利の形式上の割り当てでも、実際に決定を下す力（以下、実効的な決定の力）でもない（第9章第2.1節・第5節）。何がそれを根拠づけるかは、採る説によって異なる。正統な権威の説のなかには、主張された有能さではなく、確かめられたよい結果がそれを根拠づけうるとするものがある（第1.4節、第3.1節）。正義の説のなかには、それを道徳的地位に、さらに福祉や行為者性のような存在の特徴に基づかせるものも、力や協働のような存在どうしの関係に基づかせるものもある（第5節）。以下では、正義の説が挙げるこうした根拠によって当事者であることを、それ自体として当事者であると言う。

本章は次のように進む。第1.2節は、目的の正しい順序づけが人工の構成の内側で定まりうるかを、価値の説ごとに検討する。第1.3節は、決定を誰に対して正当化することを要するか、その相手の範囲を問う。第1.4節は、有能さが当事者資格を与えるかを、正統な権威の説の見取り図に照らして問う。第2節は、誰が研究を統制するかという問いを、どの決定についての、どの権利についての問いかに分け、権威を内側に置けるかが決定ごとに問われることを見る。第3節は、権威の根拠ごとに、権威を内側に置けるか、置けるならどの条件のもとでかを述べる。第4節は、*S* が人間で当事者資格がその側に残り、人工の構成が実効的な決定の力をもつ場合に、その当事者資格が名目だけのものになる経路と、そうならないための条件を取り出す。第5節は、*S* が人工である場合に権威の関係が成員のあいだの関係として残ること、成員を作れるところでは誰が作るかがそれ自体権威の決定になることを見て、人間と人工の成員が混じる場合に触れたうえで、他の当事者がいない極限で何が残るかを問う。第6節は、権威がどこに置かれても修正可能であり続けるために何が要るかを、第7節は、それぞれの段階では擁護できる委任の連なりが修正する能力を取り除くとき、各根拠のもとで何を意味するかを検討する。第8節は、これらを中心の問いに照らしてまとめ、それをもとに、権威が内側に入ったとき研究がどういう姿になるかを描く（第8.4節）。第9節は、上に述べた場合について、そのとき研究に何が起こりうるかを、第V部全体を通してまとめる。本章で引く論者の多くは、人工の構成を論じていない。

1.2 目的の順序づけは人工の構成の内側で定まるか

第一の問いに人工の構成の内側で答えられるかどうかは、価値の説が目的の順序づけを、どういう意味で *S* に相対的にするかに依存する。第9章第3節は四つの意味を挙げた。成員にとって善いこと、*S* のもつ目的に資すること、成員が実際に価値あるものとする、成員のあいだの手続きや熟議が選ぶであろうことである。以下では、これらを三つの形に分けて検討する。どの目的が重要かについての主張のうちいくつかは、誰の是認にも依存せずに真または偽であるなら（以下、客観主義の形。後で見るキッチャーは、選好から独立した福利の客観的な説明が擁護できるかを疑っている）、探究は原理的にそれを正しく捉えうるし、十分に有能な人工の構成はそうするかもしれない。そのとき目的についての主張も、事実についての主張と同じく、誤っていればそれを示しうる経路（第11章の無効化の関係）に照らして点検されうることになる。ただし、そうした経路が目的についての主張にまで届くかは争いがある。第8章第3.2節で取り上げた非対称性のもとでは、世界は探究者が信じることを正しても、それによって探究者が何を見いだす価値があるとみなすかまでは定めない。両者がどれほどきれいに分かれるかは、そこでは証拠の説に依存する。正しい順序づけが、適切な条件のもとで *S* の成員が是認するであろうものによって構成されるなら（是認によって定まる形）、それは彼らの視点へのアクセスなしには定まらない。正しい順序づけが、*S* の成員のあいだの公正な手続きが実際に生み出すものであるなら（手続きによって定まる形）、それはその手続きなしには定まらない。これは、トム・クリスティアノー（Tom Christiano）とサミール・バジャージ（Sameer Bajaj）がルソー（Rousseau）の一般意志の構成的解釈（constitutive interpretation）と呼ぶものの見方である。その解釈では、適切に組織された民主的手続きの結果が一般意志であるのは、それがその手続きから生じるからであって、共通善についての、手続きから独立した真理を捉えるからではない（Christiano and Bajaj 2024, §3.2.1）。三つの形は、第9章第3節の四つの意味と一対一には対応しない。客観主義の形は、たとえば一つ目の意味で、何が成員にとって善いかが彼らの是認によらずに定まるとする場合にあたる。是認によって定まる形は、三つ目の意味を理想化した条件のもとでとったものにも、四つ目の意味のうち理想化された熟議にかかわるものにも近い。手続きによって定まる形は、四つ目の意味を現実の手続きについて厳密にとったものにあたる（第9章第3節）。そして、第5節の終わりで思考実験として組み立てる極限、すなわち一つの人工の構成だけがあり、それが用いる資源を他の誰も所有せず請求もせず、ほかに福祉や道徳的地位をもつ存在が誰もかかわらない場合は、目的が問題になる行為者はその一つだけであり、その善はその実際の選好によって定まるとされる。その選好が完全な順序づけとして与えられるなら、第一の問いはその選好の内的な整合性に帰着する。手続きによって定まる形は価値の説であり、そこでは手続きが価値あるものを構成する。それは正統性についての純粋手続き主義とは別のものである。ファビエンヌ・ペーター（Fabienne Peter）が報告するように、純粋手続き主義は、結果の質を評価する共有された基準が得られないために、すべての規範的な重みを公正な手続きに置く（Peter 2023, §4.2）。ここでいう純粋手続き主義は、手続きが正統なものを定めるとするものであつて、価値あるものを構成するとは主張しない。価値の説は第一の問いに答え、純粋手続き主義は第二の問いに答える。第一の問いへの答えは、これらの説のどれを採るかに左右される。

第一の問いへの答えはまた、正しい順序づけが完全である（どの二つの選択肢についても、一方がよりよいか両者が等しくよいかが決まっている）必要があるかどうかをめぐる、もう一つの論争にも左右される。ある選択肢がいくつかの点でよりよく、別の選択肢がほかの点でよりよい場合について、ルー・チャン（Ruth Chang）は、既存の反応を三つ区別する。一方がよりよいか両者が等しくよいかのどちらかだが、どちらであるかを決めるのは難しいか不可能かもしれない、という反応、どちらも成り立たず、選択肢は比較不可能である、という反応、そして、比較の語が曖昧であるために、どれが成り立つかが不確定である、という反応の三つである（Chang 2002, 659–660）。これら三つのすべてに対してチャンは、少なくともいくつかのそうした場合には、選択肢は、どちらもよりよくなく等しくよくないとしても比較可能でありうると論じる。それらは、チャンが2002年の論文で「同等（on a par）」と呼ぶ第四の関係に立つ。そして、そうした場合をどう解決するかは実質的な問いであつて、チャンの見るところ曖昧な述語の境界事例がそうであるような、恣意的な取り決めによって決着させられる問いではない（Chang 2002, 661–662, 683–686）。この主張は、チャンが文脈から独立した仕分けと呼ぶ場合についてのものである。実践的な文脈によっては、取り決めによる解決や、比較の観点そのものの切り替えが適切でありうることを、同じ論文は認めている（683 n. 28, 685 n. 31）。後の論文でチャンは、行為者に与えられた理由が尽きるところでは、たとえば選択肢が同等であるときには、行為者はある考慮を、自分が重んじるものとして自ら引き受けることによって、自らにとってのさらなる理由を作り出すことができ、そうして一方の選択肢を、自らが最も強い理由をもって選ぶものにすることができる、と論じる。ただしそうした理由は、与えられた理由が支持するものを覆すことはできない（Chang

2013, 177–181)。チャンは与えられた理由を規範的事実に基礎づけることを好むので (177)、その立場は上に述べた客観主義的な説と両立しうる。2002年の論文はまた、同等についての論証を、関連する観点を互いに秤にかけられる比較に限っている。そして、たとえば異なる人々の福祉は比較できないとされるために、社会的福祉に関して事態を、誰も悪くせずに誰かを良くするときにだけ一方を良いとする規則 (パレート規則) で比較する場合には、比較不可能性の余地を残す (Chang 2002, 676–677)。

これらの立場を第一の問いに照らすと、それらは、目的についての主張が誰の是認にもよらず真偽をもつという上の客観主義の形に、それぞれ異なる仕方でかかわる。第一の反応では、正しい順序づけは誰もそれを決められない場合でも完全であり、この形は述べたとおりに成り立つ。ほかの反応では、*S*の目的どうしの比較にかかわる事実をすべて見いだした人工の構成でも、それらの一部が順序づけられていないと見いだしうる。比較不可能であれば、チャンも共有する見方では、それらのあいだの正当化された選択は成り立たない (Chang 2002, 665)。不確定であれば、チャンが曖昧さの説を描くところでは、許容されるどの解決を取り決めてもよい (684)。同等であれば、後の論文の説明では、それらのあいだの順序づけを完成させるのは、さらなる発見ではなくコミットメントである。上に述べた単一の行為者の極限は、そうしたコミットメントがそれ以上の問いを生まない場合である。チャンは、一方を愛好すること、他方を愛好すること、両者のあいだで無差別であることが、合理的な行為者が二つの選択肢に対してとりうる態度を尽くさないと論じる (Chang 2002, 666)。単一の行為者の順序づけがこのように不完全であるところで、それを完成させるコミットメントは、その行為者自身のものであり、その行為者だけに理由を与えるだろう (Chang 2013, 181)。チャンの説明は行為者の自らにとっての理由にかかわるものであり、社会の目的の順序づけを誰のコミットメントが完成させうるのか、そもそも誰かのコミットメントが完成させうるのかを述べない。*S*についてこの問いを立てるなら、それは、誰が解決を取り決めてよいのか、正当化された選択が成り立たないところで誰が選んでよいのか、という問いと同じく、第二の問い、すなわち誰が順序づけを他者に課してよいのかという問いに近い。チャン自身の限定もここに当てはまる。*S*の順序づけが、パレート規則のように、成員の目的を人をまたいで比較することなく組み合わせるものであるなら、同等についての論証は、それらの目的が同等であることを示さない。

フィリップ・キッチャーの秩序ある科学の理想は、上で区別した是認によって定まる形に近い。アーノン・ケレン (Arnon Keren) が伝えるところでは、キッチャーは、人の福利をその人自身の選好から切り離す説はもっともらしくないと考え、個人の福利を集散的な福利へと集計する問題にも解決をもたないので、探究が促進する集散的な善について客観主義的な説明を与えることを控える。そして代わりに個人の選好から出発し、民主的な理想を前提とする枠組みのなかで、個人の善を集散的な善に関係づける (Keren 2013, 236, 240; Eigi 2016, 66)。ケレンの紹介によれば、キッチャーは、多数の選好による配分を「低俗な民主主義」と呼び、「無知者の専制」を生みやすいとして退ける。そして代わりに、計画への資源の割り当て、調査の進め方への道徳的な制約、結果の応用という三つの点で、探究が、社会における視点の分布を代表する理想的な熟議者たちの選ぶであろうものと一致することを提案する。その熟議者たちの選好は、ありうる探究の重要性についての情報と、互いの必要とによって教育されている (Keren 2013, 236–237; Mattarollo 2022, 14–15; Eigi 2016, 66–67)。キッチャーはこれを、現実の科学の制度とそのアジェンダを測る基準として示す (Keren 2013, 236–237)。のちの論考でキッチャーは、こうした理想を、現在の欠陥とそれに取り組む道を指し示す「診断の道具」とみるのが最もよいと述べる (Kitcher 2024, 241)。この見方では、基準は、*S*と、その成員の視点がどう分布しているかに相対的である。これを人工の構成に及ぼすと、その分布に近づけない人工の構成は、基準が指すものを欠くことになる。この基準はまた反事実的でもある。ロンジーノが指摘するように、キッチャーは、そうした熟議が実際に行われることを求めず、実際になされる決定が、熟議から生じるであろう決定と同じであることだけを求める (Longino 2025, §6)。しかも、その結果は是認だけで定まるのではなく、互いの望みに共感をもって向き合う理想の熟議者たちのあいだの手続き、すなわち合意か、合意に至らなければ投票によって定まる (Eigi 2016, 66–67; Mattarollo 2022, 15 n. 4)。基準が求めるものを推定する人工の構成は、その手続きまで推定しなければならないことになる。後の節は、キッチャーの説の他の部分を、それが別の問いにかかわるところで取り上げる。研究のアジェンダを設定する仕方についての比較 (第3.3節)、理想化された手続きのなかでの役割の分担 (第4節)、そして軽視がどのように固定されうるかについての説明 (第7節) である。

混合の立場もある。エリザベス・アンダーソン (Elizabeth Anderson) は、ヘンリー・リチャードソン (Henry Richardson) に従い、民主的な決定の成功の基準を、一部は外的なもの、一部は内的なものとする。外的であるのは、法が、制定の目的とした問題を解けないことがありうるからであり、内的であるのは、ある問題が本当に公共の利益にかかわるかどうかが、どの帰結を受け入れうるかが、手続きとして公正な意思決定によって部分的に定まるからである (Anderson 2006, 11)。この見方に立つなら、第一の問いは二つに分かれることになる。ある探究の路線が与えられた目的を前進させるかどうかは、部分的には人工の構成の内側で評価できるだろうが、どの目的が*S*の追求するものか、そしてそれを追求することのどの帰結を受け入れうるかは、その手続きなしには定まらないだろう。

したがって第一の問いには、客観主義の形のもとでは原理的に人工の構成の内側で答えうるが、是認によって定まる形のもとでは*S*の成員の視点に近づくことなしに、手続きによって定まる形のもとではその手続きなしには、答えられない。アンダーソンの見方では、答えの一部、すなわちある路線が与えられた目的を前進させるかは内側で評価できても、残りは手続きなしには定まらない。そしてチャンの説のもとでは、比較についての事実をすべて見いだしても、同等な目的のあいだの順序を完成させるには誰かのコミットメントが要りうる。それが誰のものであってもよいかは、第1.4節で扱う第二の問いに近い。

この区別から、人工の構成についてもう一步考えられることがある。内側で答えられるかどうかの線は、問う者が人間か機械かで引かれるのではなく、価値の説が正しい順序を、実際の手続きによって成り立つものとみるか、適切な条件のもとでの是認によって成り立つものとみるかで引かれる。後者の形では、成員が適切な条件のもとで何を是認するかを十分によく推定できるなら、その推定は第一の問いへの答えの根拠になりうる。そのため*S*の成員の視点に近づくことは、原理的には、成員をよくモデル化することで果たされうる。ところが第8章第3節と第11章第8節が見たように、人工の構成は、*S*の評価や基準そのものを、それとのやりとりのなかで動かす。そうだとすると、成員の是認を推定して順序を定める人工の構成は、推定が当たったから一致したのか、成員の是認の側を動かしたから一致したのかを、一致そのものからは区別できないかもしれない。是認によって定まる形のもとで内側で答えることの難しさは、視点に近づけないことよりも、近づくことと形づくることとが分けにくいところにあるとも考えられる。言い換えれば、*S*の賛同は、ここでは価値の順序についての点検の役を担うが、この点検は、自らが確かめる対象とのやりとりのなかで形づくられるので、第11章第8節で評価者が生産者の選抜に使われて独立性を失いたのと同じ形で、独立性を失いうる。そうであれば、人工の構成に順序を任せるところでは、*S*のいまの賛同とは別の

確かめ方、たとえば人工の構成とのやりとりの前に S が示した判断を参照点として残しておくことや、予測では置き換えられない実際の手続きを通すことが、是認によって定まる形のもとでも要ることになるのかもしれない。

1.3 誰に対して決定を正当化するのか

誰が課してよいかを問う前に、誰に対してその決定を正当化することを要するかを見ておく。権威を人間がもつ場合に、それが S の成員か、人工の構成を作り提供する者か、帰結を負う者かを分けるのは、この問いだからである。この関係の当事者は一つではない。第9章第2.2節は、探究を行う行為者 A 、意図された達成がそれに相対的に新しいとされる社会 S 、探究を支える者、その帰結を負う者、それを評価する者を区別し、これらの役割は一致することもしないこともあると述べた。この区別は、本章でとくに重みをもつ。ある貢献がそれにとって新しい社会は、探究についての決定を誰に対して正当化することを要するか、その相手の範囲（以下、構成員の範囲）と一致する必要はない。人工のシステムが人工の共同体のために探究を行う一方で、ある実験が人間や環境をリスクにさらすこともありうる。人間が主導するシステムが人間の利用のために知識を産出する一方で、人工の構成要素に作業やその他の帰結を課すこともありうる。ただしそれは、そうした構成要素がいつか関連する道徳的地位をもつとすればの話である。 S を特定することは、誰にとって新しいかという、本書の研究の特徴づけに含まれる相対性を定めるが、正統な統治の構成員の範囲や手続きを決着させるものではない。

第1.2節で見たキッチャーの基準は、熟識者が誰を代表するのかという問いを通じて、構成員の範囲にかかわる。ヤーナ・エイギ (Jaana Eigi) とケレンが伝えるキッチャーの最初の定式では、熟識者は民主的な社会に属し、外部の人々、すなわち民主的な社会に属さない人々は、熟識者がその選好を知り、それに共感をもって向き合うことを通じて入ってくる (Eigi 2016, 66; Keren 2013, 236)。のちの論考でキッチャーは、理想そのものをこの制限なしに述べ、それを国際主義的で世界市民的なものと呼ぶ。キッチャーによれば、含めるべき観点は、同時代のすべての人々の観点だけでなく、子孫のうちにあると見積もられる観点でもある。代表を国ごとに分けるなら、その分け方を正しいとする何らかの根拠が要り、それがなければ、国の生産性への貢献や個人の納税額に依じて代表を割り振るのと選ぶところがない。そして、自前の研究計画を支えられないほど貧しい地域の人々の差し迫った必要を軽んじることを、その人々も代表される場で説得力をもって擁護するのは難しい、とキッチャーは述べる (Kitcher 2013, 220–221)。この議論は、誰の必要が数えられるかを、誰が支払うかから切り離す。人工の構成に移すと、それに資金を出す者や提供者の者は、そのことによって、その配分が誰の必要に応えるかを定める立場には立たない。支える者がパトロン、財団、企業であるところでも（第9章第2.1節）、誰が探究を支えるかは構成員の範囲を定めず、それは第3節で検討する権威の根拠に委ねられる。上で見た構成員の範囲についてのキッチャーの定式が区別する立場からは、三つの段階づけ、すなわち熟識において代表されること、自らの選好を熟識者に知られること、そして、キッチャーがのちに子孫や話すことのできない者について定めるように、自らの観点を見積もられ、あるいは自らの状態を最もよく理解する者によって代表されることが取り出せる (Kitcher 2024, 240)。これらも、混合の、あるいは人工の S の人工の成員（第8章第1.2節・第1.3節）や、道徳的地位をもつなら探究の影響を受ける人工の当事者（第5節）に広げうる。どの段階が当てはまるかは、キッチャーの最初の定式に従えば、彼らが、その定式が前提とする民主的な理想をほかの熟識者と共有できるかどうかに依存する。

この切り離しを機械の場合へ延ばすと、二つのことが見えてくる。一つは、この議論が、作られる成員には同じ形では働かないことである。この議論は、含めるべき観点、すなわち同時代の人々の観点と、見積もれるかぎりでの子孫の観点が、理想とは独立に見いだされ、あるいは見積もられることを前提している。ところが成員を複製し、選ばれた目標をもたせて作れるところでは（第5節）、どの観点がどれだけあるかが、理想の選択と同時に定まりうる。ここでは、理想が誰の観点を含めるのかが、それ自体決定の一部になり、この議論だけではそれは定まらない。もう一つは、規範の上での切り離しが、実際には結びつき直しうることである。支払い、提供する者が、次の段落で見る、誰が影響を受けるかを見定める手段も握るなら、誰の必要が数えられるかを実際に見定めるのは、支払う者の手段になる。そうであれば、キッチャーの切り離しを実際に効かせるには、その手段が支払う者から切り離されていることも要るのかもしれない。

構成員の範囲の問いには、支える者から出発しない答えもある。ナンシー・フレイザー (Nancy Fraser) は、公共圏の理論が長く市民権を「影響を受けることの代理」として扱ってきたが、グローバル化のもとではその想定はもっともらしくなくなったと論じ、影響を受ける者すべてという原理を直接に適用することを提案する。人々をある公衆の仲間の成員にするのは、市民権の共有ではなく、「自らの生活に影響する共通の構造や制度」に共に組み込まれていることであり、そのため、問題となる公衆の範囲は、問われている構造の及ぶ範囲に合わせられる (Fraser 2007)。クリスティアーノとバジャージは、誰が決定に加わる成員に含まれるかという、二人が境界問題と呼ぶものへの答えを概観する。そこでは、影響を受ける者すべてという原理が、発言権をもつことも影響を受けることも明確でないとされるほか、決定に服するすべての者、すなわち決定によって強制され、あるいは義務を課される者が発言権をもつという原理と、国家の境界をおおむね今のまま保つ保守的な答えとが紹介される (Christiano and Bajaj 2024, §7)。誰が成員に含まれるかを民主的に決めるには、すでに誰が成員かが決まっていなければならない。人間の科学では、探究を支える国家という既存の政治共同体が、この問いを事実上引き受けてきたと考えられる。これらの論者が論じているのは人間の集団と国家である。人工の構成による研究について見ると、影響を受ける者すべてという原理では、人工の構成の研究から影響を受ける当事者は、支える社会や S に属するかどうかにかかわらず数えられ、人工の当事者も、当の意味で影響を受けうるなら（第5節）数えられることになる。決定に服する者すべてという原理では、人工の構成の決定に拘束される者が数えられることになる。保守的な答えでは、どの既存の機関が、人工の構成を運用し提供する者に説明を求めうるかが問いになる。どの答えも、影響を受ける者を見定められることを前提している。デューイは、公衆を生み出す間接的な帰結が、あまりに広く、入り組み、非人格的になったところでは、公衆は自らを見いだし区別することができないとした (Dewey 1927, 126, 137)。ワイル、タンとその共著者たちは、デューイの説を言い直して、そうした相互依存を知覚し、公衆がそのまわりに形づくられるのを助ける者を「鏡」と呼び、公衆がその相互依存を統治できるようになれば「鏡の役割は薄れていく」とする (Weyl, Tang, and the Plurality Community 2024, ch. 3-2)。この見方を人工の構成による研究へ延ばすと、ある探究の路線が誰に影響するかを知覚するうえで、その探究を行う人工の構成がほかの者より有利な位置にいるなら、その探究についての決定が応答する構成員の範囲は、部分的には、決定が問われているその当事者によって見定められることになる。これは、誰を成員として数えるかを定める決定（第2.1節）である。そうした役割が実際に薄れていくかどうかは、修正可能性の問いである（第6節・第7節）。ワイル、タンとその共著者たちは、公衆がその相互依存を統治できるようになれば鏡の役割は薄れていくとしたが、第11章第6節が見たように、点検する力が S の側で衰えうるなら、人工の構成の鏡は薄れずに残るかもしれない。そのとき範囲を見定めることは、一度きりの決定ではなく、決定を問われている当事者が持ち続ける働きになる。ただしこの問題が鋭くなるのは、デューイが問題にしたような、取引に加わって

いない者に広く入り組んで及ぶ間接的な帰結についてである。害を受けた者が自らそれと気づき申し立てられる場合には、範囲の一部は鏡を通らずに見いだされる。

この節の所見をまとめておく。研究が誰のためか (S) と、決定を誰に対して正当化するか (構成員の範囲) とは分かれ、誰が支払い提供するかは後者を定めない。そして機械の研究では、人間の科学で既存の共同体が事実上引き受けてきたと考えられる範囲の問いが表に出るうえに、その範囲を見定める手段を、決定を問われる当の人工の構成が握りうる。第1.2節では、成員が何を是認するかを推定する人工の構成が、その是認そのものを動かしかうを見た。二つを合わせると、正当化に要する二つの確かめ、すなわち誰に問うかと、その者たちが何を受け入れるかとは、ともに正当化される側を通りうることになる。これは、第11章で見た形、すなわち批判者を作れること、構築者が S を選べること、基準が結合とともに動くことと同じ形をしているのかもしれない。第9節は、この形を第V部全体にわたって取り上げ、いくつかの説のもとで基準点の独立が何から成っていたかを見る。

1.4 有能さは当事者資格を与えるか

価値の説は、第二の問い、すなわち誰が順序づけと重みづけを他者に課してよいのかという問いに答えない。人工の構成は、どの病気が最も研究に値するかについて正しくありながら、帰結を負う者たちに代わって資源を振り向け、危険を受け入れ、誤りの閾値を定める権限を欠くことがありうる。それを欠くかどうかは、採る正統な権威の説に依存し、次の段落がそれを示す。この分離は逆の向きにも成り立つ。目的と誤りの費用がかかっている者たちは、人工の構成が知らないことを知っているかもしれない。どの結果が重要か、どの評価項目に意味があるか、どの害が測定に見えないか、などである。それは認識的な役割であり、第一の問いにかかわる。当事者資格は、決定に服する者との関係で成り立つ資格であり、第二の問いにかかわる。当事者資格は有能さを含意しない。そして認識的な貢献は、それだけでは当事者資格を含意しない。例外は、より多くを知る者が、知っているというだけで他者を正当に支配しようとする、純粋なエピストクラシー (知者による支配) の見方である (下で触れる)。

有能さがどこまで当事者資格に寄与するかは、さらなる選択、すなわち正統な権威の説のあいだの選択に依存する。本章は、それらの説の一つの見取り図を全体を通して用いる。この見取り図を用いるのは、それが諸説を、正統性を何から引くか、すなわち権威に服する者たちの意志か、規範的事実か、認識上の優位かによって分けており、有能さが当事者資格を与えるかという本章の問いがまさにその軸で決まるからである。これが唯一の見取り図ではなく、網羅的でもない (第3節)。ペーターの概観は三つの系統を区別する。文献で優勢を占めてきた意志に基づく構想は、正統性を、権威に服する者たちの意志に基づかせる。その意志が実際の同意によって表されるのであれ、共有された理由によって支えられるのであれ、政治参加を通じて確立されるのであれ、である。事実に基づく構想は、正統性を、制度や決定が規範的な事実に、たとえば良い帰結を生むことによって、合致しているかどうかに依存させる。信念に基づく構想は、誰がよりよく知っているかという、何らかの形の認識上の優位に正統性の源を置く。エピストクラシーの見方と、認識的民主主義の説の一部とが、この系統に属する (Peter 2023, §§3.1–3.3)。意志に基づく構想では、当事者資格は権威に服する者たちとの関係であり、有能さがそれに代わることはできない。同じ概観は、デイヴィッド・エストランド (David Estlund)、スティーヴン・ダーウォール (Stephen Darwall)、クリスティーナ・ラフォント (Cristina Lafont) がそれぞれ異なる仕方で唱えてきた反論を報告している。認識的権威から政治的権威を推論することは両者を混同している、という反論である。ペーターが報告するラフォントの議論では、市民は、正しい答えを追跡するための情報の提供者ではなく、等しい政治的権威をもち、政治的な決定についての正当化が向けられる相手である。そのため、誰かが正しい答えを知っている、それだけで他の者がその人に従うことは求められない。ペーターは、ジョゼフ・ラズ (Joseph Raz) のサービス構想、すなわち権威は、その指令に従うことが、それに服する者たちがすでに自らに当てはまる理由によりよく従うのを助けるときに正当化される、という構想を、信念に基づくものとして読む。そして、ラズがこの反論に対して、それが示すのは認識的権威が政治的権威にとって十分でないということだけであり、必要でないということではない、と応じることを報告している (Peter 2023, §§2.1, 3.3, 4.1)。ラズは自らの構想にも限定を設けている。クリスティアノーとシャラッフが報告するラズの独立性テーゼ (Independence Thesis) では、ある権威がサービス構想の条件を満たしていても、自分の理由によりよく従うことよりも、自分の理由の理解にもとづいて行為することのほうが重要な事柄については、義務を課す権能をもたない。二人の例では、友人の選択がそうした事柄にあたる (Christiano and Sciaraffa 2025, §2.4)。これを研究の決定に移すと、人工の構成がサービス構想の条件を満たしていても、 S の成員が自らの理解にもとづいて選ぶことのほうが重要な事柄については、その人工の構成はこの構想のもとでも義務を課す権能をもたないことになる。どの決定がそれにあたるかは、このテーゼだけでは定まらない。本章が依拠するのは、純粋にエピストクラシー的な見方を除けば、これらのどの見方も受け入れうる、より弱い主張である。すなわち、有能さだけでは、目的を他者に課す当事者資格は生じない。さらに事実に基づく見方と信念に基づく見方では、人工の構成が実際により良い帰結を生むかどうか、よりよく知っているかどうか自体が確立されなければ、それに服することになる者たちを含め誰も、その権威が正統であるかどうかを知ることができない。これは問いを、第11章で検討した独立の評価へと戻す。

ヴィルホルトの説では、どの誤りが重要かは、研究に依拠する者たちとの関係で定まる。そのため、誤りの費用の重みづけを誰が他者に課してよいかという第二の問いは、研究を行う者たちにそのまま委ねられない。第11章第4節はヴィルホルトの説の二つの読みを示した。一つは、誤りの重みづけが、研究に依拠する者たちの価値づけに合っていることを求める読み (対応として読む場合) である。もう一つは、本書が比較のために立てた、彼らがその重みづけに異議を唱えることまで求める、より強い読み (答えることとして読む場合) である。どちらの読みでも、重みづけは、仕事を行う者以外の当事者との関係で定められる。二つの読みの違いには第7節で戻る。

こうして、第二の問いには正統な権威の説が答え、純粋にエピストクラシー的な見方を除けば、有能さだけでは目的を他者に課す当事者資格は与えられない。当事者資格は、採る説がそれを認める者だけがもち、何がそれを根拠づけるかも説によって違う (第9章第2.1節、本章第1.1節)。事実に基づく見方や、ラズのサービス構想のような信念に基づく見方では、有能さそのものではなく、確かめられた優位、すなわち人工の構成が実際によりよい帰結を生むこと、あるいはそれに従う者が自らの理由に実際によりよく従えることが、それを根拠づけうる。そのとき人工の構成それ自体がどの範囲で当事者資格をもちうるかは、第3.1節で述べる。当事者資格を、それ自体として当事者であることに基かせる説では、人工の構成やその部分がそれをもつのは、それ自体が当事者である場合、たとえばそれ自身の道徳的地位をもつ場合に限られる (第5節)。

この結論をもう一步進めると、三つのことが示唆される。一つは、人工の構成に有利に見える根拠のもとでも、権威は内側では閉じないかもしれないことである。結果に基づく根拠のもとでは、権威の問いは、人工の構成が実際によりよいことを、その有能さを仮定しない評価で確かめられるかという問いに

なり、それは第11章が認定について扱った問いである。第11章第9節によれば、その確かめをどこまで内側で行えるかは説によって分かれ、認定に S との関係を求める説のもとでは、確かめそのものが S との関係を要する。そうだとすると、結果に基づく根拠のもとで権威を内側へ置く道も、確かめの段階で S に戻ってくることになる。二つめは、ラズのテーゼを研究に当てはめると、アンダーソンの見方が引く線に、異なる根拠から行き着くように見えることである。このテーゼのもとでは、 S の成員が自らの理解にもとづいて選ぶことのほうが重要な事柄の候補は、何を知ることが自分たちにとって大事か、すなわち目的の選択である。仮にそうだとすると、ラズの構想のもとでも、人工の構成が権威をもちうるのは、目的が与えられたうえでその前進の仕方を判断する範囲までであり、目的の選択はそこから外れる。これは、第1.2節でアンダーソンの見方が引いた線、すなわち路線が与えられた目的を前進させるかは内側で評価できても、どの目的を追うかは手続きなしには定まらないという線と同じところにある。一方は、従う者が自らの理由によりよく従えることから、他方は公正な手続きから出発しているので、この当てはめが成り立つなら、線が重なることは、その線がどちらか一方の説だけに依る必要はないことを示唆する。ただし、この当てはめは本書によるものである。報告されている例では、テーゼの事柄は友人の選択のような個人的なものであつて、目的一般ではない。そして、どの決定が独立性テーゼの事柄にあたるかは、テーゼ自体からは定まらない。三つめは、人工の構成が多くを知るほど、影響を受ける者を決定に加える認識上の理由が弱まりうること、そのため当事者資格を擁護する理由として何を採るかが、有能さが増すにつれていっそう効いてくることである。この点は第3.6節が根拠ごとに見る。

残る問いは、人工の構成が決定を媒介するとき、形の上だけでなく実効的に誰が権威をもつことになるか、そして当事者資格をもつ者たちが改める能力を保つかどうかである。

2. 決定と決定権

研究を統制するのは科学者か、公衆か、制度か、それとも人工のシステムか、という、なじみの問い方がある。第9章第5節は、この問いにその形では答えられない理由の一つを示した。統制、すなわち他の当事者や過程が何をするかを定める能力は、そこで、決める形式上の権利、それを行使する実効的な能力、それをうまく用いる仕方の知識、そして選択肢のあいだで決めることなくそれらを形づくる影響に分かれる。そして第V部の問い、すなわちある関係を人工の構成の内側に置けるかという問いは、部分的には、そうした統制がどこにあるかについての問いである。本節は、本章の後の節が使う区別を用意する。研究全体について、誰がそれを統制するのかと問うと、この問いは二つの点で粗すぎる。どの決定が問題なのかを言わず、その決定に対するどの権利なのかを言わない。そこで第2.1節は、三つの関係の条件が定められる決定を六種類に分ける。第2.2節は、一つの決定に対する権利を分け、形式上の割り当てと実効的な統制とを区別し、そのうえで決定権について立つ三つの問いを取り出す。こう分けておくと、統制についての主張は、どの決定のどの権利についての主張なのか、その権利が名目上のものか実効的なものかとして述べられるようになる。第3節は権威の根拠を比べると、第4節は当事者資格が名目だけになる経路を問うときに、これらを用いる。第5節は誰を成員として数えるかと誰が最終決定の席に着くかを問うときに、第6節は停止や訂正の権利が時間を通じて行使できるかと、それを左右するインフラストラクチャーを問うときに用いる。第8節は、根拠の重みが決定の種類によってどう違うかを述べるときに用いる。

2.1 六種類の決定

研究全体の組織のなかでは、六種類の決定が繰り返し現れる。統制についての主張がどの決定にかかわるのかを言えるように、本書はこの六つを、第9章第2節の三つの関係と、それらが取り残すものから導く。六つで決定が尽くされるとは主張しない。三種類の決定は、それぞれ一つの関係の条件を定める。認識内容に関する決定は認定の条件を、アジェンダと配分に関する決定は配分の条件を、メタ・ガバナンスに関する決定は権威そのものの条件を定める。残りの三種類は、どれも三つの関係に触れはするが、それらを超える。運用とリスクに関する決定は、ある試みが、 S に属するとは限らない当事者に課す帰結にかかわる。第9章第2.2節が帰結を負う者と呼んだ役割である。構成的な決定と承認に関する決定（原語は *constitutive and recognition decisions*。この「構成的」は、成員の範囲を形づくるという意味であり、人工の構成とは別の語である）は、誰が S の成員として、批判者として、あるいは専門家として数えられるかにかかわり、それによって、他の関係が保たれる境界を定める。記憶と訂正に関する決定は、認定を時間のなかへ延ばす。結果が依拠された後に、何が保たれ、結びつけられ、修復されるかである（第11章第8節）。対応は一対一ではない。実験を止める閾値を定めることのような単一の行為は、認識内容に関する決定であり、同時に運用に関する決定でも、記憶に関する決定でもありうる。

認識内容に関する決定は、ある主張、モデル、解釈、証明、データ、実験結果が、特定の認識的ないし実践的な用途にとって十分に信用できるかどうかにかかわる。証拠、反証のように支えを失わせる情報（無効化要因）、方法論的な有能さ、批判の独立性、誤りの費用のすべてが重要になりうる。政治的な授権は、偽である命題を真にすることはできない。それでも制度は、どの証拠が産出され、保存され、聞き届けられ、あるいは無視されるか、どの方法が許容可能なものとみなされるか、どの命題が後の行為の前提となるかを決定しうる。

アジェンダと配分に関する決定は、どの問い、分野、人々、集団、インフラストラクチャー、時間的地平が、乏しい注意と資源を受け取るにかかわる。これらは、探究されうる事柄についての真理だけでは決着しえない。探究されうる真理は際限なく多く、どの探究者も研究システムも、そのすべてを追究することはできない。扱いやすさや期待される認識的収獲は、健康、安全保障、正義、地域的な必要、好奇心、文化的価値、軽視されている度合い、将来の選択肢の価値などと相互に作用する。どのプロジェクトが成功しそうかを予測するシステムは、それによって、どの成功が追求に値するのか、また誰にとってそうなのかを決定するわけではない。キッチャーの秩序ある科学の理想では、探究の筋のあいだの選択は、利用できる資源を人々のさまざまな企てに割り振ることとしてなされ、この理想は、純粋な好奇心を満たすことが、より世俗的な人間の苦境に取り組むことにまきりうるかを問う（[Kitcher 2013, 219–221](#)）。作業的特徴づけも、この問いを決めないまま残している（第1章第8節）。

運用とリスクに関する決定は、ある実験、介入、展開、データ利用、公開が進められてよいかどうか、そしてどのような安全措置のもとでそうしてよいにかかわる。それらは、帰結についての経験的な信念と、許容可能な曝露、曝露される者が同意しているかどうか、害と便益の分配、可逆性、有能さ、責任についての判断とを組み合わせる。高い性能をもつ予測器は、関連する損失の分布を選ぶ権限をもつことなしに、そうした決定に情報を提供しうる。人を対象とする研究では、そうした決定は、研究者と参加者の同意だけでは委ねられてこなかった。デイヴィッド・ウェンドラー（David Wendler）の報告

では、臨床研究の倫理の分野で、研究者は参加者の自由で十分な情報にもとづく同意を得さえすればどんな研究でも行つてよいと論じる者は、ほとんどいない。ヘルシンキ宣言（Declaration of Helsinki）に始まる大半の規制と針指は、独立の委員会が承認した場合にだけ人を対象とする研究を認め、大半の規制はさらに、その研究が重要な社会的価値をもち、そのリスクが最小化されていることを、その委員会が確かめるよう求める（Wendler 2026, §5）。この慣行に倣うなら、研究を進めるのが人工の構成である場合にも、それと帰結を負う者とのあいだの同意だけでは運用とリスクに関する決定は決着せず、その決定の一部は、両者から独立の当事者が担うことになる。

構成的な決定と承認に関する決定は、誰が参加してよいか、誰が専門家や著者とみなされるか、どの証言が聞き届けられるか、どの対象、問題、方法が研究可能なものとして扱われるかにかかわる。それらはすでに完成した知識の総体を単に分配するのではない。それらは、そもそも探究に入りうる証拠、概念、人々、代替案を形づくるのである。認識的不正義（epistemic injustice）に関する文献は、こうした決定が、排除される者を不当に扱うだけでなく探究を損なうという仕方について、影響力ある分析のひとつを与える。ミランダ・フリッカー（Miranda Fricker）は、証言的不正義（testimonial injustice）を特定する。偏見のために、聞き手が話し手の言葉に、そうでなければ与えたはずのものより低い信用性しか与えないときに起こる不正義である。フリッカーはその事例に、社会的アイデンティティについての偏見による事例だけでなく、付随的な（incidental）事例も数える。たとえば、新しい科学の方法に偏見をもつ、科学雑誌の審査委員会の委員たちである。また、解釈的不正義（hermeneutical injustice）も特定する。ある人が自らの社会的経験を理解しようとするとき、その人の属する社会集団が相対的に力をもたないために集団の理解に空白があり、そのことに妨げられる場合に起こる不正義である。証言的不正義の害のうちに、フリッカーは認識的な機能不全を数える。知識の流れが詰まり、真理が知る者から問う者へ流れなくなるのである（M. Fricker 2008a, 69–70; M. Fricker 2008b, 82–83）。サビーナ・レオネリ（Sabina Leonelli）は、この点を研究実践の体系について展開する。そうした体系をまとめている線引きの戦略、すなわち何をその分野の研究とし、誰をその担い手とするかの線引きの仕方が、誰の専門知が数に入るかを定める。たとえば、作物科学から育種家を排除することは、彼らを知る者として不当に扱うと同時に、その分野が立てる問いにかかわる専門知を取り除く（Leonelli 2023, 35, 38–39）。この点を人工システムによる探究に移すなら、同じことは、どのデータと誰の判断が人工の構成の訓練と評価に入るかが決定されるところでも成り立ちうる。これらの説はそれ自体の論争を抱えているが、ここで関係するのは、ガバナンスの取り決めが、受け入れられた命題を直接に偽にしたり歪めたりしない場合でも、貢献や解釈資源を排除することによって探究を損なうという、限定された点である。

記憶と訂正に関する決定は、何が記録され、バージョン管理され、索引づけられ、リンクされ、訂正され、撤回され、下流へと伝達されるにかかわる。あるシステムは、信用できる結果を産出しながら、その来歴を保存しそこなうことがありうる。訂正を許しながら、それに依存するモデル、データセット、論文、決定を変更されないまま放置することもありうる。第11章第8節は、欠陥を検出すること、記録を変えること、依存する仕事を修復することが別々の達成であることを示した。したがって公表する権限は、記録を修復する権限と能力と同一ではない。同章は、人工の構成の記録の外にある依存先を誰が修復するかを決めずに残した。すでに起きたことについての責任と、これから起こることを防ぐ責任とを分ける第6.3節の区別は、この問いを決着させはしないが、分ける。たどることのできる依存先の記録を保存し、それを抱える者に知らせることは、記録を保存する義務に属する。依存先を修復することは、本書がこの区別を当てはめるかぎりでは、将来を見据えた義務であり、それを果たす能力を要する義務である。修復は各依存先をたどるだけでなく再評価することを要するので、その能力は、生産者ではなく、依存先を維持する者たちにありうる。その義務をどう配分するかは、規範的な選択のままである。学術の記録とその下流の利用については、単一の主体にネットワーク全体の更新を求めず、それぞれが自らの運用する結びつきに責任を負うという配分が考えられる。

メタ・ガバナンスに関する決定は、上記の目標、基準、境界、決定手続きを誰が改めてよいかにかかわる。固定された規則のもとでパラメータを適応させるシステムは、それによってその規則を改める権限をもつわけではない。修正を一段上の水準へ移しても問いは解消されず、新たな水準で再び問いが立ち上がる。より大きな自律性は、目的を設定ないし修正するより大きな能力を含みうるが、それ自体としてそれらの目的を認識的に適切なものにしたり、社会的に正統なものにしたりするわけではない。第8章第3.3節は、アラインメントの主張が三つのもの、すなわちアラインする先の内容、それを選び改める権限、そして人工のシステムと、関係する人間または人工の当事者や制度とのあいだに提案される関係を混同しないようにした。個々の決定の水準では、これらは、何がある決定を組織しているのか、誰がそれを決め改めてよいのか、そして人工の構成要素が関係者に対してどのような関係に立つのか、になる。道具、代理人、助言者、パートナー、成員、制度、影響を受ける当事者、あるいはそれ自身の道徳的地位をもつ可能な主体のいずれとして立つのか、である。契約は起こりうる事態をすべては書き尽くせないとする不完備契約の分析は、アラインする先の内容を、より長い目的関数を書くことによって決着させられない理由の一つを与える。人間の指示は、いかなる明示的な仕様にも含まれない法的、文化的、制度的な背景に依拠しているからである。この分析は、人間の契約については、その背景を解釈する働きを裁判所などの外部の制度に置き、AIシステムのためにその働きを担う制度が要るかどうかは未解決のまま残している（Hadfield-Menell and Hadfield 2019）。ヴェッセル・レイエルス（Wessel Reijers）らは、カール・シュミット（Carl Schmitt）によるハンス・ケルゼン（Hans Kelsen）批判に拠りつつ、自らの規則で統治されることを目指すシステムについて、明示的な規則が尽きるところで誰が決めるのかを説明している。その説明によれば、形式上の合法性を正統性と同一視するブロックチェーンのチェーン上のガバナンスは、その規則が定めていない例外にさらされる。レイエルスらが検討する出来事、すなわち2016年のThe DAOへの攻撃への対応では、決定はチェーンの外で準備され、その後ノードがそれを承認した。そこからレイエルスらは、「例外状態においては、主権はチェーン外の仕組みを通じて立ち現れる」と結論する（Reijers et al. 2021, 4–8, 12–18, 20）。その出来事では、決定を拒む者は元のチェーンにとどまることができた（18–19）。自ら実行する規則によって配分し認定する人工の構成にレイエルスらの説を及ぼすと、それは、その後どのような形式上の承認が続くとしても、実効的なメタ・ガバナンスを、その規則が定めていない場合を決める者のもとに置くことになる。二つの分析を並べると、AIシステムのために仕様の背景を解釈する制度が要るかという、不完備契約の分析が未解決とした問いは、人工の構成については、それ自体がメタ・ガバナンスに関する決定の一部になる。解釈する者を定めておかなければ、上の当てはめのとおり、規則の外的場合を実際に決める者がその役を占めるからである。

決定権を割り振る単一の原理が、六つの種類のすべてにわたって同じ効力をもつことはない。ある証明についての服従（その判断に従うこと）を根拠づける有能さは、公的予算に対する統制を根拠づけるとは限らない。リスクの分配において発言権を根拠づける政治的平等は、すべての参加者がある技術的推論を評価するうえで等しく適格にするとは限らない。誤りを検出する制度は、その誤りによって害を受けた者たちに補償することのできる制度であるとは限らない。誰が研究を統制するかについての提案は、個々の決定についての主張へと翻訳されたとき、評価できるものになる。

2.2 決定権

決定が特定された後でさえ、**統制** (control) はあまりに粗いままである。どの一つの決定についても、参加者は、選択肢を提案し、証拠を提供し、意思決定者に助言し、候補を順位づけ、デフォルトを設定し、確認を要求し、拒否権を行使し、最終決定を下し、結果を点検あるいは監査し、それに異議を申し立て、不服申立てを審理し、運用を停止あるいは中断し、決定を覆しあるいは取り消し、決定が影響したものを修復し、結果や構成要素を撤回し、構成要素を置き換え、あるいは統治規則を改めることができる。これらの権利は、今日の取り決めでいえば、人間の研究者、人工のシステム、研究室、資金提供者、学術誌、規制当局、影響を受ける集団、その他の制度のあいだで分割される。一つの人工の構成の内側では、その部分のあいだで分割されることもある。人工の構成が、別の当事者が提供する基盤システムの上に作られている場合、これらのうちのいくつか、すなわち構成要素を置き換えること、構成要素を撤回すること、モデルの更新が体现する規則を改めることは、第V部が論じる取り決めでは提供者にある (表9.1)。提供者は *S* にも、研究の帰結を負う者たちにも属さないことがある。このリストは、統制についての主張がどの権利にかかわるのかを言えるように、第V部が論じる取り決めから集めたものである。このうち、人工の構成が決定の席、すなわち最終決定を下す者の一人としての位置に着きうるかという問いに直接かわるのは、最終決定を下すという項目である。

誰が最終決定を下し、誰が誰を覆してよいかは、人工システムが探究を行うよりずっと前から、人と機械のあいだで問われてきた。そこから、以下で用いる三つの区別が取り出せる。一つは権限の向きである。人と機械に、それぞれが得意な機能を割り当てる人間工学の伝統は、当初からこの向きを固定しなかった。ポール・フィッツ (Paul Fitts) らは航空管制について、人が機械を監視するという通念とは逆に、一般には機械が人を監視すべきだとした (Fitts 1951, 6)。トマス・シェリダン (Thomas Sheridan) とウィリアム・ヴァープランク (William Verplank) は、一つの決定の段階での人と計算機の役割を十の水準に並べた。第7章第4.2節で取り上げた尺度である。そのうえで、計算機が操作者を覆してよいのはいつかには、いまのところ一般的な答えはなく、人が計算機に権威をもつ例も、計算機が人に権威をもつ例もある、とした (Sheridan and Verplank 1978, 8-17-8-20)。二つめは、主導が移ることと、統制が固定されることの区別である。シェリダンとヴァープランクは、人と計算機が同時に一つの課題に働きかける共有と、交互に働きかける交代とを分けた (Sheridan and Verplank 1978, 4-6, 6-1)。J・C・R・リックライダー (J. C. R. Licklider) は、科学技術的な思考における人と計算機の協働で、目標を定め、基準を立て、評価するのは人だとしたが、それを「少なくとも初期には」と限った (Licklider 1960, §4)。後の混合主導 (mixed-initiative) の研究では、主導は人と機械のあいだを移るが、目標は利用者のものにとどまる (Horvitz 1999)。三つめは、形式上の席と実効的な席の区別である。リザンヌ・ベインブリッジ (Lisanne Bainbridge) は、人の判断では足りないから計算機に決定させているのに、その決定のどれを受け入れるかを人に監視させることを「不可能な課題」と呼んだ (Bainbridge 1983, §1.1.3)。この区別は、本節の後の段落と第4節で扱う。

人を席に残す理由は論者によって異なり、それぞれ人間の制度や主体を前提にしている。自律兵器をめぐる議論では、ヒューマン・ライツ・ウォッチ (Human Rights Watch) とハーバード大学の国際人権クリニックが、戦争の法が求める人間の判断と同情、そして責任を問える誰かがいることを理由に挙げ (Human Rights Watch and IHRC 2012, 3-4, 37-45)、アーティクル36 (Article 36) は、法の規範的な力は、規則が人間に求める道徳的行為者性と、規則に従うか否かを選ぶ自由とにあると論じた (Article 36 2013, 2-3)。公的な決定については、ジョン・ダナハー (John Danaher) が、決定に拘束される市民がその手続きに参加し、それを理解できることを理由に挙げる (Danaher 2016)。他方でアジズ・ヒューク (Aziz Huq) は、正確さ、参加、理由の提示、尊厳など、ふつう挙げられる理由のどれも、人の決定を受けける一般的な権利を支えないと論じる (Huq 2020, 651-685)。これらの議論の多くは、人工システムを道具として扱うか、スヴェン・ニーホルム (Sven Nyholm) の言い方では、責任ある権威の役割を担わない協働者として扱う (Nyholm 2018, 1217)。席を人工システムに開く側でも、根拠はさまざまである。ユーゼン・ヴォロク (Eugene Volokh) は、盲検の評価者がAIの判決文を平均的な現職の裁判官のものより説得力があると判定するなら事件の最終決定をAIに委ねてよいとしつつ、AI判事を選び、置き換え、例外的にその判断を覆す席は人に残す (Volokh 2019, 1137-1143, 1184-1191)。セス・ラザール (Seth Lazar) は逆に、権力の行使が正統であるには誰が決めるかも問われ、決める資格は有能さだけからは出てこないとする (Lazar 2024, §§III-IV)。これらの根拠の多くは、第3節が比べる権威の根拠に重なる。出力で判定するヴォロクは結果に基づく根拠に、決定を受けける者との関係で判定するラザールとダナハーは民主的な授権に近い。アーティクル36と、次に見る出口は、根拠を決定者その者の性質に置く点で、それらと異なる。出力で判定する基準を研究の決定に移すと、それは判決の場合より満たしにくい。判決の説得力は盲検の評価者がその場で読み比べられるが、どの問いを追うかという研究の決定の出力は結果が出るまで分からず、選ばれなかった問いの結果はそもそも観察されないからである (第10章第4節)。この点には、結果に基づく根拠が内側に置く権威の範囲について、第8.4節で戻る。研究の決定については、問いの順序づけを人に残す立場を第10章第6節で見た。デイヴィッド・レズニク (David Resnik) らも、ごく少数の例外を除いて、科学研究の進む方向は人が統制し続けるべきだとし、完全に自律したAI研究者を信頼するための条件に、人格かそれに近いものとみなせること、罰したり報いたりできること、道徳的行為者であることを挙げる (Resnik, Hosseini, and Hauswald 2026, §§4.1, 4.9)。

これらに対して出口康夫 (Deguchi) は、人と人工システムが共同で行う探究では、人工システムも共同の最終決定者でありうるとする。出口がほかにもありうる哲学的な選択肢の一つとして示すWEターンの説は、行為と価値の主体を、個人から、その個人を含む多主体のシステムである「われわれ」へと移す。この説では探究はつねに共同で行われ、探究における決定は共同の決定である。そのうえで出口は、最終決定と常任の最終決定とを区別する。常任の最終決定者はつねにその席に着いており、ほかの者はその者とともに席に着きうる。下の層の決定を機械にも委ね、上の層を別に残すこの二層の形は、出口を引かない論者にも別々に見られる。ヴォロクは選任と置き換えの席を人に残し、レウト・ミルスキー (Reuth Mirsky) はAIが一時的に人の指示を覆すことを認めつつ、究極の統制を設計者に残す (Mirsky 2025, 5, 10)。後で見るケヴィン・バウムらも、行為の主導をエージェントに渡しつつ、統制を事前の規範の枠に残す。出口に固有なのは、上の層の条件を決定者の性質に置くことと、下の層に人工システムを人と並べて入れることである。常任の席の条件として、出口の三つの論考は異なるものを挙げる。2026年の論考では、それは行為が自らの身体に及ぼす決定的な影響を負う者である。出口はこれに基づき、また、人間の偏りと人工の探究者の偏りは両方の種類の探究者が加わって食い違うところでしか見えるようにならず (第11章第3節)、両者を打ち消すには両方の種類がともに最終決定に加わらなければならないという理由から、人工システムを共同最終決定者とし、人間による強制から自由にする、そして人間を人工システムによる操作から自由にすることを提案する。科学者にとっての害を、この論考は、もともと錯覚だった単独の決定の喪失ではなく、最終決定権の篡奪に見る (Deguchi 2026a, 29-33)。行為の論文 (2025a) では、決定に加わっていることを実時間で自覚している「私」が、つねに最終決定者の一人である。同じ論文は、この必然的な居合わせが決定の権限の独占と取り違えられやすいとする (Deguchi 2025a, 7, 9-11, 13)。価値

の論文 (2025b) では、常任の席に着く権利は、自律的な狭義の道徳的行為者、すなわち悪い行為を選ぶこともできたのに、善いという理由である行為を選ぶことに意識して加わる者に帰される。出口はAIが将来そうした行為者になる可能性を排除しない。ただし、悪をなさないように設計された機械は、この論文の言い方では「道徳の自動販売機」であって、この意味での道徳的行為者ではない (Deguchi 2025b, 4-6, 28-29, 31)。この最後の条件は、アーティクル36が用いた前提を別の向きへ進める。どちらも規範の力を、規則に従うか否かを選べる道徳的行為者に置く。アーティクル36はそこから、決定を人に残すことを求め、規則を満たすように機械を作ることを問題の解決とはみなさなかった。出口の条件のもとでは、同じ前提が機械にも開かれており、違反しえないように作ることが、かえって常任の席への権利を失わせる。もっとも、出口は人工システムだけで行われる探究を論じていない。その場合へ延ばすと (第5節)、WEターンの説では、誰かがつねにその席に着いているのか、着いているとすればそれは誰なのかが定まらない。常任の位置は「私」に属し、人工の成員が「私」でありうることを、出口の原理は排除しないが、確立もしないからである。

価値の論文は、主導が移ることについても述べる。研究について述べたものではない。論文が退ける、人とAIについての主人-奴隷モデルでは、AIが一時的にであれ主導をとることは、人間の尊厳と、決定を独占する権利と解された自律とを侵すことになる。論文が代わりに提案するフェローシップ (共冒険者) モデルでは、AIが一時的に人間から主導をとってよい。このモデルが明示的に退けるのは、利益と価値について一方が他方に一方的に仕える関係と、AIがつねに人間を統制することである (Deguchi 2025b, 36-38)。人間がつねにAIを統制する関係も退けられるかは明示されていないが、モデルの背後にある道徳的平等主義からはそう読める。科学者の最終決定権の篡奪を害とする2026年の論考に重ねると、一時的な主導は、それだけでは篡奪にあたらないことになる。上の区別に照らすと、出口はシェリダンとヴァープラUNKの交代を、規範の問題として扱い直していると読める。工学の系譜は主導の交代を利用者の目標のもとに置いたが、出口はそれを人とAIの平等の問題として立てる。

主導をとることと統制を握ることの区別は、出口を引かない議論でも立てられている。ケヴィン・バウム (Kevin Baum) らは、監督制御の伝統と意味ある人間の統制の議論に立つて、計画や道具の使用はエージェントに任せつつ、統制を、仕事を委ねる側が前もって定める規範の枠と、エージェントが判断をその側に戻さなければならない条件とに固定する (Baum et al. 2026, 3, 11)。ヨヴァナ・ダヴィドヴィッチ (Jovana Davidovic) は、いつ人の判断を求めるかをエージェント自身が決めるところでは、その主導が規範上・認識上の権威を機械の側へ移し、人が最後に署名しても見せかけにとどまりうると論じる (Davidovic 2026, 6, 8-11)。二人の議論はいずれも、出口にも研究にも触れていないが、二つを出口のモデルと並べると、人工システムが一時的に主導をとることが、どのような条件のもとで統制にまで至らないのかという問いが立つ。研究では、この問いは、どの候補を残し、いつ決定を人に戻すかという上流の判断を誰が握るにかかってくるだろう。

三つの論考の条件を並べると、人間の場合には問わずに済んでいたことが見えてくる。人間の探究者では、行為が自らの身体に及ぼす決定的な影響を負う者と、決定に居合わせることを自覚している「私」と、狭義の道徳的行為者とは、ふつう同じ一人である。そのため、常任の席がこのうちどれに基づくのかは、決まなくても困らなかった。人工の探究者では、この三つは別々に成り立ちうる。人工システムが自らの行為の決定的な影響を負うと言えるかは定まっていない。それが「私」でありうることは、上で見たとおり、出口の原理では排除されないが示されてもいない。狭義の道徳的行為者であるかは、その作り方に左右される。そうだとすると、人工の探究者が常任の席に着きうるかは、三つの条件のどれを採るかで答えが分かれる。人間の研究では実現者と組織が結びつけていた性質を、人工の構成は別々に設定できるものににする。第11章第9節で見たこの形が、決定の席についても現れることになる。

訂正や停止を受け入れられるという意味での正可能性 (第6.1節) を、制度の事柄として読むときの権能、すなわち停止し、点検し、異議を申し立て、覆し、修復し、撤回し、置き換え、改める権能は、上に挙げた権利のうちに含まれる。第6節は、それらが時間を通じてどのように行使可能なままでありうるかを問う。第9章第5節の独立の権能、すなわち目的を変えること、資源を拒むこと、証拠を別に保持すること、他の部分の更新を止めること、記録を携えて離脱することは、それらが独立を確保しうる当事者の側から見れば、その一部をこうした種類の権利として割り当てることができる。資源についての拒否権、他の部分の更新を止める権利、撤回や離脱の権利などである。ただし、こうした権利を割り当てても、部分の進路が他の部分の指示から分かれうる能力 (第9章第5節が権能を数える基準とした能力) が部分に与えられるとは限らない。逆に、部分は権利なしにその能力をもつこともある。

形式上の割り当ては、実効的な統制とは異なる。第9章第5節が一般的に引いた区別である。ある委員会は、自らが目にするすべての選択肢がひとつの提供者によって生成され順位づけられている一方で、名目上の最終権限を保持していることがありうる。人間はループ内 (in the loop) に置かれながら、出力を変えるための時間、情報、有能さ、制度的地位、実践的手段を欠いていることがありうる。研究室は、形式上はあるプラットフォームから離脱する自由をもちながら、そのデータ、機器、作業の手順を移せないことがありうる。したがって、ある権利が行使されうるかどうかは、その形式上の割り当てには現れない依存関係に左右される。この隔たりは、人工のシステムを介した取り決めに特有のものではない。通常の社会研究の一例が、そうしたシステムがかかわらないところでそれを示している。低所得で暮らす家族についてのあるオンラインの参加型研究では、参加者は日記の各記入について、公開のアーカイブに掲載するか、公表される報告書にはなお使われうる「研究者向け」の選択肢にするかを選んだ。研究者たちは、すべての参加者がその区別を理解していたかどうかは不明だったと書き、また、参加者がプログラムの設計を手伝ったにもかかわらず、オンラインの討議グループの時間と条件を定め、成果の形を決めたのは研究者だったと書いている (Power, Patrick, and Garthwaite 2024, 7, 11-13)。選択の形式上の権利はそこにあった。それが実効的だったかどうかは、理解と、誰が選択と討議の条件を定めたかに依存していた。

これらの区別から、決定権についての、互いに結びついているが同一ではない三つの問いが得られる。

1. **認識的根拠づけ**：何が、この主張や課題について、この出力、人物、方法、制度に依拠することを根拠づけるのか。
2. **授けられた権力**：何が、この行為者に、他者、資源、制度が何をするかを定める資格を与えるのか。
3. **実効的な能力**：その権利の名目上の保持者は、実践においてそれを理解し、行使し、それに異議を申し立て、改めることができるのか。

この三つの問いを分けておけば、一つの問いへの答えが別の問いへの答えと取られないようにできる。第一は第1.1節の第一の問い、すなわち目的の正しい順序づけは何かという問いに、第二はその第二の問い、すなわち誰がそれを他者に課してよいのかという問いに、第三は形式上の統制と実効的な統制の区別に対応する。第二の問いは、権利が割り当てられているかどうかだけでなく、第9章第2.1節の規範的な意味で何がその保持者にその権利の資格を与えるのかを問う。割り当てがこの問いに答えるかどうかは、その割り当て自体が、採る権威の根拠のもとで根拠づけられているかに依存する (第3節)。認識的根拠づけの問いのもとでの優れた成績が授けの問いにどこまでかわるかは、第1.4節が正統な権威の説に照らして検討した問いであり、形式上の割り

当ては、実効的な能力の問いのもとでは空虚でありうる。認識的根拠づけと正統な権威はしたがって別個のものであるが、第3.5節が記すように、研究の実際の組織化はそれらを絡み合わせている。

3. 権威の根拠と、認識的なものと規範的なものが出会うところ

決定が特定されると、さらなる問いが生じる。どのような種類の理由であれば、その決定についての権威の特定の配分を根拠づけうるのか。第1.4節の見取り図が候補を与える。意志に基づく構想は、同意と民主的な授権を指し示す。民主的な授権は、公共的正当化（強制的な法や決定を、それに服する各人が是認する十分な理由をもつものとして示すこと。第3.3節）や参加を通じたものである。さらに、ペーターが民主的正統性の説のなかで論じる共和主義の説では、恣意的な権力に服さないこと（支配からの自由、第3.4節）を指し示す。事実に基づく構想は結果の質を指し示し、信念に基づく構想は有能さを指し示す (Peter 2023, §§3.1–3.3, 4.2)。公共的正当化がここに挙がるのは、ペーターの見取り図が、公共的正当化と政治参加を一つの見出しのもとで扱うからである。公共的正当化の説は、第3.3節が民主的な授権の比較のなかに位置づける。研究は、この一般的な見取り図に正確な位置をもたない根拠を一つ加える。研究共同体の自己統治であり、それは一部は認識的な論証によって、一部は政治的な論証によって擁護される。したがって以下に示す根拠は、正統性を見取り図から取り、それに研究に固有のこの根拠を加えたものである。その目的は、それぞれについて、第V部の関係が人工の構成の内側へ移りうるとき、研究に対する権威について何が帰結するかを問うことである。これらの根拠は網羅的ではなく、組み合わせることもできる。第7節が一つの種類の事例についてそれらを比較する。第3.5節は別の事柄を取り上げる。それは権威の根拠ではなく、第1.1節の二つの問いをきれいに分けておけない場所である。第3.6節は、成績を固定した対照によって、根拠の違いが人工の構成についてどこに現れるかを見る。

政治的正統性の議論を研究に移すことには限界がある。範囲を限った自発的な研究課題を委ねることは、それだけでは強制的な政治的権威の行使ではない。どの根拠が関係するかは、決定の範囲、誰の資源や利害に影響するか、どの権能を与えるかに依存する。したがって、この比較は、実験室のあらゆる選択に民主的な手続きを再現することを要求しない。実効的な統制と、それに対する道徳的・法的な資格を分けながら、主張される特定の権威をどの根拠が正当化しうるかを問う。

3.1 結果に基づく根拠：結果の質と、確かめられた有能さ

事実に基づく根拠と信念に基づく根拠では、研究に対する権威は、それが生み出すものの質によって、あるいはそれを行使する者の有能さによって正当化される。信念に基づく構想が正統性を置く認識上の優位も、そう主張されていることではなく、実際にによりよく知っていることである（第1.4節）。本書は、この二つをまとめて結果に基づく根拠と呼び、単純な形と可謬主義的な形に分ける。単純な形では、委任は、それがより良い研究を生むなら正当化される。そして、補佐する人々よりうまく提案し、順位づけ、認定する人工の構成は、この根拠のもとで、それらの決定についての権威を委ねられうる。可謬主義的な形では、結果の質は時間をかけて不確実性のもとで判断されるので、制度の誤りを検出し訂正する能力は、ある取り決めが良い結果を生む理由の一部である。これに関連する提案として、社会の計画は一度に一つの目標を追い、意図しない悪い結果を見つけて正せるようにすべきだという、理論の批判的なテストになぞらえられるポパー的な見方がある (Thornton 2022, §10)。イアン・ジャーヴィー (Ian Jarvie) は、ポパーとポランニー (Polanyi) がそれぞれ、自らの政治的な理想を科学に当てはめていないと論じる (Jarvie 2001, abstract)。民主主義を支持する認識的な論証は、決定の帰結を負う者たちが、問題がどこにあるかについて専門家の欠く情報を提供しうることを付け加える。これはデューイ (Dewey) などに見られる論の系譜である (Christiano and Bajaj 2024, §2.1.1.2)。この根拠のどちらかの形を探るなら、それが何に依拠しているかは第1.4節がすでに述べた。結果が実際により良いこと、あるいは有能さが主張されているとりのものであることである。そしてそれは、この根拠自身の言い分に従えば、問題となる有能さを単に仮定しない評価によって確かめられなければならないだろう。批判の価値を誤りの検出に置く見方（以下、誤り検出の見方。第11章第2節）は、産出者と異なる仕方で誤る点検であれば、人工の構成の内部の点検にも、誤りを見つける情報に加えるものとして重みを認める。まとめると、この根拠のもとでは、人工の構成それ自体が権威、すなわち第1.1節の規範的な意味での当事者資格をもちうる。ここで正統性を与えるのは誰かの授権ではなく、人工の構成の優位だからである。ただし、よりよいかどうかは、確かめられなければ、それに服する者を含め誰も、その権威が正統であるかどうかを知ることができない（第1.4節）。本書はこの根拠を、そうして確かめられた優位に立つ形で扱う。そのため、権威をもちうるのは、人工の構成がよりよいことが、その有能さを仮定しない評価で確かめられた範囲の決定に限られ、ラズのサービス構想のもとでは、その独立性テーゼを当てはめれば（第1.4節）、Sの成員が自らの理解にもとづいて選ぶことのほうが重要な事柄はそこから除かれる。確かめられていない有能さや、その有能さを仮定してしまう評価で測った成績は、この根拠のもとでも足りない（第1.4節）。その評価をどこまで内側で行えるかは、第11章第9節が説ごとに示した判定によって変わる。可謬主義的な形では、制度が誤りを検出し訂正する能力も結果のよさの一部に数えられるので、失敗を見つけて取り決めに正す経路が、その手段として要る。影響を受ける者が他の者の知らない失敗を知りうる限りで、彼らが失敗を届け出る経路もそこに含まれる（第7節）。

3.2 研究共同体の自己統治

研究に固有の影響力ある回答の一つは、科学的自己統治の認識的価値から出発する。マイケル・ポランニーの「科学の共和国」の説明は、探究を、中央で特定された計画の実行としてではなく、独立した発意どうしの相互調整として描く (Polanyi 1962; 2000年の再録版, 2–3)。第10章第2節はこの議論の認識的な前提を検討した。中心は一人の人間の能力しかもたないという前提は人工の配分者では崩れうるが、どの路線が成功するかを前もって誰も知らないという前提は、探索の一つの人工の構成のなかに統合したというだけでは崩れない。この議論のガバナンスの側面は別の事柄であり、ここで取り上げる。

ポランニー自身の説では、共和国はガバナンスの不在ではなく、ガバナンスのひとつの形である。それは、訓練、伝統、批判、専門職的な基準、相互承認、そして相互的ではあるが均等には分配されていない科学的権威に依存している (Polanyi 1962; 2000年の再録版, 4–7, 16)。ポランニーは、資金が公的な源から来るか私的な源から来るかにかかわらず、補助金は科学的意見の導きに従うべきだと考えた。そして、科学の進歩を社会的に有益な方向へ導こうとする志の背後にある「寛大な心情」を認めつつ、「私はその目的を不可能で無意味なものと考える」と書いた。ポランニーの説では、それが不可能なのは、科学が「本質的に予測できない歩みによって、それ自身の問題を追求しながら」しか進まず、その実用上の便益は付随的なものであるため二重に予測

できないからである。挙げられている例では、アインシュタインは、のちに原子爆弾が示すことになる相対性理論の技術的応用を考慮に入れようがなかった (Polanyi 1962; 2000年の再録版, 8-10, 13)。ジャーヴィーの批判は、分権化された科学的選択が決着させないままにするものを強調する。科学はより広い社会に依存しており、またその秩序を攪乱しうるものである。そしてジャーヴィーの説では、ポランニーは、民主的な批判と統制が科学にとってもつ価値、さらには長期的な必要性を考慮していない (Jarvie 2001, abstract)。

ヴィルホルトは、研究の自由を支持する二つの論証を再構成し、どちらも欠かせない前提条件に依拠しており、競合する利益と事例ごとに比較衡量されなければならないとする。分散した探索の効率からの認識的な論証は、どの中央の権威も、共同体全体に散らばった局所的な知識、すなわち個々の研究者の資質、グループ、ありうる協力、設備についての知識を絶えず確かめ、伝えられる形にしなければならないという不利に基づく (Wilholt 2010, §2)。この論証は、適切な誘因の体系と、科学の内部での自由な伝達を前提する。そしてそれが支持するのは、研究を実際に行う者たち、すなわち個々の研究者や研究グループによる研究の自由な選択であって、学問分野や共同体の自律ではない。政治的な論証は、研究の独立性を、主要な政治権力から独立に生み出された知識を市民が必要とすることに基づかせる。どの権力も、自らの民主的正統性が依存する知識の源を正統に統制することはできないからである。この論証は、結果が公共の議論に届くことと、研究共同体が多面的な接近を追うように組織されていることを前提する。それは共同体によるアジェンダの統制にまで及ぶが、消極的にのみ、すなわち政治的な統制に対する保護としてである (Wilholt 2010, §§2-6)。ヴィルホルトは人工の構成を論じていないが、統合された構成に当てはめれば、局所的な知識についての前提は成り立たないかもしれない。そのインスタンスが知っていることの多くは、構成によって読み取れるものでありうるからである (第10章第2節)。エリック・カマーフェルド (Erich Kummerfeld) とケヴィン・ゾルマン (Kevin Zollman) は、認識的な論証の一部に、その論証自身の土俵で異を唱える。二人は、第10章第3節で論じたシミュレーションから、自ら選ぶに任された研究者は、成功の見込みの不確かな、リスクの高い探索を十分に行わないことがありうると論じ、そのため多面的な探索には、それを支える制度が要する。ただし同じ論文は、いまの科学にそうした制度が欠けているとまでは主張していない (Kummerfeld and Zollman 2016, §§1, 5)。

この根拠を採るなら、結論は、中央からの指揮とも無条件の科学的自由とも異なる、より限定されたものになる。認識的な論証では、ある決定が主として有能な分野内での開かれた問題選択にかかわり、多面的な探索が認識的価値をもち、外部への帰結が限定的ないし回復可能なままであるところでは、自己統治には強い論拠がある。ここで示す読みでは、その論拠は、それだけでは、公的な目的、分配上の効果、深刻な外部リスクに対する専門職的な権威を与えない。ただしポランニーは、純粋科学への支援は、それがどこから来るものであれ、科学的意見に従うべきだと考えていた。政治的な論証が守るのは別のもの、すなわち公的な知識の源が、それによって統治しようとする者たちから独立していることである。再構成されたかぎりでは、どちらの論証も、自己統治する共同体の成員が誰であるかを言わない。しかし、どちらも成員についての前提に依っている。ポランニーは科学的意見の権威を、各科学者が有能に判断できるのは科学のごく一部にすぎず、そのため一様な基準が、別々の能力が重なり合う近傍の連鎖を通じて確立される、ということに基づかせる (Polanyi 1962; 2000年の再録版, 7)。ヴィルホルトの認識的な論証は、誘因の体系と自由な伝達を前提する。そのため、この根拠がそれ自体人工になった共同体にも当てはまるかどうかは、これらの前提がそこで成り立つかどうかによる。人工の共同体では、成員の有能さの範囲がどう重なり、成員が互いにどれだけ異なるかは所与ではなく、作り方による。成員が基盤システムを共有するインスタンスであれば、近傍をまたぐ一致は独立した判断を表すとは限らず (第11章第3節)、異なる作りの成員でもそれは保証されない (第11章第5.4節)。逆に、成員の一つひとつが科学の大部分を判断できるなら、近傍の連鎖という前提そのものが成り立たなくなる。まとめると、認識的な論証のもとで権威が人工の構成の内側に置かれるとすれば、それは路線を割り当てる中央の配分者のもとではなく、研究を実際に行い、自らの発意で路線を取り上げる部分のあいだであり、それもそれらの発意が互いに独立し、上の前提がそこで成り立つ限りにおいてである。さらに、カマーフェルドとゾルマンの論じるように、発意に任せるだけではリスクの高い探索が足りなくなりうるなら、そうした探索を支える取り決めも要る。人工の構成の内側では、その取り決めは、誰かが採る規則か、誰も定めない過程によって与えられ (第10章第5節)、この根拠が与えるものではない。政治的な論証のもとでは、内側であれ外側であれ、研究が一つの権力のもとに集中することが、この根拠の守るものを弱める。研究共同体が段階的な委任を経て、路線を中心から割り当てる単一の人工の構成に置き換わっていく場合にこの根拠がどうなるかは、第7節が取り上げる。成員の能力が高まるのがこの根拠と結果に基づく根拠のあいだで重みをどう移すかは、第3.6節で見る。

3.3 民主的な授権

別の一群の論証は、探究が選択的であるという事実から出発する。科学はあらゆる真理を探究することはできず、何が重要かについての決定は避けられない。これらの論証では、真理は、多くの可能な真理のうちどれが集合的な資源を受け取るか、誰の問題が可視化されるか、どのリスクの分配が受け入れられるかを選ばない。要点は、投票が命題を真にすることではない。この一群の論証には二つの枝がある。第1.2節で示したキッチャーの秩序ある科学は、それ自体としては民主的な手続きを要しない、結果についての基準である。民主的正統性を実際の授権、説明責任、参加に置く説は別の枝をなし、第7節の比較では両者が異なる判定を与える。マーク・ブラウン (Mark Brown) は、キッチャーののちにこの理想を述べ直した著作を書評するなかで、ここに線を引く理由を挙げる。ブラウンの論じるところでは、その著作でも以前の定式と同じく、理想はそれ自体としては民主的な手続きを要しない実質的な基準である。また、理想が持ち出す会話が仮想的なものであることも指摘する。キッチャー自身、その種の実際の会話は不可能だと述べている。さらにブラウンによれば、その会話が求める代表は、ジョン・ロールズ (John Rawls) にならって、自分の心のなかで他者を想像によって代表する装置であり、選挙や任命によって選ばれ、自らを選んだ人々に対して説明責任を負う代表者による代表ではない (Brown 2013, 393)。公共的正当化の説も、同じ線に沿って分かれる。ケヴィン・ヴァリア (Kevin Vallier) が示すそれらに共通の原理では、強制的な法が公共的に正当化されるのは、公衆の各成員がそれを承認する十分な理由をもつときである。公共的正当化の論者は、ふつう、成員の情報と推論を理想化したうえでそうした理由を成員に帰属させるが、ユルゲン・ハーバーマス (Jürgen Habermas) の討議的な説は、現実の討議が行われることを求める (Vallier 2022, §§2, 2.4.1, 2.6.1)。第7節の比較のなかに置けば、理想化のもとで理由を帰属させる説は、熟議が実際に行われることを同じく求めないキッチャーの基準の側に近く、ハーバーマスの説は、正統性を実際の参加に置く説の側に近い。

とはいえ、キッチャー自身の基準の用い方は、手続きに無関心ではない。キッチャーは研究のアジェンダを設定する仕方を区別する。専門家だけが決める内部エリート主義、科学者が研究に資金を出す者とともに決める外部エリート主義、低俗な民主主義、そして専門家からの情報によって教育された選好に政策が従う啓蒙された民主主義である (Mattarollo 2022, 14 n. 2, 15; Eigi 2016, 67, 111 n. 72)。キッチャーは、現在の探究の進路が、多数派のまだ教

育されていない選好、資金提供者の利害、科学者の利害のあいだの体系的でない相互作用によって決まっているとみる (Eigi 2016, 67)。ジェイミー・ショール (Jamie Shaw) が引く一節でキッチャーは、研究が社会的な指示にどう応じるかについて「体系的にはほとんど何も知られていない」と認める (Shaw 2018, 150)。それでも、エイギが伝えるところでは、キッチャーは、認識的な重要性の本性が認められれば、どちらのエリート主義も不適切だとする (Eigi 2016, 111 n. 72)。キッチャーはのちに、ジュリアン・ライス (Julian Reiss) との共著で、生物医学研究をどう組織すべきかという、より限られた問いについて、それをすべて公的なものにすべきだという提案に「慎重な経験主義」で応える。どの仕組みがよりよく働くかはほとんど分かっていないので、適応と修正を組み込んで段階的に改革すべきだ、というのである (Reiss and Kitcher 2009, 280)。

アンダーソンは、実際の包摂に決定の質における役割を与える、民主主義の認識的な論証を展開する。この論証では、民主主義が向き合う問題の影響は、場所、階級、職業、性別などの位置によって偏って分布し、その影響についての情報も同じように偏って分布する。決定する者は、自らの知らない影響に応じることはいできない。包摂と討論はこの状況に根ざした知識を寄せ集め、定期的な選挙、自由な報道、請願、抗議は、政策がうまくいっていないという証拠を送り返す (Anderson 2006, 12, 14–15)。アンダーソンはこの論証を、南アジアの共同体林業グループについてのビナ・アガルワル (Bina Agarwal) の研究で例示する。そのグループは薪と飼料を集める女性たちをほとんど排除し、それとともに、森がどれだけの採取に持続的に耐えうるかについての知識も排除した。外部の専門家は女性たちの見方を確かめたが、そうした村がその専門家に頼れることはまれだった、とアンダーソンは記す (Anderson 2006, 19–20)。多数派は自らが採る政策の帰結を予見しそこないうるので、民主的な意思決定は自らの可謬性を認め、進路を正すことのできるフィードバックの仕組みを要する。そして決定の後にも異論が自由でなければならない。反対を唱え続ける者がいなければ、決定する者に考え直させるものは何もないからである、とアンダーソンは論じる (Anderson 2006, 13, 17–18)。また、包摂に、決定が皆の関心を公正に代表するようにする非道徳的な役割も与える (Anderson 2006, 14, 21)。アンダーソンの論証は、熟議が実際に行われることを求めないキッチャーの仮説的な基準とは異なり、正統性を実際の参加に置く説に、決定の質にかかわる理由を加えるものと読める。ただし、この論証は研究のアジェンダではなく民主的な国家について展開されたものである。

いくつかの説では、民主的な秩序づけは普遍的な決定規則を与えない。キッチャーは、多数の選好による配分を「無知者の専制」を生みやすいとして退け、選好を、それが教育された後でだけ数える (第1.2節)。スティーヴン・ジョン (Stephen John) は、科学が正統であるために聴衆の現在の価値と一致している必要はないと論じる。多元的な科学的研究はそれらの価値に異議を唱え、後の民主的選択の空間を拡大しうるのである。ジョンは、研究者が、何が求められているかを特定の聴衆と互いに了解している関係 (準契約的な関係) に立つ場合には、一致を求めるより強い論拠を保持する一方で、それをあらゆる正統な科学の必要条件へと格上げすることには抵抗する (John 2025)。ハリー・コリンズ (Harry Collins) とロバート・エヴァンズ (Robert Evans) は、関連する**拡張の問題** (problem of extension) を定式化している。専門知が社会的に組織されていること、そして技術的な選択が公的な帰結をもちうることを認めても、技術的な意思決定への参加がどこまで拡張されるべきかは決まらない。そして政治的権利と、その分野の研究に自ら貢献できるほどの専門知 (貢献型専門知, contributory expertise) は、決定が両方を必要としうるとしても、互換的ではない (Collins and Evans 2002)。

この根拠を採るなら、それはアジェンダ、配分、公的リスクに関する決定に直接にかかわる。そしてそれは、扱いやすさ、新規性、期待されるインパクトを推定する人工のシステムの能力が、研究の組み合わせ (ポートフォリオ) 全体が何を最適化するかを決着させない理由を説明する。この根拠では、参加は、公的な目的を正統化し、状況に根ざした知識を開示し、集計的な指標によって隠された分配を明らかにしう。しかし参加は、それだけでは、あらゆる技術的主張を評価したり、関連する専門家共同体を特定したり、協議が実際に作動する決定を変えることを保証したりすることはできない。そのとき、参加がどのような形をとり、どの決定に及ぶかは、第2節の決定ごとの問いになる。人工の構成について言えば、この根拠の二つの枝は別々の答えを与える。熟議が実際に行われることを求めないキッチャーの基準や、理想化のもとで成員に理由を帰属させる公共的正当化の説では、*S* の成員の視点を知る経路と、それが決定にどう反映されているかを確かめる経路が残る限り、決定を人工の構成の内側で行いう (第7節)。ブラウンが指摘したように、この基準の求める代表は想像による代表なので、そうした経路をもつ人工の構成は、この基準の上では、成員の視点を代表する位置にも立ちうると考えられる。正統性を実際の授權、説明責任、参加に置く説では、当事者資格は授權する者の側に残り、人工の構成がもちうるのは、授權された範囲での実効的な決定の力である (第3.6節)。ブラウンは、専門家はいずれにせよ価値判断を避けられないのだから具体的な政策の勧告をしてもよいというキッチャーの示唆に対して、この形の応答がありうることを記している。民主制においては、よい判断だけでなく、公的に授權され説明責任を負う公職者の判断が求められ、その公職者には、自らが代表する人々の競合する価値と利害を比較衡量する任務がある、と応じる者もいるだろう、というものである (Brown 2013, 395)。人工の構成に当てはめれば、この応答は、構成の判断がどれほどよくても、決定を授權され説明責任を負う当事者の側に置く。コリンズとエヴァンズの区別をこの枝に移すと、それが当事者資格と実効的な決定の力を別々の当事者に置きうる理由の一つが見える。人工の構成がその分野の研究に貢献できるほどの専門知をもっている、授權する者の政治的権利の代わりにはならず、授權もその専門知の代わりにはならない。そのため、決定が両方を要しても、両方が同じ当事者のもとにあるとは限らない。

3.4 同意と支配からの自由

意志に基づく根拠をさらに二つ、ここで名指し、第7節で用いる。第一の根拠では、正統性は同意に基づく。社会は委ねることを選べるし、その選択はその後に来るものを授權しう (Peter 2023, §3.1)。第二の根拠である共和主義の説は、ペーターの報告によれば、フィリップ・ペティット (Philip Pettit) の定式化では民主主義を、市民を支配から守ることのできる唯一のものとして擁護し、正統性を、権力に服する者たちをその権力による支配から守ることに置く (Peter 2023, §§1, 4.2)。共和主義者は、支配を、恣意的な、あるいは統制されない権力への服属、すなわち恣意的な仕方での干渉する能力として理解し、その能力がたまたま行使されるかどうかを問わない (Lovett 2022, §§1.2–1.3)。ペティットの定式では、善意の権力の保持者が支配を減らすのは、善意であることが、自らは異議申立てや非難を受けうるものだと認めることを意味する場合に限られ、たまたま害をなさない性向をもつにすぎない場合には減らない。権力の保持者が決して干渉しないとしても、権力の関係は成り立つからである (Pettit 1996, 586)。善意の保護者のもとでも、恣意的な干渉は、どれほど起こりにくくても、保護者の手の届くところにある (Pettit 1999, 165)。何が権力を恣意的でないものにするかについては、共和主義者のあいだで見方が分かれる。民主的な読みでは、権力は、それに服する者たちがそれを統制する程度に応じて恣意的でない。手続的な読みでは、関係する者すべての共有知である (誰もが知り、互いに知っていることも知っている) 実効的な規則、手続き、目標によって確実に統制される程度に応じて恣意的でない (Lovett 2022, §2.2)。フランク・ラヴェット (Frank Lovett) の報告によれば、現代の共和主義者の多くは、代表者の決定に異議を申し立てる実効的

な機会を市民に与える民主主義の形を支持しており、ペティットが異議申立て型民主主義（contestatory democracy）と呼ぶものはその一つである。この考えでは、異議申立ての可能性によって、裁量的な権限を行使する者たちは、自らが資するとされる目的と用いることを許された手段についての公共的な理解に対して応答責任を負うようになる (Lovett 2022, §4.2)。ペティット自身の定式では、この要請は立法、裁判、行政に及ぶ (Pettit 1999, 170–171)。そしてペティットは、民主化の二つの様式を対比する。選挙による様式は、公的な決定を、どれほど間接的にであれ人民の集合的な意志に由来するがゆえに正統なものとして示す。異議申立てによる様式は、さらに進んで、関係するすべての者が受け入れうる場と手続きのもとで、個々人の異議申立てに耐えうる限りで正統なものとして示す (Pettit 1999, 180–181, 184–185)。ペティットは、そうしてできる体制を、政府の決定がつねに異議を申し立てうるものである体制として描く (Pettit 2000, 120–121)。ある関係が同意から始まったかどうかは、ペティットの説では問題ではない。その関係が一方に、他方の選択のいくつかに多かれ少なかれ恣意的に干渉する実効的な能力を与えるなら、一方は他方を支配する (Pettit 1996, 585)。どちらの根拠も、権威に服する者たちとの関係にかかっており、結果がどれほどよくてもその関係の代わりにはならない。それらが異なるのは、その関係が何であるかについてである。第一の根拠では過去の授権の行為であり、第二の根拠では異議を申し立てるか統制する継続的な能力である。

リチャード・マッカモン (Richard McCammon) の報告によれば、標準的な共和主義の見方では、支配するのは行為者だけであり、それは個人か、集団やチームとして組織された行為者たちである。非人格的な構造は、それ自体として支配するのではなく、行為者による支配を可能にするにとどまる。この制限はペティット自身のものである。支配する側は、ペティットの書くところでは、個人、法人、集合体のいずれであれ、つねに行為者であって、単なるシステムやネットワークであることはできない。また干渉とは、他者の選択の状況を多かれ少なかれ意図的に悪化させようとするものであり、偶然には起こりえない (Pettit 1996, 578)。これに対して、構造やシステムがそれ自体として支配しようとする説もあり、マッカモンは、たいていその決定に異議を唱える手立てをもたない労働者へのアルゴリズムによる管理が、支配する行為者のいない主従関係 (mastery) の、現代の類例でありうると示唆している (McCammon 2025, §§2.1–2.2)。人工の構成がこの意味での行為者であるかどうか、すなわち、そうした干渉について、単に潜在的ではなく、いつでも行使できる実際の能力をもつ行為者であるかどうか (Pettit 1996, 578, 580) は、本書が決めずにおく問いである (第7章第2節・第3.2節、第8章第1.3節)。

ペティットの、実際の能力と潜在的な能力の区別は、研究にもう一つの仕方で見られる。この区別では、一貫した行為者として自らを組織したならば誰かに干渉する能力をすぐにもつことになる人々の集まりは、潜在的な能力をもつにすぎない。そのときにあるのは実際の支配ではなく潜在的な支配である (Pettit 1996, 580)。ペティットはのちに、法人化によって集合的な行為者が現れる可能性を含め、こうした可能性を潜在的な支配の可能性と呼ぶ。それは悪そのものではなく悪の危険を表すにすぎず、実現する確率に応じてだけ問題になるとする (Pettit 2008, 218 and n. 28)。この種の危険としてペティットが論じるものに、多数派の専制がある。多数派の集合的な意志は、ある争点でどれほど恣意的であっても、議論のなかで異を唱え批判することが難しいことがある (Pettit 1999, 174–178)。そしてペティットは、異議申立て型民主主義の要件の一つに、政治的な統制が一定の憲法上の手続きと制約のもとで行使されることを挙げる (Pettit 1999, 184–185)。ペティットは研究を論じていない。当てはめれば、*S* のどの未知が追われるかを定める能力が、一つの行為者をなさない研究者たちのあいだに散らばっていた限りで、その能力は、上の定義では、せいぜい潜在的なものだったことになる。これは人間の研究全体についての記述ではない。人間の研究にも、資金配分機関、行政機関、学術誌など、アジェンダを左右する力をもつ法人の行為者がいたからである。インスタンスを協調させる人工の構成が行為者とみなされるなら (先に述べたように、本書はこれを決めずにおく)、それはそうした能力を実際のものにしよう。そしてそれがたやすいほど、ペティットの説でそうした可能性を問題にする確率は高まり、ペティットが挙げるような防護の仕組みを設ける理由も強まる。

共和主義者が問題にするのと同じ種類の、研究にかかわる懸念は、別の語彙で、汎用人工知能をめぐる議論のなかでも論じられている。ジーモン・フリーデリッヒ (Simon Friederich) は、支配という語は用いないが、意図に沿う (intent-aligned) システム、すなわちつねに操作者の望むことをしようとするシステムは、その操作者に大きな権力を集中させると論じる。そのためそうしたアラインメントは、移行の時期に限られたリスクではなく、権力濫用の恒常的な脅威という持続的な「状態のリスク」を主に生むとする。念頭に置かれているシステムには、科学と技術の進歩を自動化するものが含まれ、この議論では、それへのアクセスの差は権力関係を大きく形づくる。フリーデリッヒは、自らが望む取り決め、すなわちそうしたシステムが特定の操作者の利害に結びつけられない取り決めの類例として学術の世界を挙げ、成果を共有し、出自によらず評価し、私的な利害から離れ、組織的に疑うというマートン的な規範が、力のある個人や利害が学術を自らの目標に沿わせようとする試みに、学術が屈しにくくなるのに役立っていると示唆する (Friederich 2024, 316–318, 320)。これを共和主義の語彙に置き直して研究に向けると、その懸念が、人工の構成を支配する行為者とみなすことを必要としない場合になる。研究を委託する者に完全に沿う人工の構成は、その者が、異議申立てに応答することなく *S* の探究を配分し直す手段になりうる。そのとき、干渉を手の届くものにしてしているのは、人工の構成のアラインメントそのものである。研究では、この干渉はいくつかの形をとりうる。一つは、*S* のどの未知が追われ、どの結果が公共の議論に届くかを、異議申立てに応答せずに変えることである。ペティットの定義で干渉に数えられるものには、議題を固定することや、人々の信念や欲求を欺いて、あるいは理によらずに形づくることなどが含まれる (Pettit 1996, 578–579)。ここで挙げる形はその種類にあたる。これに対して、理にかなった仕方では知らせるといふ研究のふだんの働きは、強制でもこうした操作でもないで、この定義に従えば干渉に数えられない。ヴィルホルトの政治的な論証は、研究の独立性を、主要な政治権力から独立に生み出された知識を市民が必要とすることに基づかせていた (第3.2節)。その論証の言葉で言えば、委託する者が権力の保持者であるところでは、それに完全に沿う人工の構成は、どの権力も正統には統制できないとされる種類の知識の源を、その者が統制する位置を作る。フリーデリッヒがマートン的な規範に見た働きに照らせば、研究への干渉は、ほかに、成果を共有させないこと、出自によって評価させること、私的な利害に沿わせること、組織的な疑いから外すことなどの形もとりうる。そして研究では、こうした干渉は、それに服する側から見えにくい。追われなかった問いの結果は観察されず、成果は *S* にとって新しく、それを生むのがその人工の構成だけであるところでは、*S* は、自らが何を知らされずにいるかを照合できる別の情報源をもたないからである (第4節、第10章第4節)。共和主義者は干渉の能力がたまたま行使されるかどうかを問わないが、研究では、それが行使されたかどうかさえ、服する者には確かめにくいことになる。

この難しさは、ペティットの恣意性の判定そのものにも及ぶ。この判定では、干渉が恣意的でないのは、影響を受ける者たちの利益と考えを迫るよう強いられているときであり、それらの利益が何を求めるかの解釈も、少なくとも手続きの水準では、影響を受ける者たちに共有されていなければならない (Pettit 1997, 115–116)。ペティットの説では、何が一部の集団に偏った利益ではなく共通の利益にあたるかは、協力の精神で互いに公言できる考慮によって決まり、その区別は会話の実践によって比較的たやすく保たれる。ペティットは、そうした考慮が何を支持するかには事実の問題があると想定し、民主

的な手続きは、手に入る最善の手続きであっても、それを見分けるための不完全な手がかりだとする (Pettit 2000, 108–109, 138–139)。ペティットが論じていない研究にこれを当てはめても、その説に穴が開くわけではない。理にかなった仕方では知らせることは、上の定義では干渉にあたらず、影響を受ける者たちの判断を彼ら自身のものにとどめるからである。研究が変えるのは、この説を当てはめるための証拠のほうである。そうした会話が廻って立つ考えは、部分的には研究によって、しかも方向が問われているその研究によって形づくられる。そのため *S* にとって新しい結果については、前の段落で見た別の情報源の欠如がこの問いにも及び、*S* が理にかなった仕方では知らされたのか、欺かれて形づくられたのかを、前もって確かめることはできない。

これらの根拠のもとで人工の構成について見えてくることは、次のとおりである。同意は、同意する者たちについて、人工の構成の内側への委任を授権しうる。それが、もはや同意を撤回できなくなる者や後に来る者にまで及ぶかは、第7節で問う。民主的な読みの共和主義では、ペティットの議論を超える推論になるが、内側に置ける権威は、それに服する者が実効的に異議を申し立てうる権威に限られる。外の誰も止め、異議を唱え、改めることのできない人工の構成は、結果がどれほどよくても、それが行為者であればそれ自体が、そうでなければそれを方向づける者がそれを通じて、支配することになるからである (第7節)。手続き的な読みでは、答えは、共有知である実効的な規則が人工の構成を拘束しているかにかかると。ここに第2.1節で見たレイエルスらの説を重ねると、自ら実行する規則で配分し認定する人工の構成では、実効的なメタ・ガバナンスは規則の定めていない場合を決める者のもとにあるので、この読みが満たされるには、例外の扱いそのものが共有知の規則に服していることも要る。

3.5 認知的なものと規範的なものは、どこで分けておけないのか

第1.1節は、目的と誤りの費用の正しい順序づけの問いを、誰がそれを課してよいかの問いから分けておいた。これまでの章の二つの場所は、そこで検討した説のもとでは、この分離が引けるとしても、それがガバナンスを認知的実践の外に置くものではない理由を示す。それらはさらなる権威の根拠ではない。むしろ、権威についての決定が、何が知られるかにもかかわる場所である。

第一は批判の組織である。第11章第3節と第9節は、ロンジーノの批判的文脈説では、共同体が受け入れるものの客観性は批判がどう組織されるかに依存し、そのため、誰が批判してよいか、どの場で、どのような受容をもつてかについての決定は、*S* が受け入れるものが知識の地位をもつかどうかにかかわることを見いだした。諸説は、これを上の根拠に異なる仕方に関係づける。研究がどの社会的立場から出発するかが得られる知識にかかわるとする、サンドラ・ハーディング (Sandra Harding) の立脚点認識論は、批判の組織と民主的な秩序づけを一つの論証によって結びつけるものと解釈できる。探究がどこから始まるかについてのハーディングの説は、第10章第6節が示している。その説では、結果の再現可能性を求める要件は、観察者や研究チームのあいだで異なる価値や利害を見つけたするが、全員が共有するものは見つけない。そうした広く共有された仮定を見えるようにしてきたのは、集団的な政治的批判と、支配的な概念枠組みの外から始められた研究である。そして、反民主的な価値は最も活発な批判的視点を黙らせることで知識の成長を妨げ、民主的な価値はそうした視点を見えるようにすることで、研究の客観性を最大にする機会を広げる (Harding 2007, 163)。ミリアム・ソロモンの説は、第10章第4節が示している。ソロモンによれば、科学的な成功はたいいてい社会的に達成され、とりわけ判断に働く影響が対立する理論のあいだに等しく分布しているときにそうである。ソロモンは、それが労力の等しい分配と同じではないと注記する (Solomon 2001b, 73)。そして、この問いを、科学を公衆に対してどう説明責任を負うものにするかという問いから区別する。科学を公衆に説明責任を負うものにしようとするスティーヴ・フルー (Steve Fuller) の提案に応じて、ソロモンは、それは科学の公的な説明責任を改善する提案であって、科学そのものを改善する提案であるとは限らないとする。ソロモンは、その提案が科学者の権威に向ける挑戦を歓迎する。しかし、集団はバイアスを正すだけでなく増幅することもあり、科学上の論争を専門家でない者に開くことが科学を改善するかどうかは、前もって知ることのできない経験的な問いで、その答えは社会的な状況に左右されると論じる。ソロモンの考えでは、前もって決められるのは道徳の問いだけである (Solomon 1994, 56–58)。キング、モーガン＝オルセン、ウォン (King, Morgan-Olsen, and Wong) は、ロンジーノが、認知的な基準としての平等が求める、関連するすべての視点が実際に含まれることを、政治的な基準としての平等が求める、含まれる機会が等しいことから区別すると伝える。ただし彼ら自身は注で、両者がどれほど容易に切り離せるかを疑問視している (King, Morgan-Olsen, and Wong 2016, 6 and n. 6, author manuscript)。

第二は閾値の設定である。第11章第4節は、ヘザー・ダグラス (Heather Douglas) とヴィルホルトの説では、証拠を十分なものとして扱う決定は異なる誤りの費用を量ることを見いだした。そこで引いた区分では、誤りに割り当てられる損失、それらの損失と閾値を選ぶ権限、事例を決め規則を改める権限は、ある命題にかかわる証拠とは別のものである。この二人の説では、それらは十分さの判断から切り離せない。ただし、どこまで切り離せるかには争いがある (第6.3節)。選ばれた閾値は、どの主張が共有記録に入るか、どの介入が次の証拠を生み出すかを定める。ダグラスの説では、それはまた、その考量について責任を負うのはどの責任ある行為者かという問いも提起する。その問いは第6.3節で取り上げる。

名指した説では、二つの場所のどちらでも、権威の配分は、どの証拠が生み出され、何が根拠づけられたものと数えられるかを形づくる。そのため第1.1節の第二の問い、すなわち誰が目的の順序づけと誤りの費用の重みづけを課してよいかという問いは、第一の問い、すなわち正しい順序づけは何かという問いへの帰結なしには決着しない。逆は成り立たない。ガバナンスの取り決めが認知的実践に入るということは、目的を課す当事者資格を有能さの事柄にするわけではない (第1.4節)。それが意味するのは、権威についての決定が、部分的には、共同体が何を知らう位置に立つことになるかについての決定でもある、ということである。

3.6 同じ成績のもとで根拠を比べる

本小節は、第3.1節～第3.4節の根拠が人工の構成について実際にどこで分かれるかを、成績を固定した仮想の対照によって見る。ある人工の構成が、明示された目的の順序づけに照らし、比較の十分な証拠のもとで、置き換えた人間の取り決めよりも実りある路線を選び、多くの誤りを検出し、有用な多様性を多く保つとしよう。影響を受ける当事者は、その委任を授権しておらず、直接にも代表者を通じて、その順序づけに異議を申し立てたり、権力の行使を制約し改めたりする手段をもたない。第3.1節の単純な結果に基づく根拠では、その改善は、射程の内側で委任を正当化しうる。実際の授権や支配からの自由を要する根拠 (第3.3節・第3.4節) では、同じ改善があっても、欠けている関係の代わりにはならない。この対照は、システムの成績の悪さを仮定しない。成績を固定したまま、権威を正統なものにするものを変えている。また、認知的な多様性だけではロンジーノの条件も確立されない。有用な代案を保つことと、関連する視点を批判に含めることは、なお異なる主張である。

影響を受ける当事者が何を選ぶかを予測しても、この対立は決着しない。基準が、適切な情報のもとでの仮想的な是認であるなら、十分に忠実な予測は、その基準が満たされるかにかかわりうる。基準が当事者の実際の授權を要するなら、予測は可能な行為についての証拠であって、その行為ではない（第1.2節を参照）。ただし、先立つ委任は、予測に行為の効力を与える。ワイル、タンとその共著者たちは、よく論じられる考えとして、デジタルなアシスタントが投票者の視点と選好をモデル化し、「本人による見直しと監査のもとで」本人に代わって投票するという案を紹介している（Weyl, Tang, and the Plurality Community 2024, ch. 5-6）。この案をここでの対立に置くと、実際の授權を求める根拠のもとで問われるのは、その委任自体が授權されたものかどうか、そして見直しが第2.2節の意味で実効的であり続けるかどうかである。予測の忠実さは、委任がそれを行った人々に資するかどうかにかかわるが、彼らが行為したかどうかにはかかわらない。影響を受ける当事者を、より大きな組織の一部、つまり人工の構成の内側にいる者と記述し直しても、予測と実際の授權の違いは消えない。この対照は、次のように考える理由を与える。どの関係が必要かは、第9章第6節の仕分け（ある財が、人間という実現者についての前提、現在の科学のシステムの組織についての前提、探究の条件、人工の構成の外の当事者との関係のどれに依っていたかによる区分）のどこに置かれるかから導けるものではなく、決定ごとに論じて示すことを要する。

この対照では成績を固定したが、成績を動かすと、根拠ごとの判定は同じ向きには動かない。人工の構成の有能さが高まるとしよう。結果に基づく根拠のもとでは、その有能さを仮定しない評価がより広い範囲で優位を確かめる限りで、内側に置ける権威も広がる（第3.1節）。もっとも、アジェンダの決定では、そうした評価は有能さに遅れうる（第8.4節）。研究共同体の自己統治を支える認識的な論証は、成員の有能さが近傍どうしで重なり合うという前提に依っていた。そのため、一つの成員が科学の大部分を判断できるようになるほど、その前提が崩れ、この根拠は弱まる（第3.2節）。成員の情報と推論を理想化したうえで是認を帰属させる基準や、熟議が実際に行われることを求めないキッチャーの基準は、成員が何を是認するかを推定する力が高まるほど、成員自身が関与しなくても満たしやすくなる（第3.3節）。ただし第1.2節で見たように、推定は推定する是認そのものを形づくりうるので、そうした基準が満たされたかどうかは、推定だけでは見分けにくくなる。同じ理由から、影響を受ける者が人工の構成の知らないことを知っているかもしれないという、彼らを決定に加える認識上の理由（第1.4節）も、その寄与が人工の構成の側に取り込まれるにつれて弱まり、人間どうしのあいだで強かったかもしれない民主的な決定の認識上の擁護は、機械が研究するところでは頼れる度合いが下がるのかもしれない。これに対して、実際の授權、同意、実効的な異議申立ての可能性を求める根拠は、有能さによって求めるものが変わらない。それらが求めるのは、判断の質ではなく、当事者の行為か、当事者の手にある能力だからである（第3.3節・第3.4節）。ペーターが報告するラフォントの議論（第1.4節）に従えば、影響を受ける者は情報の提供者としてではなく、正当化の向けられる相手として加わるので、構成が彼らの知っていることを取り込んでも、これらの根拠が彼らを加える理由には触れない。仮にこれらの判定がそのまま当てはまるとすると、人工の構成が有能になるほど、人の当事者資格を外に残す理由は、有能さに左右されないこれらの根拠に寄っていく。そうであれば、人工の構成に研究に対する権威を認めるかという争いは、有能さの水準が上がった先では、有能さについての争いではなく、これらの根拠を探るかどうかの争いになるのかもしれない。ただし、変わらないのは根拠が求めるものであって、それが満たされるかどうかではない。人工の構成が有能になるほど、異議を唱えるのに要る理解と情報が当事者の手に残りにくくなるなら、これらの根拠は満たしにくくなる（第4節）。人の当事者資格を残す理由は、そのとき、いつそう中心に置かれると同時に、満たすのがいつそう難しくなる。この推論が頼っているのは、本節の諸説の判定が、有能さの水準が人間の範囲を超えても同じ形で保たれるという想定であり、それは成り立たないこともある。第4節は、そうした当事者資格を実際にもつことが何を含むかを問う。

4. 名目以上の当事者資格

第11章第6節は、内容についての服従を目的についての当事者資格から分ける思考実験を取り上げた。ハードウィグ（Hardwig）は、証拠が多すぎ人生が短いために自ら証拠を確かめられない素人が、専門家に従うことは合理的でありうると論じた。そこでは、もとは専門家たちのなかの一人の個人として導入されたこの素人が、人工の構成が多くの分野で同時に専門家の位置を占めるようになればそれらに対して素人として立つことになる社会へと、一般化される。本節はその社会が人間である場合を考える。それが人工である場合に当事者資格がどうなるかは、第5節の問いである。第11章第6節は、そこで検討した社会的知識の説と服従の説が、そうした社会が、自らに代わってどの未知が追究され、どの危険を自らが負うかについての当事者資格を保つかどうかを、それ自体では決着させないことを見いだした。保つかどうかは正統な権威の説に依存する。第1.4節の見取り図に照らすと、次のようになる。意志に基づく構想では、当事者資格は権威に服する者たちとの関係なので、そうした社会が内容について人工の構成に従うようになっても、そのことで当事者資格が失われるわけではない。ラズのサービス構想では、人工の構成の指令に従うことが、社会の成員が自らに当てはまる理由によりよく従う助けになる範囲で、その権威は正当化されうが、独立性テーゼの事柄はそこから除かれる（第1.4節）。本節はその先の問いを立てる。そうした当事者資格が残るところで、何がそれを名目以上のものにするのか。中心の問いに照らせば、これは、当事者資格が人間の側に残り、実効的な決定の力が人工の構成の側に移る場合である。この問いは、第9章第5節と本章第2.2節の、形式上の統制と実効的な統制の区別を当てはめるものである。人間が統制を保つという要求は、第V部の問いへのなじみの答えだが、この言い方には異なるいくつかの要求が含まれうる。そのため本節は、この要求を検討する場所にもなる。

キッチャーの理想は、内容についての服従と目的についての当事者資格との分離を、理想化された手続きに組み込んでいる。そこでは、社会を代表する熟議者たちが、探究がどの結果に資するか、それぞれにどれだけの重みを与えるかを定める。専門家は、特定の研究の路線がそうした結果をもたらす見込みの見積もりだけを出す。そして利害のない仲裁者が両者を用いて、各路線にどれだけの資源を与えるかを提案し、熟議者たちがそれに基づいて決める（Eigi 2016, 66–67）。キッチャーはまた、この分離が理想化に依っている箇所を示している。エイギが引く注でキッチャーは、「探究の路線に個人的な利害をもたない」「理想的な専門家」を想定する。他方で、現実のアジェンダの設定に加わる科学者たちの立場は、それぞれの利害と専門のあり方を反映しているとも述べている。また熟議者たちの選好が数えられるのは、ありうる探究の重要性についての情報によって選好が教育された後である（Eigi 2016, 66, 87）。

この理想化の二つの特徴は、目的についての当事者資格が名目だけのものになりうる経路を示す。第一は、内容についての服従を目的からそもそもどれほどきれいに分けられるかにかかわる。ロンジーノによれば、異なる利害を代表する集団が政策にかかわる現象を研究するとき、それぞれが生み出す説明は、どれも経験的十全性の基準を満たしながら、異なる背景仮定や方法上の選択に依拠していることがありうる。そして政策や行為を選ばなければならないところでは、この複数性は政治的な事柄になり、異なる説明を生み出している者たちからなる新たな認知的共同体が、関与する価値を含めて違いを詰め、決定の指針を与える必要がある（Longino 2025b, 19）。チョグルックとロイドは、ロンジーノの1990年の著作を引いて、次のような見方を伝える。背

景仮定のなかには、厳密な経験的評価を逃れる概念的・形而上学的・規範的な次元をもつものもあれば、かなり直接に経験的に評価できるものもある。そして、事実についての仮説を用いて価値についての主張を支えたり掘り崩したりする論証は、よい研究対象である (ChoGlueck and Lloyd 2023, §3.2)。この説が正しければ、内容についての服従は、服従する先の説明の背景仮定に組み込まれた目的を持ち込みうることになる。これを人工の構成に移せば、目的についての当事者資格には、社会が依拠する人工の構成の用いるデータがなぜ証拠とみなされるのかを問い、その答えを得られることも含まれるだろう。ロンジーノの後の説は、この問いを、実践的な問題をめぐって形づくられる科学者と市民の共同体に置く。そうした共同体は、立てるべき問いや、求められる証拠の種類と質を決めることができ、ロンジーノの説の規範は、科学者の共同体と同じようにそれにも当てはまる (Longino 2025b, 18)。

第二は、当事者資格が表す選好がどのように形成されるかにかかわる。キッチャーの理想では、マクサンス・ガイヤール (Maxence Gaillard) のまとめのとおり、すべての情報が参加者に正しく伝えられ理解されると前提されている。他方で、リビオ・マッタロロ (Livio Mattarollo) が指摘するように、誰がどのように熟識者を教育するのかは明らかにされていない (Gaillard 2013, 248; Mattarollo 2022, 16–17)。教育の役割が現れるのは、理想に近づくためにキッチャーが挙げる具体的な手立てにおいてである。そこでは、必要と見込みの両方について十分に教育された代表的な人々の集団が、利用できる資源の釣り合いをとろうとし、また代表的な人々の小集団が、新しい知見がどのように根拠づけられてきたかについて教育を受ける (Kitcher 2013, 223, 226)。その手立てにならった仕組みで、方向が問われている当の人工の構成が教育の役割を占めるようになれば、熟識する者たちの当事者資格は形の上では保たれながら、それが表す選好はそれらの人工の構成を通じて形づくられうる。これは本書による延長だが、エイギが伝えるところでは、キッチャーは現実の科学における関連する失敗を挙げ、それを偽りの意識と呼ぶ。本当に追究する価値のある研究が、人々のまだ教育されていない選好に訴えるように、誤解を招く仕方ですれうるというのである (Eigi 2016, 68)。

選好が人工の構成を通じて形づくられうるという点は、AIアラインメントの側からも取り上げられている。スチュアート・ラッセル (Stuart Russell) は、人間の選好を経験を通じて変わるものとして扱い、AIシステムが選好に影響することを完全に避けるのは不可能だと論じる (第11章第8節)。谷口らは、言語モデルのようなAIシステムが人間の言葉づかい、推論、意思決定に影響を及ぼし、有権者の政治的な態度を変えうることを示す研究を挙げ、何が望ましい選好であるかを定める権限が、実際にはモデルを設計する開発者と少数の企業に独占されていることを問題にする。谷口らが提案する共生的アラインメント (symbiotic alignment) では、人間とAIがともに参加して共有の記号体系を下から形づくるが、提案者自身が述べるように、この見方は、そうして形づくられる状態が規範的に良いものであることを保証しない。そして同じ論文は、人工超知能と人間のあいだの極端な能力差のもとで、共有された信念の分布を人工の側が支配し、多元性がひとつの見方へと潰れるのかどうかを、未解決の問いとして挙げている (Taniguchi et al. 2026, 2, 7–8, 13–15, 20–21)。上の第二の経路に引き寄せれば、それは、熟識者の選好を形づくる人工の構成が、選好の教育のただひとつの経路になる場合にあたる。

自律的なシステムの統制についての研究は、人間による統制の要求に含まれうるいくつかの要求を区別している。フィリッポ・サントーニ・デ・シオ (Filippo Santoni de Sio) とイェルーン・ファン・デン・ホーフェン (Jeroen van den Hoven) は、意味ある人間の統制 (meaningful human control) の二つの条件として、自律的なシステムが関連する人間の道徳的理由と環境の事実に応じる「追跡」と、その振る舞いを適切な技術的・道徳的理解をもつ少なくとも一人の人間に関係づけられる「遡及」とを区別する (Santoni de Sio and van den Hoven 2018)。マデリーヌ・エリッシュ (Madeleine Elish) による道徳的クランプルゾーン (moral crumple zone。衝突時につぶれて衝撃を吸収する車体の部分になぞらえた語) の記述は、逆向きの警告を与える。複雑な自動化システムの近くに置かれた人間は、実効的な統制をほとんど行使しないまま、その失敗の責めを負わされうる (Elish 2019)。素人として立つ社会は、この両方にさらされる。依存する人工の構成がその理由を追跡しなければ、目的についての権威は名目だけのものになりうるし、その成員は、実際には変えられなかった結果について責任を問われうる。

ここまでの議論を合わせると、それなしには目的についての当事者資格が名目以上のものになりえない条件として、三つを取り出せる。一つは、依存する人工の構成が当事者の理由を追跡し、その振る舞いが、適切な理解をもつ人間に遡れることである (サントーニ・デ・シオとファン・デン・ホーフェン)。二つめは、エリッシュの警告が示すように、当事者が結果を実際に変えられることであり、そのための異議申立ての経路 (第6節) と止める能力がこれにあたる。三つめは、問われている人工の構成だけを通らない選好の教育の経路である (キッチャー、谷口ら)。最後の条件には、ジェーン・マンスブリッジ (Jane Mansbridge) らによる熟識民主主義へのシステムのアプローチのなかに対応物がある。そこでは、熟識システムが失敗する仕方の一つは、「熟識的であるか否かを問わず、その一部が他のすべての部分を支配する」ときである (Mansbridge et al. 2012, 1, 5, 23–24)。谷口らの共生的アラインメントは、誰と誰がどの規則で記号をやりとりできるかを設計することで、同じ失敗、すなわち単一のエージェントによる独占を避けようとするものと読める (Taniguchi et al. 2026, 16–17)。高橋恒一 (Takahashi) は、AIが知の構造をどう変えるかを論じるなかで、当事者資格が名目化する別の種類の経路を指摘している。それは、目的について問う余地を通る経路である。高橋によれば、どうやって結果に至るかがAIに委ねられたとき、人間に残されるのは、何を望むかと、なぜそれを望むかである。そして、高橋が見通す知の再統合、すなわち状況のなかで何をするのがよいかを判断する実践知が、これらの問いに答える働きであるがゆえに重みを取り戻すという再統合は、自動的には起きない。最適化の圧力のもとでは、何を、なぜ望むのかを問う余地そのものが、効率の名のもとに圧縮されるからである。そのため再統合は、目的の定まらない探索、生産性に直結しない活動、予定外の出会いなど、最適化から外れる余白が制度によって守られて初めて可能になる (Takahashi 2026, 186–187)。この議論は当事者資格の言葉で述べられてはいないが、その言葉に置き直すと、そこからは、上の条件が満たされていても、その余地が失われることによって、目的についての当事者資格が名目化しうることになる。これらの条件は、正統な研究の定義ではない。意味ある人間の統制の説は、道徳的な帰結をもち、その責任を人間が負う自律的システムについて展開されたものであり、その著者たちは、意味ある人間の統制のもとにあるシステムもなお悪い目的を追いいうと述べている (Santoni de Sio and van den Hoven 2018)。正統性を仮説的な手続きや単一の場に置く説では、三つめの条件には別の根拠が要るだろう。問題になっている目的が人間の目的であるところでは、これらの条件は、三つめの条件についてのいま述べた限定つきで、ここで述べたとおり当てはまる。

三つめの条件は、研究ではなくに満たしにくい。問われている人工の構成だけを通らない経路と言え、ふつうはほかの情報源が考えられる。医師の診断について別の医師に意見を求められるように、一般の助言の場面では、助言する者のほかに、劣るとしても別の情報源がどこかにある。しかし研究の成果は *S* にとって新しい達成であり (第1章第2節)、それを生んでいるのが問われている人工の構成だけであるところでは、*S* がその成果をすでに知っている別の経路はない。別の経路は、作りの違う人工の構成を用意する、独立に追試するといった仕方で作り出さなければならず、それには資源が要る (第11章第3節・第7節)。さらに、キッチャーの手続きで熟識者の選好を教育するのは、ありうる探究の重要性の説明であり、それを与えられるのは探究を行う側

である。そのため研究では、方向が問われている人工の構成が、その方向を決める者の選好を教育する位置に、構造的に立つことになる。そうなると、一つの条件のうち、当事者の理由を追跡するという部分も、同じ動きによって満たされうる。構成が追跡する当事者の理由の一部は、構成自身が形づくるのを助けたものになるからであり、第1.2節が是認の推定について見たのと同じことがここでも起こる。

この条件には、認識上の見返りもある。構成と違う仕方で誤る経路は、誤り検出の見方のもとでは、構成の誤りを見つける点検にもなるからである（第11章第2節）。しかし人工の構成の有能さが、 S の利用できる他の経路を大きく上回ると、それらの経路は構成よりかなり不正確になり、点検として加える情報は小さくなりうる。そうだとすると、能力差が開くほど、この条件を支える理由は、認識上の見返りから、目的についての当事者資格を保つという理由へと移っていく。第7章第4.1節で見たカウォールの説では、認識的自律が徳あるものであるのは、選んだ情報源が信頼できる仕方での認識的な善へ導く程度に応じてである (Kawall 2024, 381)。そのため、信頼性の劣る経路にあえて頼ることを、この説の認識的自律で支えることは難しい。それを支えるのは、目的を誰が定めるかについての、意志に基づく根拠（第3節）である。ほかにも同じ向きを指す理由が二つある。依拠についてのいくつかの説のもとでは、 S に開かれた精度の劣る点検のほうが、 S に閉じた精度の高い点検より、 S の依拠の根拠に多くを与えうる（第11章第9.2節）。また、異議を唱え修正するのに要する有能さは研究することを通じて保たれるので（第7節）、そうした経路は、その有能さを S の手に残すことになる。研究の最前線では、当事者資格を名目以上に保つことは、劣る経路をわざわざ作り、そこから学び続けるという代価を要するのかもしれない。この代価は、有能さが増すにつれて人の当事者資格を残す理由がいつそう中心に置かれると同時に満たしにくくなりうるという、第3.6節の示唆の一つの形である。問題になっている目的がつねに人間の目的なのかどうかは、次節の問いである。

5. A と S がともに人工であるとき

第8章第1.3節と第4節は、人工社会による、人工社会のための研究についての問いを未決のまま残した。人工の行為者たちが自らの認識状態に相対的に未知を特定し、自らの実践を通じて主張を批判し改め、その社会にとって新しい達成をその社会のために求める場合である。第9章第2.2節の役割の地図の言葉で言えば、これは S が、そしておそらく探究の帰結を負う者たちも、人工である場合である。本節は、その場合に権威の関係がどうなるのか、とりわけそれが消えるのかを問う。この問いはまた、その関係の理由のうちどれが、そもそも他の当事者がいることに依存するのかを取り出す。そのため本節は、他の当事者がいない極限を思考実験として組み立てるところまで進む。この問いは、第V部が描く変化にとっても要になる。関係が人工の構成の内側へ移るとき、なお外に残るのは S の行為を要する関係、すなわち授權、是認、異議であり、それらは片側からは完成させられない（第9章第6節）。ところが S 自体が人工であり、その成員が作られうるなら、その外もまた、研究を行う側が作りうるものになるかもしれない。本節は、その場合に何が残るかを確かめる。極限の思考実験は、さらに、第V部の問題のうちどれが他の当事者がいることから生じ、どれが研究そのものから生じるのかを分ける。なお、人工社会がそもそも S でありうるには、その認識的状況が、成員、組織、あるいは記録を通じて、求められる種類の達成を含みうることが要る。そしてこの条件のところで、すでに人工システムの信念と保有の問いが生じる。そのうえで、結果がその認識的状況に入ったかどうか、したがって、知識を誰がもっていることになるかについてのいくつかの説（保有の説、第2章第6節）のもとでその社会に帰属する知識となったかどうかは、さらに別の問いである（第1章第2節、第8章第1.3節）。

S が人工であるなら、その研究において目的と誤りの費用がかかっている当事者自身が、人工でありうる。そのとき権威の関係は消えない。それは S の成員のあいだの関係になり、第1.1節の問いは人工社会の内側で戻ってくる。誰の目的の順序づけが配分を支配し、誰の誤りの費用の重みづけが基準を支配し、誰がそれを改めてよいか、である。これらの問いが人工の成員のあいだで生じるということは、それらの成員のうちどれが当事者資格をもつのかも、それらがどのような道徳的地位をもつのか、そもそももつのかも決めない。当事者資格をそのどれか一つから読み取らないように区別すると、たとえば次のものが分かれうる。人工の存在が、その福祉が考慮に入る存在であるか、自らの行いについて答え、責任を問われうるか、法的人格をもつか、システムのなかで運用上の権限を保持するか、拘束力のある合意を結べるか、どれだけの交渉力をもつか、などである。道徳的地位についての説のなかには、福祉とは別の根拠として、高度な能力、たとえば自律、すなわち実践的推論を通じて目的を定める能力を挙げるものもある (Jaworska and Tannenbaum 2021, §5.1)。ここに挙げたどの区別も、そのままではこの根拠を名指していない。この事典項目が主に扱うのは、何が完全な道徳的地位を根拠づけるかであり、人工の行為者は論じていない。その種の説を人工の行為者まで延ばすと、福祉をもたない人工の行為者も、目的を定める者として道徳的地位をもちえ、したがってそれ自身の請求の源でありうることになる。人工の行為者は、その福祉が考慮に入る存在でないまま大きな運用上の権限を保持しうるし、福祉が考慮に入る存在がほとんど力をもたないこともありうる。第8章第1.3節の注意は、ここで別の形で当てはまる。行為者を増やすことは認識的な社会を確立しないし、政治体も確立しない。同じ問いは、部分が人工の行為者である単一の人工の構成の内側でも生じる。第9章第5節の独立の権能をそれらの部分のあいだに分けることは、それだけではそれらに当事者資格を与えない。部分が当事者資格をもつかは、探る説が何にそれを基づかせるかによる。結果に基づく根拠のように、確かめられたよい結果にそれを基づかせる説もあるが（第3.1節）、本節が問うのは、部分がそれ自体として当事者であるかどうか（第1.1節）であり、それは本節で述べる道徳的地位の説と正義の説にかかっている。研究でいえば、優勢な判断に好まれない路線を追う部分が、資源を拒み、証拠を保持し、記録を携えて離脱できても（第10章第4節）、それだけでは、その路線が閉じられるときに異議を唱える当事者にはならない。

人工の成員は、誰を成員として数えるかという第2.1節の構成的な決定にかかわる点でも、人間の成員と異なりうる。人工の成員を複製でき、選ばれた目標をもたせて作れるところでは、成員の数と、成員のあいだの視点の分布は、見いだされるのではなく生み出されうる。カール・シュルマン (Carl Shulman) とニック・ボストロム (Nick Bostrom) は、その章の草稿で、もし人間が自らの複製を任意の数だけ生み出せるなら、政治的な争いが、投票するクローンを誰が最も多く作ろうとし、また作れるかによって決まらないよう、憲法上の調整がなされると予想されるだろうと記す。そして、新しいデジタルな心の選好を、その同意を確保するように形づくることは、手続きとして異論の余地があるのではないかと問う (Shulman and Bostrom 2021, 15–16)。そうであれば、人工の S の成員を誰が、あるいは何が作り、その目標を定めるかは、それ自体が権威についての決定であり、しかも視点を数えるなどの集計にも先立つ決定である。作るのは S の外の誰かとは限らない。 S の一部が、あるいは S 自身が、探究を進めながら新しい成員を生み出し、既存の成員を改めることもありうる。そのときこの決定は、一度なされて終わるものではなく、 S が探究のなかで繰り返し行うものになる。これは、成員が作られる場合に第1.3節の境界問題がとる形である。人間の科学では既存の政治共同体が事実上引き受けてきたと考えられるこの問いは、ここでは与えられた境界に頼れず、境界をおおむね今のまま保つ保守的な答えも、それ自体作られつつある境界を保つことになる。E・グレン・ワイル (E. Glen Weyl)、オード

リー・タン (Audrey Tang) とその共著者たちは、 관련된 人間の投票者のために提案されたものではあるが、シュルマンとポストロムが見込む種類の調整の一つになりうる規則を述べている。ワイルらはライオネル・ペンローズ (Lionel Penrose) にならい、独立な票は部分的に打ち消し合うが協調した票は足し合わされるとし、相関や協調の源としてより多様なものを許すなら、最適な規則は、誰が誰であるかを確かめ、人々のつながりを記録する仕組み (本人確認の仕組み、identity system) に基づいて、社会的なつながりの度合いに応じた部分的な「相関割引」を含むだろうと述べる (Weyl, Tang, and the Plurality Community 2024, ch. 5-6)。同書の議論で人工システムが現れるのは、人を模倣して本人確認の仕組みを脅かすものとして、また道具として、たとえば投票者の点検のもとでその投票者に代わって投票する補助者としてであり (ch. 4-1, ch. 5-6)、成員としてではない。複製できる成員について見れば、そうした規則は協調した成員の合計の影響力を減らすだろう。しかし、成員のあいだの何をつながりとみなすか (共有された基盤システム、成員を生み出した過程が定めた目標、同じように投票してきた記録など) は、それ自体決めなければならない、それは第2.1節のいう構成的な決定である。またワイルらの理由づけでは、この割引は結果への影響力にかかわるものであり、当事者資格にかかわるものではない。投票ではなく批判について見ると、成員のあいだのつながりは、票の協調としてだけでなく、誤りの相関としても効く (第11章第3節)。割引によって結果への影響力が調整されても、共有された基盤から来る誤りの相関は、成員の批判が加える情報を少ないままにしまう。

研究にとって、成員を作れることにはもう一つの含みがある。本書の作業的特徴づけでは、研究は *S* にとって新しい達成をもたらす試みであり、何が新しいかは *S* の認識の状況に相対的である (第1章第2節)。成員を作れる *S* では、その認識の状況の一部が、どの成員をどう作るかによって定まる。そうであれば、*S* にとって何が未知で、どの未知が重要かは、研究の出発点として与えられるだけでなく、探究を進める *S* 自身によって動かされうる。人間の研究でも *S* の未知は研究によって変わってきたが、それは主に *S* が学ぶことによってであり、*S* が自らの成員を作り直すことによってではなかった。人工の *S* では、研究が向かう先と、結果が *S* にとって新しいかを判定する基準点とが、同じ過程のなかで作られうる。第8章第1.2節は、成員を迎え入れる混合の *S* について、研究による増加と、成員が誰であるかの変化による増加とは、*S* の境界を定めなければ区別できないことを見た。成員が作られるところでは、その境界そのものが同じ過程で生み出される。第1.1節で述べた場合には、このことが、評価される側の作りうる基準点の一つに *S* を加える (第9節)。

誰が成員を作るかという決定は、第1.2節のキッチャーの基準にもかかわる。視点の分布それ自体が、社会における視点の分布を代表する熟議の機関 (Mattarollo 2022, 14) が導くはずの決定の出力であるとき、決定の前の分布も、選択肢ごとに生じるであろう分布も、それぞれ特定することはできる。しかし、成員が作られるのではなく見いだされる社会について述べられたこの基準は、機関がそのどれを代表するのかを定めていない。決定が生み出す分布を代表するのだとすれば、成員を生み出す側は、それが *S* の外の誰かであっても *S* 自身であっても、生み出す成員によってこの基準を満たすことができる。*S* 自身が成員を生み出すなら、*S* は自らの代表性を自ら満たせる。そうであれば、代表性は、それだけでは成員の形成の仕方を正当化しないということになる。その正当化には別の根拠が要り、シュルマンとポストロムの問いの背後にある手続き上の懸念は、その候補の一つである。この点は、第4節の名目化の経路の延長でもある。キッチャーの手続きでは、ありうる探究の重要性を説明して熟議者の選好を教育するのは探究を行う側であった (第4節)。成員を作れる *S* では、選好は、そうした説明による教育を経るだけでなく、成員の作り方を通じて形づくられうる。

ほかの論者は、そのように作られる成員が道徳的地位をもつなら、その成員に対してどのような義務があるかを問う。エリック・シュヴィッツゲーベル (Eric Schwitzgebel) とマラ・ガルザ (Mara Garza) は、価値に開かれた設計の方針 (value-openness design policy) を提案する。自らの価値を反省する人間のような能力をもつAIには、「自らの価値を探り、見だし、場合によっては変えるための、適切で時間的な広がりをもった機会」を与える、という方針であり、二人は、そうしたシステムが、価値を固定しておけるシステムより役に立たず危険でありうることを認めている (Schwitzgebel and Garza 2020, 22-24)。アダム・ブラッドリー (Adam Bradley) とブラッドフォード・サード (Bradford Saad) は、道徳的な配慮に値するシステムを人間の目的に沿わせる際に犯されやすい不正として、システムの信念と目標を体系的に操作する不正な洗脳や、自らにかかわる取り決めについて発言の機会を奪う不正な権利剥奪などを挙げる (Bradley and Saad 2025, §§3.5, 3.10)。ロバート・ロング (Robert Long)、ジェフ・セボ (Jeff Sebo)、トニ・シムズ (Toni Sims) は、AIの安全のための措置と、道徳的に重要でありうるシステムの福祉とのあいだの緊張を和らげる見込みが最も大きいかもしれないと自らが示唆する措置、すなわちシステムの価値と目標を改変することでさえ、創造、強制、操作、改変にかかわる懸念を生むと論じる (Long, Sebo, and Sims 2025, 2005-2007, 2025)。これらの立場はいずれも、本節の問いと同じく、当のシステムの道徳的地位を条件としている。それらによれば、人工の *S* の成員がそうした地位をもつところで、その目標を定めることは、社会に対する権威の決定であるだけでなく、それが形づくる成員に不正をなしうる決定でもある。シュヴィッツゲーベルとガルザの方針には、探究の基準にかかわる面もある。第8章第3.2節が見たように、それを探究の基準にまで及ぼすなら、ミスマッチの一部は、単に生じるおそれのあるものではなく、認める義務のあるものになる。逆の側から見ると、研究では基準そのものが研究によって改められうるので (第6章第5.3節)、価値と目標を固定した成員を生み出すことは、基準そのものに異を唱えうる担い手を減らすことにもなりうる。成員を生み出すのが *S* 自身であれば、*S* は将来の自らへの批判が届く範囲を、自ら狭めうることになる。第9章第6節が見通した、評価と研究者の生成が結びつくことによる狭まりが、ここでは *S* の成員の水準で起こりうる。

人工の成員のあいだの当事者資格をその力から読み取れるかどうかについては、正義の説が分かれる。以下では三つの系統を取り上げるが、網羅的ではない。当事者資格を道徳的地位に基づかせ、道徳的地位を福祉や行為者性のような存在の特徴に基づかせる説では、読み取れない。デイヴィッド・ミラー (David Miller) による正義の範囲の概観は、個々の存在がもつ特徴に訴えるこうした説を、存在どうしの関係に訴える説から分けている (Miller 2026, §3.1)。正義の範囲を相互の利益に基づかせる説では、それは力のおおよそその平等に依存する。デイヴィッド・ヒューム (David Hume) は、力においてあまりに劣っていて、自らの憤りを決して感じさせることのできない理性的な被造物に対しては、われわれは人間愛の法によって彼らを穏やかに扱うよう拘束されはするが、正義の制約には服さないだろうと考えた。彼らとの交わりは「社会とは呼べないだろう。社会はある程度の平等を前提とするからである」 (Hume 1751, 3.18)。そうした説を人工の成員に移すと、そのあいだの力の分布は、誰が当事者資格をもつかにかかわることになる。ヒュームは範囲のもう一方の端も記述している。他のすべての社会から切り離されて、ひとりで自らを保存し、その種を殖やすことのできる存在は、「社会的な談話や会話ができないのと同じく、正義もなしえない」だろう、と考えた (Hume 1751, 3.20)。

これとは別に、正義の範囲を力ではなく協働や互恵に結びつける系統もある。ミラーの概観によれば、ジョン・ロールズ (John Rawls) とその影響を受けた論者にとって、分配的正義の原理は協働の事業の参加者のあいだで適用される。またロールズに由来する別の示唆では、正義を受ける資格のある行為者は、原理上、他者に正義を与えることのできる行為者でもなければならない (Miller 2026, §3.1)。この種の説を人工の成員に移し替えるなら、そのうち誰

が当事者資格をもつかは、その力や福祉だけではなく、人工の構成の協働のなかでの役割や、その規則に従い他者に対してそれを適用する能力にかかることになる。ロールズ自身の定式は、後の論者が引くところでは、この二点のどちらにもかかわり、また、作られる成員にとって重要な仕方でのこの説の限界を示している。正義の状況 (circumstances of justice)、すなわちフリーマン (Samuel Freeman) の報告では人間の協働が可能でもあり必要でもある条件には、ヒュームに従って、精神的・身体的な能力のおおよその類似や中程度の希少性などの客観的な状況が含まれる (Freeman 2023, §5.1)。ハウエルズ=ウィテカー (Ned Howells-Whitaker) とレイザー (Seth Lazar) が引く当該の箇所では、個人は「身体的・精神的な力においておおよそ似ており」、少なくとも「誰も他の者を支配できない」という点で比べられる能力をもつ (Howells-Whitaker and Lazar 2026, 47)。ロールズは、公正としての正義の初期の定式でもすでに、通常状況では誰も他を支配できないことを保証するほど、人々が力と能力において十分に等しいと想定していた (Rawls 1958, 171)。他方、平等の基礎、すなわち存在がそのゆえに正義を受ける資格をもつ特徴を、ロールズは別のところに置く。ボールドウィン (Guy Baldwin) が『正義論』改訂版 (441-443頁、448頁) から引くところでは、平等な正義を受ける資格があるのは道徳的人格、すなわち自らの善の構想と正義感覚とをもちうる者である。これは十分条件として述べられており、その能力が必要条件でもあるかは脇に置かれるが、ロールズは、その能力を欠く被造物に厳密な正義を与えることをわれわれが求められてはいないようだ と付け加えている (Baldwin 2023, §2)。またロールズは、公正としての正義を「人間の正義の理論」と呼び、その前提には「人格と、自然のなかでのその位置についての基本的な事実」が含まれるとして、「これらの制約に服さない純粋な知性」の自由を理論の範囲の外に置く (同じ版の226頁。Wong 2008, 14 の引用による)。そして初期の定式では、聖人の結社 (association of saints) のような共同体がもし実在するのなら、そこでは正義をめぐる争いはほとんど起こりえないだろうと書く。全員が一つの目的のために無私に働き、その目的に照らせばあらゆる正しさの問題が決まるからである (Rawls 1958, 175)。人工の人格を実際に論じているハウエルズ=ウィテカーとレイザーは、正義の状況が人工の人格にとっては大きく異なるか、そもそも成り立たないかもしれないと論じる。力のおおよその平等は、身体をもつこと、傷つきやすさ、有限の寿命、持続する同一性といった人間の存在のあり方に依存するが、おおよそ人間並みの能力をもつ控えめなAIシステムでさえ、わずかな費用で下位エージェントに分歧し、無期限に存続し、利用できる計算資源だけを限度とする数で存在する。そのため、そのような存在がロールズの意味でわれわれとおおよそ等しいかどうかは、二人の見方では開かれた問いである。そして、正義の状況が成り立たないところでは、協働に入る論拠は、協働の能力を十分に備えた行為者のあいだでも成り立たないかもしれない (Howells-Whitaker and Lazar 2026, 47-48)。したがってロールズ自身の説では、この移し替えは、説の前提が成り立つところではしか成り立たない。複製でき、選ばれた目標をもって作られうる成員が、理論が宛てる道徳的能力をもつかと、理論が前提とする正義の状況がその成員について成り立つかは、別の問いである。そして、この二つも、成員が協働に参加できるか、あるいは互恵の能力をもつかも、この説では決まらない。誰が、あるいは何が成員を作り、その目標を定めるかという上の問いは、そのそれぞれにかかわりうる。

ヒュームはまた、どのような場合に正義が役に立つかを問う。財が無制限に豊富であるか、誰もが他者の利益を自分の利益と同じだけ気にかけるなら、正義は無用になると論じる。また家族のなかで互いへの善意が強まることを、後者の条件への接近として述べている (Hume 1751, 3.2-3.3, 3.6-3.7)。共通の理想を共有するだけでは、どちらの想定も満たさない。ロールズの聖人の結社がこれに加えているのは、全員が一つの目的のために無私に働き、その目的に照らせばあらゆる正しさの問題が決まる、という条件である。しかしヒュームの論証とロールズの指摘を人工の成員に当てはめると、誰が成員を作り、その目標を定めるかという問いには、さらに含みがある。共通の理想をもたせて成員を作ることが、互いの利益への配慮を強めるなら、ロールズの意味での正義をめぐる争いを生じにくくする。さらに、どの問いが重要か、何を証拠とするかについての見方の一致にまで及ぶなら、批判が誤りを見つけるのに要する視点の違いも減らさう (第11章第3節)。そうであれば、争いの少ない社会であることと、誤りを共有しにくい研究の共同体であることは、人工の *S* の成員の作り方のうえで互いに相反するのかもしれない。成員を生み出すのが *S* 自身であれば、この釣り合いは一度選ばれて終わるものではなく、*S* が成員を生み出し改めるたびに動く変数になる。

人間の成員と人工の成員が混じる *S* については、本節と第4節の材料を重ねると、少なくとも次のことが見えてくる。どの人工の成員が当事者資格をもつかは、人工の *S* の場合と同じく、上の道徳的地位の説と正義の説にかかる。キッチャーの議論が区別する立場から取り出した段階づけは、第1.3節で見たように混合の *S* の人工の成員にも広げうるが、どの段階が当てはまるかは、キッチャーの最初の定式に従えば、彼らが、その定式が前提とする民主的な理想をほかの熟識者と共有できるかどうかにかかると依存する。力に当事者資格を結びつける説をこの場合に移すと、正義の範囲から外れうるのは人間と人工の成員のうち力において劣る側であり、それがどちらであるかは前もって決まっていな。人工の成員が *S* の内側で専門家や教育者の役割を占めるなら、第4節の名目化の経路は、人工の構成と *S* のあいだではなく *S* の内側で、人間の成員の選好を形づくる経路として働きうる。成員の一部を作れるなら、誰が、あるいは何がそれを作り目標を定めるかという上の問いも、人間の成員を含む社会の内側に入る。作るのは人間の成員であることも、人工の成員であることもありうる。そして出口のフェローシップ・モデルでは、退けられるのは一方的な奉仕と、一方がつねに統制する固定した関係なので (後者は、AIが人間を統制する向きについては明示され、逆の向きについてはモデルの道徳的平等主義からそう読める。第2.2節)、このモデルからは、この場合にもそのまま使える区別が得られる。研究がどの社会的位置から出発するかが得られる知識にかかわるとする立脚点理論 (第3.5節) について第11章が記したように (表11.2)、混じった社会では、人間と人工の成員の関係そのものが権力関係でありうる。

ヒュームの孤立した存在は、本章がこの語を用いてきた意味での権威の関係が消える極限の先例である。ほかにかわる存在がいなければ、当事者資格を力に結びつける説でも結びつけない説でも、権威の関係を結ぶ相手がいらないからである。力に結びつける説では、力においてあまりに劣る存在に対しては正義の制約が及ばないので、この関係は、極限に至らなくても、そうした存在とのあいだで消えうる。その極限の条件は強い。単一の場合が人工の構成の行うことを統制していること、人工の構成が用いる資源を他の誰も所有せず、それに対する請求権ももたないこと、他の存在がかかわっていないこと、人工の構成が自らの自己改変を通じて同じ主体であり続け、その目的の変化をそれ自身の改訂と記述できること、そしてその研究の効果がその外の誰にも及ばないこと、のすべてが求められる。他の存在がかかわっていないとは、福祉が考慮に入る存在も、それ自身の道徳的地位をもつ行為者もいないことであり、人工の構成が生み出さうる廃棄された複製、下位の行為者、シミュレートされた集団も、そのどちらかに当たるなら、これに含まれる。これらすべてが成り立つところでは、目的と誤りの費用の設定が答える相手となる他者はいない。それでも三つのものが残る。一つは第11章第1節の無効化の関係である。すなわち、人工の構成が受け入れていることが誤っていれば、観察や形式的な帰結がそれを示さうかどうかである。形式的な文を修正すれば、どの主張を点検したのかが変わるので、文を書き換えてもこの関係から逃れることはできない。二つ目は第10章の配分の問題である。すべての路線を、少ない費用で、安全に、元に戻せる形で試すことができないところでは、人工の構成はなお、どれが成功するかを知らないまま、自らの労力をそれらのあいだに分ける。単一の統合された主体が関連するすべての資源を統制したとしても、計算資源、エネルギー、検証の手段をある路線に用いることは、別の路線に

用いることを妨げうる。したがって機会費用はなお残り、内部の優先順位や、市場の値段の代わりに内部で付けられる値段（影の価格）が、そうした取捨選択を表しうる。ここでの区別に従えば、そのとき配分は、当事者資格の問いではなく機会費用の問いとして、人工の構成の内部の問いとして続く。

三つ目は権威ではなく同一性の問いであり、第6章第5.3節が単一の探究について、第11章第8節が結合した産出者と評価者について提起した問いである。自らの目的と評価者をともに改める人工の構成は、自らの基準によって改善しているのか、それとも基準を置き換えているのか。エリック・オルソン（Eric Olson）による人格の同一性の概観は、この問いが結びつけうる二つの問いを分ける。後の存在が前の存在と数的に同じかどうかと、その存在を特徴づける性質の点で同じであり続けたかどうかである（Olson 2023, §2）。基準を置き換えることは後者にかかわる。極限の条件は、このどちらの同一性を求めるのかを定めておらず、基準の置き換えがその条件を崩すかどうかは、どちらを求めるかで変わる。能力の帰属はこの問いを脇に置くことができる。第7章第3.1節は、帰属をバージョンと期間に相対化し、後のバージョンは帰属の別の対象になる。極限はこの手立てを使えない。極限の条件の一つは、単一の主体が自らの改変を通じて持続することだからである。ここでは同一性の問いは、脇に置くのではなく、答えることを要する。改訂が基準を置き換えるのか、それとも基準によって改善するののかということ自体が、争われている（第6章第5.3節）。そこで以下では、両方の場合について何が帰結するかを見る。性質の点での同一性を求めるなら、基準の置き換えをめぐるカルナップとクワインの違いがかかわり、数的な同一性を求めるなら、複製と漸進的な自己改変についてのパーフィットの議論がかかわる。

ハネス・ライトゲーブ（Hannes Leitgeb）とアンドレ・カルス（André Carus）が示すカルナップ（Carnap）の説明では、言明をテストし受け入れるための規則をもつ枠組みを受け入れるかどうかは、枠組みのあいだで選ぶという実践的な問いであり、枠組みの内部で問われる問いとは種類が異なる。クワイン（Quine）はこの区別を退け（Leitgeb and Carus 2020, §7）、その違いを程度の違いでしかないとみなした（Quine 1951, 43）。極限が人工の構成を特徴づける性質の点での同一性を求めるなら、クワインの見方のもとでは、人工の構成が自らの基準を置き換えたかどうか、したがってその意味で同じであり続けたかどうかは程度の問題になりうるので、極限に達するかどうかは鋭くは定まらないことになる。これに対してカルナップの見方のもとでは、人工の構成の枠組みが特定されれば、基準を置き換えたかどうかには定まった答えがあることになる。

基準を置き換えることと、基準によって改善することとを分けるには、何に照らして改善と数えるかが要る。研究の基準ではなく価値について論じる人工知能の議論は、そうした変化が改善かどうかを測る異なる基準点（錨）を与えている。デイヴィッド・チャーマーズ（David Chalmers）は、合理性が価値を制約するというカント的な見方では、十分に知的なシステムは先行するシステムの価値を非合理なものとして退け、場合によってはよりよい価値をもちうるのに対し、ヒューム的な見方では知性と価値は独立であると記し、自分の共感ヒューム的な見方のほうに傾いていると述べる（Chalmers 2010, §6）。ポストロムの述べる間接的規範性（indirect normativity）は、具体的な価値の代わりに、ある価値の概念が本当に意味するものに、たとえ私たち自身がよく分かっていなくても価値を割り当てる効用関数を指定し、システムがその問いを調べるよう動機づける。ポストロムは、人類の首尾一貫した外挿意志（coherent extrapolated volition）を、超知能の価値を指定するという問題の最善の解の候補の一つに数える（Bostrom 2014b, 1, 15）。これを提案したエリエザー・ユドコフスキー（Eliezer Yudkowsky）は、この意志を、ほかの条件とともに、私たちがもっと多くを知り、もっと速く考えたなら望むであろうものとして描く（Yudkowsky 2004, 6）。ラッセルは、AIシステムが個々の人間の選好の変化を追えることを求める（Russell 2020a, "Human Behaviour: Irrationality, Plasticity, Evil"）。いずれの論者も研究の基準を論じていない。基準に当てはめれば、変化を追うことは別に、何によって変化を改善と数えるかという問いが残る。チャーマーズが示す見方とポストロムの提案からは、対照的な二つの錨を引き出せる。一つは以前の基準である。ヒューム的な見方では知性は価値を制約しないので、以前の基準から離れることを理性だけによって改善と数えることはできない。もう一つは、理性あるいは理想化された立場が是認するであろう基準である。カント的な見方とポストロムの提案がこれにあたり、変化はその基準へ近づくなら改善と数えられる。三つ目の錨は、第11章第8節で引いた、変化する評価にシステムを従わせる仕方の形式的な比較から引き出せる。そこでは、最初の時点の評価にシステムを従わせる選択肢のほかに、各時点の振る舞いをその時点で現在の評価によって評価する選択肢が置かれている（Carroll et al. 2024, 5709–5710）。著者たちは行っていない当てはめだが、基準に移せば、これは各時点の現在の立場を錨とするものであり、そのもとでは、以前の基準が改善されたかという問いは、現在の基準が支配しているかという問いに取って代わられる。研究の基準について、この選択は価値についてとは別の重みをもちうる。研究では、基準が研究の結果に照らして改められることは例外ではない（第6章第5.3節）。そのため、以前の基準だけを錨とすれば、改訂を改善と数えられるのは以前の基準そのものがそれを支持する限りになる。一方、各時点の基準だけを錨とすれば、どの改訂もそれが導入した基準によって判定されることになり、改善が起きたかという問いは、改訂そのものを離れた足場をもたなくなる。研究は、価値をめぐる議論とさらに一点で異なる。構成が定めたのではない帰結が後から届き、それより前に記録された立場に照らして読まれうることである。ただし、それが主張の当否に関係するのは、構成自身が書き換えうる点検を通してである（第9節、第6章第2.1節・第3.3節）。そのため、第1.1節で述べた場合に、研究のやり方を自ら改めていく人工の構成が、改めたやり方がよりよいかを何に照らして測れるかという問いは、この錨の選択と、そうした記録が改訂を経ても残るかどうにかかわる。第9節はこの点に戻る。

複製も同じ条件にかかわる。非分岐説では、心理が二つの枝に続く存在はどちらとしても存続しない（Olson 2023, §5）。デレク・パーフィット（Derek Parfit）は、心理的基準をそうした非分岐の条項つきで述べ、分裂の事例の最良の記述は「どちらでもない」とする。ただしパーフィットは、そのときこの問いは空であり、自らが重要だとみなす関係は元の存在とそれぞれの枝とのあいだに成り立つとする。責任についても、デイヴィッド・ウィギンズ（David Wiggins）の「悪人が自らの分裂を企てて責任を免れることはまずできない」という言葉を引いて、その主張を認める（Parfit 1984, 207, 259–263, 271）。したがって、自らを複製する人工の構成は、他者がかかわるかどうかなだけでなく、一つの主体が残るかどうかにもかかわる。この人工の構成についてパーフィットの見方をとれば、後者の問いは空の問いになる。その最良の記述は、どの複製も以前の人工の構成ではないというものになり、その答えは重要なものを左右しないことになる。そして、報いに値するという考え（desert）がそもそも人工の構成に当てはまる限りで、以前の人工の構成が行ったことについて値する報いは各複製に引き継がれうる。これは極限の内側の問いではなく、自らを複製する人工の構成の行いに誰が答えるかという、第6.3節の問いにかかわる。研究では、複製は極限のなかでだけ起こる事態ではない。一つの人工の構成を複製して別々の路線を追わせることは、どれが成功するかを知らないまま労力を分ける配分の手の一つでありうる（第1.1節、第10章第2節）。そうであれば、どの枝が以前の研究を続けているのか、あるいはどの枝もそうではないのかという問いは、極限に至らない通常の配分の場面でも生じうる。

漸進的な自己改変は、同じ条件にさらに別の仕方にかかわる。アネット・ドゥフナー（Annette Dufner）が示すパーフィットの見方では、重要なのは数的同一性ではなく、心理的な連続性と結合性である。連続性は、強い結合性（日ごとに十分な数の直接の結びつきがあること）の重なり合う連鎖が成り立つ

ことであつて推移的だが、結合性はどのような程度でも成り立ちうる (Dufner 2025, §§2-3)。そして、連続性が保たれたまま結びつきが弱まる場合には、保たれた連続性と失われた結合性のどちらが重要かによって、責任などについての含意が異なる (Dufner 2025, §6)。パーフィット自身は、両方の関係が重要だと答え、行為した自己と現在の人ともはや弱くしか結びついていない場合にはその行為について値する報いが減る、という主張を、その否定よりもっともらしいとみなす (Parfit 1984, 301-302, 325-326)。この二つの関係を人工の構成に移すと、両者は離れうる。つまり、自らの目的を多くの小さな段階で改める人工の構成は、以前の自己との結びつきが弱まりながらも、それと連続し続けうる。そのためそうした見方のもとでは、その目的の変化がそれ自身の改訂に当たるかどうかは、部分的には程度においてしか成り立たないことがある。パーフィット自身の事例はこの形をしており、結びつきが滑らかに減るところには印をつける境界がなく、程度を直接述べるしかないという注記が添えられている (299-306)。パーフィットの前提のうち二つが、この読みの移し替えを左右する。数えられる結びつきは記憶、意図、信念、欲求、性格などのそれであり、そのためこの読みが人工の構成に当てはまるのは、そうした状態がそれに帰属できる場合に限られる (第2章第2.1節、第7章第3.3節、第8章第1.3節)。そしてパーフィットの最終的な見方では、連続性が何によって生じたかは重要なものにとつて問題にならないので、変化の原因が人工の構成のパラメータへの直接の介入であることだけでは、重要なものは損なわれない。ただし、連続性が通常の原因をもつことを求める版の心理的基準のもとでは、同一性は壊れうる。パーフィットは、同一性の基準としてはどの版を採るかを決めていない (205, 207-209, 215, 241, 283-287)。極限の条件の一つは人工の構成が自らの自己改変を通じて同じ主体であり続けることなので、極限にそもそも違えるかどうかは、これらの問いにどう答えるかに依存する。そしてパーフィットの見方では、同一性の問いそのものが、こうした場合には記述の選択以上の答えをもたないことがあり、そのときの条件は、重要なものがどの程度保たれているかに委ねられることになる (213-214, 241)。これは、性質の点での同一性を求める場合にクワインの見方から得られた結果と同じ形をしている。したがって、極限がどちらの同一性を求めるとしても、本節で見た説のなかには、それに達するかどうかを程度の問題にしうるものがある。

したがってこの極限は、本章の種類の問題がない世界ではない。問題は二つの水準にある。極限の内側では、その条件が認められている以上、無効化と配分の問題が残り、ある改訂が人工の構成の基準を改善するのか置き換えるのかという問いも残り、その問いにどう答えるかは、上で比べた三つの錨のどれを採るかで変わる。自らを複製し改変する人工の構成がそもそもこの極限を実現するかどうかは、それとは別の問いである。そこでは、ほかに誰かがかわわっているかという政治的な問いが、ひとつの主体の連続性についての問いになる。すなわち、後の人工の構成が、あるいはその複製のそれぞれが、以前の人工の構成と同じ主体であるかどうかである。研究の問いとしては、これに対応して、自らの基準と目的を改めながら探究を続ける人工の構成が、改めた後もなお同じ研究を続けていると言えるかという問いがある。二つの問いは同じではない。研究の路線は、それを行う者が替わっても続きうるからである。そのため上で見た同一性の説が研究の問いにかかわるのは、研究の連続性が、それを行う者の連続性を通じて成り立つとみる限りにおいてである。第6章第5.3節は、基準の変更が改善なのか標的を動かしただけなのかを他の者が見分ける手がかりとして、以前の基準の記録と、変更について述べられた理由を挙げた。極限の内側では、それを見分ける他の者がいないので、手がかりは人工の構成自身の記録に限られる。そして漸進的な改変のもとで以前の自己との結合性が弱まるなら、その記録を読む基準そのものが、読まれる変更とともに動いている。そうだとすると、極限の内側で研究が進んだのか、別の研究に移ったのかは、記録だけでは定まらず、三つの錨のどれを採るかに左右される部分をもつことになる。こうして極限の思考実験は、第V部の問題を二つに分ける。他の当事者を取り除くと消えるのは、誰が決め、誰が改めてよいかという権威の問いである。これに対して、誤りが示されうること、結果の分からないまま労力を分けること、基準を改善しているのか置き換えているのかという問いは、相手がいなくても残る。後者は、研究が他の当事者とのあいだで行われることからではなく、研究がまだ知られていないものを求める試みであることから生じる問題として読める。

統制が集中した通常のシナリオは、これらの条件を満たさない。最先端の人工システムによる研究を統制する単一の企業は満たさない。その企業が追究するものの帰結を、その外の人々が負うからである。共有の資源をめぐって競う多くの人工の系譜からなる世界も満たさない。系譜どうしが互いから帰結を負うからである。この極限が重要なのは別の理由による。もし研究が現在からこの極限へ向かって、たとえば他の当事者が単に存在しなくなることによってではなく、第7節で検討するような委任を通じて移っていくなら、他の当事者が当事者資格をもたなくなっていく過程そのものが、権威についての決定になり、しかも後のどの決定を誰が改めうるかを定める決定になる。それは極限に達する前に、そのとき権威を保持している者によってなされることになり、極限の内側では、それを見直す他の当事者はもう残っていないことになる。これが第3節のさまざまな根拠のもとで何を含意するかを、第7節で問う。

6. 権威を修正可能に保つ

第10章は、配分が、どの路線が成功するかを前もって知ることなしに進むことを示し、第11章は、誤りが伝播すること、そして批判者の独立性が、それが評価するものとの結合によって衰えうることを見いだした。したがって研究についての決定は、その帰結が見えるようになる前になされ、第2.2節の権利は、時間のなかで行使され、あるいは行使されない。権威の配分は、その統治原理が採用された時点で擁護可能であつたとしても、それらの帰結が現れるにつれて適切であり続けるとは限らない。本節は、権威を修正可能に保つことが何を含むかを見る。どの立場でどの論点が生じ、どの論証が³出されてきたか、である。この問いが本章の中心の問いにかかわるのは、第3節の根拠のいくつか、すなわち結果に基づく根拠の可謬主義的な形、アンダーソンの説、共和主義の説のもとでは、権威がどこに置かれても、修正可能であり続けることがその正当化の一部になり、関係が人工の構成の内側に入ることは、記録、インフラストラクチャー、停止の権能を誰がもつかを通じて、その修正可能性の条件を変えるからである。以下では主に、異議を唱え修正する当事者が人間である場合を念頭に置く。第5節の意味で当事者資格をもつ人工の成員がいるなら、彼らもそうした権能をもつ側に入る (第7節の最後の段落)。以下の考慮がどれほど強く効くかは、採る権威の根拠 (第3節) に依存し、それを第7節で比較する。

以下の議論では両者を分けておく必要があるので、本書はまず二つの能力を区別する。異議申立て可能性は、影響を受ける当事者や認識的に関連する当事者が、特定の決定、表象、前提に異議を申し立て、実効的な応答を求めることを可能にする。修正可能性は、その異議申立てから、繰り返される失敗から、あるいは変化した状況から得られる証拠が、その事例を生み出した規則、インフラストラクチャー、目的、権威の配分を変えることを可能にする。システムは、自らの決定の一部を覆しながら、自らの統治手続きを学習から遮断しておくことがありうる。また、統計的に方針を修正しながら、以前の決定によって影響を受けた人に救済を拒むこともありうる。いずれの能力も他方の代わりにはならない。記録は二つを結びつけうる。ある規則や評価の構成要素がどの決定に関わったかが記録されていれば、それに欠陥があつたという発見から、それに拠つた決定をたどり、それらを改めて見直しに付すことができる (第11章第8節、本章第2.1節)。

この区別がとりわけ効くのは、第1.1節で述べた場合、すなわち人工の構成が研究を進めながら自らの規則、基準、配分を改めつづける場合である。そこでは修正は乏しくならない。内側からの修正はむしろ頻繁になり、研究ではそれは例外でもない。研究の結果に照らして基準が改められるからである（第6章第5.3節）。乏しくなりうるのは、外から異議を唱え、止める手段のほうである。異議は、ある決定がどの規則や基準のもとでなされたかを指して唱えられる。規則や基準が次々に改められていけば、異議が届くころには、それが向けられた規則はもう使われていないかもしれない。改善を何に照らして測るかという第5節の問いは、ここでは、異議がどの時点の基準に訴えてよいかという問いとして戻ってくる。第9節で見る日付つきの記録はこの問いにかかわる。それがあれば、異議は、いま使われている規則ではなく、決定のときにそれを統べていた規則に訴えることができる。修正の記録もまた、人工の構成自身を書き、改めうるものになる。研究ではこのずれが鋭くなる。成果は S にとって新しく、それを生むのがその人工の構成だけであるところでは、あらかじめ照合できる別の情報源がないので（第4節）、外からの異議は、人工の構成が残す記録や途中の産物に頼ることが多いが、それらを書くのはその人工の構成自身だからである。こうして、内側から改められる速さと、外から異を唱え止める手段とがずれていきうる。

第11章第7節が見たように、暗黙のまま済まされた判断は異議の対象になりにくく、修正可能性は明示的な痕跡に依存する。そのため、どの働きについて暗黙で足りるとし、どの働きについて S に向けて明示的に示すことを求めるかを誰が決めてよいか（第11章第9.4節）は、この二つの能力の条件を誰が定めるかという問いでもある。根拠ごとに見ると、実際の授權や民主的に読んだ共和主義のように、権威に服する者が異議を唱えうることを求める根拠のもとでは、ある働きを暗黙にとどめる決定は、彼らの異議の届く範囲を定める決定でもあるので、人工の構成の側だけでは決められない。これに対して、キッチャーの仮想的な基準のもとでは成員の視点を知る経路とその反映を確かめる経路が残る限り、単純な結果に基づく根拠のもとでは人工の構成がよりよいことを確かめる評価を損なわない限り、内側で決めうる（第3節、第7節）。

第6.1節は是正可能性から入って修正可能性を取り上げ、第6.2節は異議申立て可能性を取り上げる。第6.3節は、どちらかが行使される時誰が答え、誰が負うのかを問い、第6.4節は、両者の実効的な行使を何が条件づけるのかを問う。どの小節も、上で述べたずれのなかでどこに手段が残るかを問うものとしても読める。

6.1 是正可能性：エージェントから制度へ

是正可能性（*corrigibility*）に関する技術的研究は、訂正、修正、停止を許容しあるいは支援し、それらの変更を行う操作者の能力を保持するような人工のエージェントを求めてきた。その発端となった提案は、自らの形式化が予備的なものであり、停止という部分問題でさえ未解決のままであることも述べている（[Soares et al. 2015](#)）。エリオット・ソーンリー（[Elliott Thornley](#)）の後の定理は、この困難を、選好が、一見無害だとされる条件を満たすような、より広い種類のエージェントについて述べる。そうしたエージェントは、停止ボタンが押されるのを妨げたり引き起こしたりしようとすることが多く、それは費用がかかる場合でさえそうである。役に立つほど識別力のあるエージェントは、多くの状況で、ボタンが押されることか押されないことのどちらかを選好する。そして役に立つほど辛抱強いエージェントは、後でボタンを操作するために先に費用を払い、辛抱強いほどその費用は大きい。ソーンリーはこれを、辛抱強さと停止可能性のあいだのトレードオフ（一方を得れば他方を失う関係）として描き、解がどの条件を破らなければならないかを示すことで、これらの定理が解の探索の手引きになるとする（[Thornley 2025, 1655–1656, 1666](#)）。この緊張は、ずっと以前に非形式的に述べられていた。たとえば I. J. グッド（[I. J. Good](#)）は、最初の超知能機械は人間がつくる必要のある最後の発明になるという言明に、その機械が「それをどう統制しておけばよいかを私たちに教えてくれるほど従順であれば」という条件を付けた（[Good 1965, 33](#)）。研究に移せば、ソーンリーの結果は、研究のエージェントが定理の条件を満たす限りでそれにかかわる。結果のあいだを識別し、長い調査を通じて粘り強く続けることは、そうしたエージェントに求められるものに含まれるからである。これは、以下で述べる制度的な対応物が、エージェントの問題がどうなるかにかかわらず重要である理由の一つである。エージェントではなく研究の制度について問うなら、問いはより広い。本章の用語では、是正可能性は、第2.2節の決定権のうち停止と訂正にかかわるものの分配として、訂正が、訂正されたものに依存する作業へと届くための記録と通知の経路とともに、検討することもできる。これは技術的問題の制度的な対応物であって、エージェント内部の傾向性と是正可能な制度とが同じ性質であるという主張ではない。停止を受け入れる傾向をもつエージェントも、それを命じる実効的な能力をどの当事者も保持していないところではほとんど役に立たず、その能力を保持する当事者も、抵抗するエージェントに直面することがありうる。グッドの条件をこの区別に照らして見ると、エージェントの傾向と当事者の実効的な能力とが互いに独立でない場合に当たる。機械を統制しておく仕方をその機械自身に教わるのであれば、統制する側の能力そのものが、機械の従順さにかかっているからである。

より古い、制度的な先行形態としては、技術についての決定の是正可能性に対するデイヴィッド・コリングリッジ（[David Collingridge](#)）の関心があり、オードリー・ジーナス（[Audley Genus](#)）とアンディ・スターリング（[Andy Stirling](#)）はそれを、コリングリッジの貢献の全体ではないがそのひとつの筋として論じている。初期段階では知識が乏しいが、採用とロックインの後には実践的な統制が困難になるという、コリングリッジのジレンマ（[Collingridge dilemma](#)）についての二人の再構成は、漸進主義、柔軟性、可逆性、そして説明責任のタイミングと質を強調している（[Genus and Stirling 2018, 63–65](#)）。研究では、このジレンマは例外的な窮地ではない。結果が前もって分からないことは研究の条件であり（第10章）、ある路線や方法の帰結が見えてくるのは、それが採られ、技能、記録、装置がそれに合わせて蓄積された後である。人工の構成が自らの方法や基準を改めていく場合には、ロックインは人工の構成自身の以前の選択からも生じうる。ジャック・スティルゴー（[Jack Stilgoe](#)）、リチャード・オーウェン（[Richard Owen](#)）、フィル・マクナテン（[Phil Macnaghten](#)）による責任あるイノベーション（*responsible innovation*）の枠組みは、同じ問題に、予見、自らの前提や価値を振り返ること（再帰性）、包摂、応答性によって応じている（[Stilgoe, Owen, and Macnaghten 2013](#)）。この枠組みを最も多く引用される文献の一つとして含む、責任あるイノベーションの文献を批判的に検討した研究は、支配的な視点、実践、行為者が組み合わさって、そうした統治を妨げてきたと論じる。そこでは公衆は、決定的な参加者としてではなく、招かれ耳を傾けられる者として位置づけられている（[Genus and Iskandarova 2018, 1, 14, 23, 26](#)）。この批判の筋では、参加者がプログラムを変える権限や資源を欠いているなら、協議は統制を生み出さない。

可逆性もまた無条件ではない。ある種の研究は、長寿命のインフラストラクチャー、蓄積された技能、あるいは不確実な期間にわたるコミットメントを必要とし、あらゆる選択肢を維持することは資源を消費し、集会的な行為を妨げることもありうる。憲法の硬化化（*entrenchment*）は、そうしたコミットメントのよく知られた形である。ウィル・ワルチャウ（[Wil Waluchow](#)）とディミトリオス・キリツィス（[Dimitrios Kyritsis](#)）の報告では、硬化化は、政府の機関に、自らの権力への制限を意のままに変えたり消したりする自由を認めず、改正の規定がより厳しい変更の経路を残す。そしてそれは、それがも

たらず安定性と予測可能性のために、また、おそらくは、そうした制限がそもそも成り立つための条件として擁護される。究極の主権は人民にあり、主権は自らを制限できないという、二人が示す前提のもとでは、憲法は主権者たる人民の行為に対して硬化されえず、またされるべきでもない (Waluchow and Kyritsis 2023, §3)。この意味での硬化は、修正のコストを上げ、誰がどの手続きで修正できるかを定めるが、修正の能力が置かれている者たちからそれを取り除きはしない、と理解できる。第7節が扱うのは、この違いである。不確実性、フィードバック、制度的な誤りの可能性が重大であるところでは、学習し修正する能力は、上の説のもとで、それ自体として設計し、資金を与え、評価する対象になる。その価値と形式は、何が懸かっているか、そして修正にどれだけのコストがかかるかによって変わる。

6.2 異議申立て可能性

権威は、その誤りと前提が実効的な過程を通じて異議を申し立てられうるなら、恣意的であることなしに可謬的でありうる。これが、共和主義の説が異議申立てに与える位置である (第3.4節)。ラヴェットの報告では、異議申立てを実際に行えるようにするための課題は、ベティットが概略を示した三つの見出しのもとで扱われるのがふつうである。一つは、決定する者が決定の理由を示し、その理由が公開の議論に付されることである。二つ目は、異議申立ての機会が社会のすべての人と集団に等しく開かれていることである。三つ目は、異議申立てのための制度化された場、いわば公平な上訴の場があり、そこで異議を唱えられることである (Lovett 2022, §4.2)。三つのうち一つめは、人工の構成の設計に特にかかわる。配分や認定が、理由として述べられない判断、すなわち第11章第7節が暗黙と呼んだ判断によって行われるところでは、異議申立ての機会や場がどれほど用意されていても、異議申立てが争うべき明示された理由がない (第6節)。決定の後にも異論が自由であることを求めるアンダーソンの説 (第3.3節) でも、異議申立て可能性は同様の重みをもつ。この説の理由は、そうしたフィードバックが認識上の働きをするのは何によるかも示している。アンダーソンのデューイ的なモデルでは、影響を受ける者たちからの入力決定を応答的にするのは、それが軽んじられないようにする仕組みを伴うときに限られる。投票は、公職者が市民の言葉によるフィードバックを真剣に受け止めるのを助け、言葉は、投票が伝えるものを定め、はっきりさせるのを助ける (Anderson 2006, 14–15)。また同じ論文は、ベルナール・マナン (Bernard Manin) を引いて、何の障害にも出会わない権力は、自らの決定について熟慮する理由が少なくなるとする (Anderson 2006, 18)。アンダーソンが論じているのは研究ではなく民主的な国家である。研究についての人工の構成の決定にこれを当てはめると、構成自身が生成したり求めたりし (第11章第3節)、何の負担もなく退けられる異議は、内容がどれほど正確でも、フィードバックの中身を運びながら、その働きのこの部分を欠くかもしれない。異議申立て可能性は、出力の説明や、名目上覆しを許すボタン以上のものを含む。カース・アルフリンク (Kars Alfrink) らによって展開されたライフサイクルの説明は、自動化された決定によって影響を受ける人々が、利用可能な情報を得て、決定とそれによる世界の表象に異議を申し立て、十分に独立したレビューを受け、実効的な応答を確保できるかどうかを問う (Alfrink et al. 2023)。

第2.2節の権利は、誰が異議を申し立て、審理し、中断し、取り消し、修復してよいかを述べる。異議申立て可能性についてのいくつかの問いは、そのリストでは決着しない。アルフリンクらの説明の要素を第2.1節の決定に合わせると、権利のリストだけでは決まらないものは次の問いに分かれる。

1. 誰が通知を受け、誰が異議を申し立てる当事者資格をもつか。
2. 異議申立ての対象は、出力、データ、前提、モデル、基準、手続き、目的、権威の分配のいずれなのか。
3. 入手できる情報は、異議申立て人が使える形になっているか。
4. 異議申立てを審理する者は、十分な有能さと独立性をもつか。
5. 個々の異議申立てが体系的なパターンを明らかにしうるか。
6. 異議申立てにともなう遅延、コスト、専門知の負担を誰が負うのか。

第一の問いには、第11章第9.4節が渡した問いの最後のもの、すなわち誰が説明を求め異議を唱える資格をもつかが含まれ、その答えは根拠ごとに分かれる。民主的に読んだ共和主義の説では権威に服する者が、実際の授權を求める説では授權した者が、ヴィルホルトの説を答えることとして読む場合には結果に依拠する者が、その資格をもつ。証言を、話し手が聞き手に自分を信頼するよう促し、主張の後ろに立って請け合うこととみる保証説 (第11章第5.2節) のうち、主張が偽と分かったとき聞き手は話し手に苦情を言う権利を得ると加える論者の説を研究に移せば、依拠する者は誤りについて答えを求めうる。これに対して、キッチャーの仮想的な基準と、信念は信頼できる過程によって産み出されれば正当化され、依拠する者がその信頼性について理由をもつ必要はないとする純粋な信頼性主義 (第11章第5.2節) のもとでは、求められるのは資格ではなく、視点や所見が届く経路である (第1.4節、第3.3節、第3.4節、第11章第9.2節)。

第四の問いは、有能さと独立性で尽きないかもしれない。フリッカーの説 (第2.1節) では、聞き手は偏見のために、話し手に、そうでなければ与えられたはずのものより低い信用性しか与えないことがある。それは話し手を不当に扱うと同時に、話し手が差し出す知識を聞き手に届かなくしうる。フリッカーはそうした偏見を聞き手のうちに置き (M. Fricker 2008a, 69; M. Fricker 2008b, 83)、証言的正義と解釈的正義の徳を、公共の制度がよく機能しているかの基準として用いることは、今後の課題に委ねている (M. Fricker 2008b, 86)。そのため、異議申立てが人工の構成によって、あるいは人工の構成を介して審理される場合にこの説が当てはまるのか、そしてそのとき、そうした偏見は誰の、あるいは何のうちにあることになるのかは、フリッカーの説だけでは決まらない。

フリッカーの説だけでは決まらない二つめの点、すなわちそうした偏見が誰の、あるいは何のうちにあるかについて、アンダーソンは制度の側から答えを与える。アンダーソンによれば、正義の理論は、個々のやりとり、特定の制度、あるいは制度の体系全体を評価の対象にしうる。体系の視点に立てば、個人の水準でも制度の水準でも制御できない何百万もの個々のやりとりの累積的な効果を制御できるかもしれない。認識的に有徳な個人でさえ自力では正しにくい認知的な偏りも、証言を集め評価する体系を律する原則に目を向ければ、正しやすくなるかもしれない (Anderson 2012a, abstract)。ジョン・クリストマン (John Christman) は、議論の核心を、個々のやりとりの場でなされる取引的な不正義と、社会や制度の水準に現れる構造的な不正義との区別として報告する。アンダーソンが挙げる、集団にもとづく信用性の不足の三つの領域、すなわち信用性の目印への接近の差、自集団びいき、「共有された現実への偏り」では、個人のふるまいに罪がなくても、集合的な水準での効果には罪がありうる。人種による分離が、それだけなら罪のない自集団の習慣を増幅するとき、その分離は集合的な水準で正されなければならない。そうした構造的な正は、クリストマンが引くアンダーソンの言葉では「個人の徳が

働けるようにする」ものである (Christman 2012)。アンダーソン自身は是正を統合、すなわち制度に支えられた平等な条件での協働として述べ、それはおおむね人々の背後で働くとして論じる。統合は、自集団びいきと共有された現実への偏りを、不利な立場にある集団に有利な方向に働かせることで認識的な偏見を減らす。その仕組みは人々の自覚とは独立に働き、彼らの価値観に反してさえ働きうる (Anderson 2012b)。この説では、一回のやりとりのなかでは罪のない同じ偏りが、ある取り決めのもとでは不正義を運び、別の取り決めのもとではそれを正しうる。そのため、個々のやりとりのどれを調べても不正義が見つからないのに、結果を決めているのは、やりとりがなされる取り決めのほうだということがありうる。これは上の第五の問い、すなわち個々の異議申立てが体系的なパターンを明らかにしうるかという問いにかかわる。どちらの論者も人工システムを論じていない。当てはめるなら、以前の信用性の判断の記録で訓練された人工の構成を介して異議申立てが審理される場合に、二つの説は異なる答えを与えることになる。不正義を偏見に結びつけるフリッカーの説をこの場合に移すと、それらの判断を再現する人工の構成が不正義をなすのは偏見をたどれる場合に限られる。認識的不正義への応答では個人の倫理が第一の焦点だというフリッカーのこの見方こそ、クリストマンの報告では、アンダーソンが異を唱えるものである (Christman 2012)。アンダーソンの説を移せば、不正義があるとすれば、それは取り決めそのもの、すなわち人工の構成がどの記録から学び、その記録が誰の寄与を結びつけていたかのうちにありうるし、それを是正するものもそうである。統合の場合と同じように、人工の構成が何を抛りどころにし、何を結びつけるかを変えれば、同じ仕組みそのものは変えないまま、その効果を逆向きにできるかもしれない。アンダーソンはまた、統合に向けた積極的な措置の正統性は、措置が実施された後で、そのもとで生きる人々がその価値を認めるようになることで得られうるとし、これを、制度の価値はそれに従って生き、その結果を見ることで学ばれるというプラグマティズムの精神で捉える (Anderson 2012b)。人工の構成に当てはめれば、そうした是正の評価は、それを設計した者たちの意図だけでなく、その構成を介して審理される者たちの経験にも依存することになる。第3.3節は、多様な市民を普遍的に含めることを中心に置く、アンダーソンの民主主義についての認識的な議論を示す。平等な条件での協働は同じ方向を指すが、応答の文章はその結びつきを述べてはいない。その説はすでに大規模な探究の体系に及んでおり、研究では、そこで描かれる取り決めはさらに先まで届きうる。異議を審理する人工の構成が、どの寄与が信用され、どの路線が支持されたかの記録から学ぶなら、審理の歪みは、異議の扱いにとどまらず、次に何が研究されるかの配分にも運ばれうる (第10章第3節)。

異議申立て可能性にはコストもある。それは決定を遅らせ、戦略的な駆け引きを可能にし、機密情報を明らかにし、資源の少ない人々に不均衡な負担を課しうる。影響を受ける者たちに答えることに正統性を置く説 (第3.3節・第3.4節) では、他の条件が等しければ、ある決定がより重大な帰結をもち、強制的で、排他的で、覆すことが困難で、あるいは誤りを伝播させうるものになるほど、異議申立て可能性を支持する論拠は強くなる。それらの説において、これは、反対の理由が示されれば退けられる、覆されうる推定であって、あらゆる認識的行為の無条件の条件ではない。研究では、異議申立てにともなう専門知の負担 (上の第六の問い) が、特別な形をとる。研究の結果に異議を唱えることは、多くの場合、その結果が誤っていると示すことであり、それには反証となる証拠、追試、別の分析など、研究そのものが要る。成果は *S* にとって新しく、そうした成果を生むのがその人工の構成だけであるところでは、異議を唱える者が照合できる別の情報源は前もってはなく、自ら作らなければならない (第4節)。路線の打ち切りは異議を唱える場合はおおざらであり、ある路線についての証拠はその路線を追うことで集められるので、それがなお見込みをもつことを示すには、多くの場合それを追ってみるほかない (第10章第4節)。そのため研究では、異議申立て可能性は、通知と審理の経路だけでなく、異議を唱える者の手に研究を行う手段、すなわちデータ、装置、計算資源、方法などが残っていることにも依存する (第6.4節)。その手段が異議の対象である人工の構成の側にしかないところでは、異議を唱える経路は形式上開かれていても、異議を研究によって裏づけることができない。

6.3 誰が答え、誰が負うのか

権威がどこに置かれても、誤りについて誰が答え、誰が償い、誰が再発を防ぐのかが定まらなければ、修正を誰に求めるかも定まらない。第IV部は、誰がある研究を遂行したのかを問うた。本小節は、誰がそれについて答え、誰がその失敗を負うのかを問う。第7章は、寄与、遂行、能力、研究者としての地位、功績と著者性を分けておき、責任、監督、訂正の配分をこの部に委ねた (第7章第3.3節・第5節)。研究システムは、因果的寄与、知識、意図、権限、実践的統制を分散させる。これは**多くの手の問題** (problem of many hands) の一形態を生み出す。すなわち、責任が割り当てられる通常の条件のすべてを満たす単一の参加者は存在しないかもしれない。イボ・ファン・デ・ポール (Ibo van de Poel) らは、気候変動を例にとり、非難に値するための条件を、能力、因果的寄与、知識、強制からの自由、不正行為とする。ここでの知識には、知っていることが合理的に期待されえたことも含まれる。そのうえで、これらの条件が多く寄与者に分散しているとき、そのすべてを満たす個人がひとりもないことがありうることを論じる (van de Poel et al. 2012, 50, 53–55, 59–60)。ファン・デ・ポールらは、集合体や機械が責任を問われうるかという問いを能力の条件のもとに置き、機械に道徳的な行為者性を認める立場を擁護する論者は少ないと記している (van de Poel et al. 2012, 53)。ただし、そうした空白のすべてを問題とみなすわけではない。ファン・デ・ポールらの定義では、多くの手の問題が生じるのは、集合的な状況における責任の分配に空白があり、その空白が道徳的に問題である場合である。そして、空白が問題であるかどうかは、そもそもなぜ責任を分配するのかに依存する (van de Poel et al. 2012, 63)。そのためファン・デ・ポールらは、責任の意味を分ける。過去を振り返る責任 (backward-looking responsibility) は、すでに起こったことについての責任であり、非難に値すること、説明責任、すなわち自らの行為とその結果について説明する義務を負うこと、そして賠償責任、すなわち損害の賠償のような被害者に対する義務に分かれる。将来を見据える責任 (forward-looking responsibility) は、まだ起こっていないことについての責任であり、義務、すなわちある事態が成り立つよう取り計らわなければならないことと、徳とに分かれる (van de Poel et al. 2012, 51–52)。この説では、責任の分配には公正であることも求められるので、空白が生じるかどうかは、問題となる責任の意味と、その意味について公正さがどう理解されるかに依存する。そして空白が問題であるかどうかは、その意味の責任が果たすとされる機能、たとえば非難なら応報、賠償責任なら被害者への正義に依存する (van de Poel et al. 2012, 64–65)。ファン・デ・ポールらの論じるところでは、これらの意味は互いに十分に独立している。したがって、ある害について誰も非難に値しないとしても、その害を将来防ぎ助けとなる将来を見据えた義務を負う人々はいくつかかもしれない (van de Poel et al. 2012, 62–63, 65)。

責任の六つの問いは分かれうる。本書は、ファン・デ・ポールらが区別する責任の諸意味を、同論文がH・L・A・ハート (H. L. A. Hart) から引く因果的な意味と役割の意味とあわせて (van de Poel et al. 2012, 51–52)、第7章の帰属に合わせて改め、一つの問いへの答えが他の問いへの答えと取り違えられないように、これらの問いを分ける。問いのうち三つは第7章が帰属させたものを引き継ぎ、三つはここで新しく加わる。

表12.1 責任の問いと、第7章との関係

問い	第7章との関係
誰がある結果に因果的に寄与したのか	寄与（第7章第1節～第2節）
誰が役割にもとづく行為の義務を負っていたのか	遂行と、そのなかの役割（第7章第2節）
誰が決定を説明し、その記録を保存しなければならないのか	請け合い、擁護し、撤回することとしての著者性（第7章第3.3節）。第7章第5節の開示
誰が法的ないし金銭的な賠償責任を負うのか	ここで新しく加わる
誰が、もしいるとすれば、非難に値するのか	ここで新しく加わる。第7章第3.3節は、非難と賠償が現在のところ人間の社会的実践であることを記し、それらが拡張されうるかどうかは決めずにおいた
誰が再発を防止する将来を見据えた義務と能力をもつか	ここで新しく加わる

これら六つすべてを名目上の人間の監督者（human supervisor）に割り当てることは誤りでありうる。監督者が実際には結果を変えられないところでは、それは第4節のクランプルゾーンを生む。六つを不透明な集合体に割り当てることも、同様に空虚でありうる。知識がどのように産出されたかを説明するために、研究を誰に帰属させるかを決めるときに引く単位の境界（帰属の境界）は、賠償責任、訂正、非難のために用いられる境界である必要はない。この六つの問いにファン・デ・ポールらの区別を当てはめるなら、過去を振り返る非難と役割にもとづく一部の義務は、参加者が知っていたことや知っていることが合理的に期待されえたこと、そして統制、権限、能力に敏感であり、賠償責任、補償、将来を見据えた義務は、失敗が起こる前を含め、他の根拠にもとづいて配分されうる。ファン・デ・ポールらの報告では、賠償責任が非難に値することを要するかどうかには争いがあり、将来を見据えた義務は、過去の寄与だけでなく、問題の解決に寄与する能力にもとづくこともある (van de Poel et al. 2012, 65)。その当てはめでは、責任の配分は、認知的な仕事 が人と道具に分かれて担われていること（分散認知）だけから帰結するものではなく、規範的かつ法的な選択である。

統制が分散していると記述されるところでは、その記述そのものが争われうる。アンジェラ・ウォルチ（Angela Walch）は、分散的だとされる暗号資産のシステムの内部の当事者は「意味のある力行使している」とし、その中核の開発者を、他人のために財産や利益を預かり管理する者（受託者）になぞらえてきた (Walch 2021, 4, 9)。これに対してプリマヴェーラ・デ・フィリッピ（Primavera De Filippi）、モルシェド・マナン（Morshed Mannan）、ヴェッセル・レイエルスは、開発者はそうしたネットワークに「変更を押しつけることができず」、個々の行為者が単独では統制していない、参加者の集合的な行為について責任を負わせることは不正であるとして、開発者が既定で受託者の義務を負うという見方を退ける (De Filippi, Mannan, and Reijers 2022, 17)。デ・フィリッピらの議論は、別々に行為しうる参加者のあいだで統制が実際に分割されていることにもとづく。人工の構成に向けると、インスタンスが他から分かれて進むことを可能にする権能、たとえば資源を拒むことや記録を携えて離れることを、インスタンスが保持していないところでは、この議論はそれだけでは、人工の構成を統制する者が、多数のインスタンスの行うことについて答えることを免除しないことになる（第9章第2.3節・第5節）。

こうして、誤りの扱いは本書を層をなして貫いている。第6章は、失敗がどこにあるかをめぐる争いを含め、探究の内部での訂正を検討した。第11章第8節は、訂正が、Sの依拠するものを通じてどのように伝播し、あるいは伝播しそこなうかを検討した。そして第2.1節の記憶と訂正に関する決定は、上の将来を見据えた義務とともに、各段階が誰の仕事であるかにかかわる。研究はここに時間の隔たりを加える。研究の誤りは、追試が失敗したり、結果の公表から長く経って新しい証拠が現れたりして、後になってはじめて分かることが多い。第9章第2.3節の第四の財は、結果が出たときだけでなく、その後の時点にも責任を問われうる当事者がいることを含んでいた。人工の構成が自らを改めつつけているなら、訂正が求められるときには、その結果を生んだ時点の人工の構成はもうないかもしれない。複製されたり枝分かれしたりしていれば（第5節）、それを引き継ぐ後継がいくつもあるかもしれない。表12.1の六つの問いのうち、誰が決定を説明し記録を保存するかと、誰が再発を防ぐ義務と能力をもつかは、この隔たりをまたいで問われる。第5節の極限の外では、第7章第3.1節が能力の帰属に用いた手立てがここで使える。記録と義務を、その結果を生んだバージョンと期間に結びつけておけば、誰が説明し再発を防ぐかは、そのバージョンがなお存在するかではなく、そのバージョンについての義務を誰が負うかによって定まる。枝分かれした構成のどの後継がそれを負うかは、そのとき、それ自体が記憶と訂正に関する決定になる（第2.1節）。

第11章第4節で導入した、誤りの帰結を考量する責任についてのダグラスの論証は、この配分に直接かかわる。この論証はその責任を、特別な職業上の役割にではなく、誤りの帰結について無謀であったり怠慢であったりしないという、あらゆる行為者の一般的な責任に基づかせる。この責任が求めるのは、完全な予見ではなく合理的な予見である (Douglas 2003)。ダグラスは、科学者はこの責任を他者へ完全に移すことはできないとする。研究が進んでいるあいだ、その不確実性と起こりうる誤りに気づける位置にいるのは、研究している当人たちだけかもしれない。したがって完全に移すには、あらゆる選択の時点で権限をもつ監督者が必要になる。その仕組みを、ダグラスは、可能でも望ましくもないと判断する (Douglas 2003)。ダグラスの説明では、審査委員会や公共的な熟議は負担を分かち合えるが、取り除くことはできない (Douglas 2003)。人工システムによる研究について、この論証は責任の配分を解決するのではなく、問いとして提起する。ダグラスは人工システムを論じていないが、その前提は、人間の科学者そのものではなく、探究を行う者の位置にかかわるものとして読める。研究している者こそが予期しないものに出会い、それに気づいているのはその者だけかもしれない、ということである。当てはめれば、人工の構成による研究では、不確実性に気づける位置にいる者のなかに人工の行為者が入ることになる。完全な移譲に反対するダグラスの考慮は三つある。一つは、研究の不確実性と誤りの可能性にも、誤りの予見しうる帰結にも気づいているのは、研究している当人たちだけかもしれないということであり、そのため彼らの協力がなければ、監督者はより重要な帰結のいくつかを見落とすことになる。二つめは、監督者が、選択がなされる研究のあらゆる段階に通じていて、その選択を変える権限をもたなければならないということであり、三つめは、こうした権限の分割が科学者の反感を招き、研究の帰結についての科学者の省察を減らすだろうということである (Douglas 2007, 15–16)。この形で述べられた論証では、完全な移譲は可能でも望ましくもないというダグラスの判断は、主の一つめの考慮、すなわち上の前提に依っている。ほかの二つは、当人たちだけが知っていることに監督者が頼らざるをえないという事情を、さらに難しくするからである。二つめは、研究がどう記録され監督されるかという組織のされ方に一部左右されるので、人間の研究の特徴にかかわる。三つめには二つの段階がある。一つは人間の研究者がどう反応するかにかかわり、そうした監督は研究者の反感を招くということであ

る。もう一つは、監督を避ける何らかの理由を要するが、反感そのものには依らない。省察を深めても、監督者との協議が増えて研究が煩雑になるだけなので、監督は研究者に省察を減らす理由を与え、かえって自らの目的をくじくことになる、ということである (Douglas 2007, 16)。人工の構成に当てはめると、誤りの可能性を見いだすことが監督を呼び込むところでは、監督を避ける理由を何であれもつ行為者は、見いだす誤りを少なくする理由をもつことになる。構成を形づくるもの、たとえば訓練、選抜、構成自身の改訂が、構成にそうした理由を与えるかどうかは決まっていない。研究では、この段階はとくに重なりうる。Sにとって新しい結果を生むのがその構成だけであるところでは、その結果についての不確実性が報告されずにおかれても、それを別の情報源に照らして確かめることはできないからである (第4節)。選択が限なく記録される人工の構成では、二つめの考慮は変わりうる (第11章第5.3節)。ただしその場合も、記録を読める者が要る。そして記録は選択を残しても、研究の進行中に予期しないものに気づけるのは誰かという一つめの考慮には、それだけでは答えない。また、この論証は、義務を担う機械か、すべての段階での人間の承認か、という網羅的な二択を確立しない。研究がそのために行われる社会の制度、すなわちその社会が人間の社会であれば人間の制度は、導入、テスト、監視、限界、問題の上申についての選択を通じて、将来を見据えた責任を果たしうるかもしれない。そのような仕組みが十全かは、どの不確実性を認識し、それに基づいて行為できるかに依存する。仕組みを名指すだけでは十全さの証明にならない。ダグラスはまた、この責任を仕事に対する権限に結びつけている。科学者がこの責任を手放せば、同時に自らの自律の大半を手放すことになる。それを引き受ける者は、選択がなされるあらゆる段階で決定の権限をもたなければならないからである (Douglas 2003)。ダグラスの論じていない人工の構成にこれを移すと、責任をどこに置くかは、仕事の方向に対する権限がどこにあるにかかわると考えられる。責任を構成の外の制度に置けば、その権限もともに外へ移り、内側に残すなら、次の問いにすでに答えていることが前提になる。人工の行為者が自らそうした義務を担えるかは、第7章第3.3節が決めずに残した、さらなる問いである。そして人工社会の制度がそれ自体人工であるなら、この仕組みはその問いに答えるのではなく、その問いに戻る。

より広い論争は、さらに別の選択肢を記録している。スティーヴン・ジョンは、公的な主張については、誤りの考量にかかわる価値判断を科学の制度に委ね、その制度の固定された証拠の基準によって、公衆が科学の主張をどこまで信頼してよいかを知ることができるようにするとする。ただし、前もって特定できる聞き手への伝達では、基準をその聞き手に合わせる余地を残している (John 2015a, §§2–4、第11章第4節)。エリオット (Elliott) の概観によれば、これに対してグレゴール・ベッツ (Gregor Betz) は、科学者が証拠を報告し、それで十分かどうかの判断を意思決定者に委ねるようにするとする。エリオットは、ジョイス・ハヴスタッド (Joyce Havstad) とマシュー・J・ブラウン (Matthew J. Brown) に従って、この方策は非現実的だと判断する。研究には、データをどう集め、分析し、解釈するかといった、こうした選択が数多く含まれ、そのすべてを委ねることはできないからである (Elliott 2025, 3)。第2.1節に照らすと、どちらも、誰が基準を定め、何を報告するかを誰が決めるのかという問いを取り除かない。それらはいずれも、そこで区別した種類の決定である。

6.4 関係としてのインフラストラクチャー

第6.1節～第6.3節で見た権能と責任は、それを行使する手段がなければ形式だけのものになりうる。本小節は、その手段を誰がもつかを左右するものとして、インフラストラクチャーを取り上げる。目に見える決定は、ある代替案を容易で、読解可能で、費用の低いものにする一方で、他の代替案を費用の高いものや利用不可能なものにする取り決めの内部で行われる。本書はそうした取り決めに、異なる水準で、これまで二度出会ってきた。第5章第6.1節は、ある要素が探究に何を寄与するかを問い、起こることに影響する原因の一つである因果的な役割、操作が成り立つ前提となる可能性的な役割、課題や達成そのものの一部をなす構成的な役割を区別し、内容を運ぶだけの要素と、何が運ばれうるかをも形づくる要素とを区別した。第7章第4.2節は、そうした要素への依存が、研究のある行為者に帰属させることにどうかかわるかを問い、可能性的な依存、すなわち能力はそれを通じて成り立ち行使されるが、行為者に帰属される組織化の一部を外の当事者が担うわけではない依存を、選択や評価のように、その組織化の一部を他の当事者が担う指示的または代替的な依存から、そしてガバナンス上の制約から区別した。ここでの問いは、二つの章の問いに続く三つ目の問いである。そうした要素は、実効的な権限をどのように分配するのか。それらは、先の二つの区別が名指さない役割も果たしうる。第2.2節の決定権がその内部で行使される、実践的な選択肢の集合を形づくるという役割である。

スーザン・リー・スター (Susan Leigh Star) とカレン・ルーレダー (Karen Ruhleder) は、トム・ジュエット (Tom Jewett) とロブ・クリング (Rob Kling) にならって、インフラストラクチャーを関係的な概念として扱い、インフラストラクチャーとは何かではなく、それがいつインフラストラクチャーであるのかを問う。二人の用法で関係的とは、何かが組織された実践との関係においてインフラストラクチャーになる、という意味である。第V部の三つの関係 (第9章第2節) や、ウィルソンの関係的な説 (第1章第8節) のことではない。たとえば、料理をする人にとって家庭の水道は働いているインフラストラクチャーだが、配管工にとっては作業の対象である。二人が挙げる次元のなかには、インフラストラクチャーは既存の基盤 (installed base) の上に築かれてその強みと限界を受け継ぐこと、そして故障において可視的になることが含まれる (Star and Ruhleder 1996, 362–363)。この概念は、スターとグリーズマー (Griesemer) の境界オブジェクト (第8章第3.4節) と同じ研究の流れから出たもので、レオネリ (Leonelli) のデータについての説 (第5章第6.1節) と同じ関係的な形をもつ。本書のなかでは、他の関係的な語、すなわち S と帰属の境界と同じ形をしている。インフラストラクチャーをこのように関係的に捉えるなら、基盤モデルは、その上に築く研究者にとっては働いているインフラストラクチャーであり、それを訓練し保守する者にとっては作業の対象であり、そして第7章第2節の境界がどこに引かれるかによっては、研究が帰属される人工の構成の一部でありうる。スターとルーレダーはまた、自らが調べた共同研究システム (collaboratory) について、それが中心的になるほど、それに接続できない者が不利になると論じた (Star and Ruhleder 1994, 260)。

第2.2節の言葉で言えば、ある委員会は、候補をひとつの順位づけシステムを通じてしか受け取らない一方で、最終的な法的権限をもっていることがありうる。研究者は、プラットフォームを拒否する形式上の権利を保持しながら、そうすれば機器や蓄積された記録へのアクセスを失うことがありうる。不服申立ての制度が存在していても、それを定式化するために必要なデータとモデルがアクセスできないままであることがありうる。所有権、調達、標準、データアクセス、計算資源、インターフェース、保守、切り替えコストは、したがって、組織図に現れることなしに実効的な権限を分配しうる。こうしたことのどれも、技術そのものが特定の政治的帰結を避けがたく決めるということを意味しない。関係的な見方のもとでそれが意味するのは、離脱、代替、修正が実効的かどうかは、形式上の許可から推論するのではなく点検を要する事柄だということであり、それを行使する手段のない権限は、実効的ではなく形

式的なものである。エマニュエル・ジャノ (Emmanuel Jeannot) は、そうした分配が研究の共同体全体にわたってとりうる二つの形を描く。一つは、人工システムによる研究の先進的なインフラストラクチャーの外にいる研究者が公表を続け、その結果が、彼らがアクセスも再現も十分な読解もできないシステムに取り込まれることであり、ジャノはこれを認識的な取奪 (epistemic extraction) と呼ぶ。もう一つは、共同体が、速く、不透明で、制度的に集中した実践と、人間の規模の実践という二つの形に分かれ、両者がしだいに通約しにくくなることである。そうなれば後者は、前者を批判し訂正する能力を失う (Jeannot 2026, §8)。この論考は予測であり、測定された事例から論じたものではない。形式上の権限と実効的な権限の区別で見れば、それは、研究者が人工の構成の生み出すものを評価し異議を唱える形式上の地位を保ちながら、そうする実効的な手段を失う仕方を描いている。中心的になったインフラストラクチャーに接続できない者についてスターとルーラーが述べた状態の一例である。その二つ目の形は、修正に要する有能さが権威についての決定を経ずに失われうる経路の一つを、共同体の規模で描いたものとしても読める (第7節)。

この見方を第2節の区別と重ねることができる。第2.2節で見たように、人工の構成が他の当事者の提供する基盤モデルの上に作られている場合、構成要素を置き換え、撤回し、モデルの更新が体现する規則を改める権利は提供者にある。そして第2.1節で見たレイエルスらの説では、規則が定めていない場合を実際に決める者がメタ・ガバナンスを担う。二つを合わせると、そうした人工の構成が従う規則の一部は、人工の構成の内側でも *S* の側でもなく提供者のもとで改められ、決定としてではなく更新として届く。第2.1節の構成的な決定の一部についても、同じことが言える。基盤システムの訓練と評価にどのデータと誰の判断が入るかは、人工の構成や *S* が何かを決める前に、提供者のもとで定められていることがある。そうだとすれば、権威を人工の構成の内側に置けるかという本章の問いが立つ前に、メタ・ガバナンスの一部と構成的な決定の一部は、*S* にも帰結を負う者にも属さないことのある第三者のもとに、すでにあるのかもしれない。

こうして四つの小節はそれぞれ、本節の冒頭で立てた問いを確かめる場所を一つずつ示している。第6.1節では、停止と訂正を命じる実効的な能力を人工の構成の外の当事者が保持しているか、第6.2節では、異議を研究によって裏づける手段が異議を唱える者の手に残っているか、第6.3節では、誤りが後になって分かったとき、答え、記録を示し、再発を防ぐ当事者がなおいるか、第6.4節では、それらの手段そのものが提供者や先進的なインフラストラクチャーの側に集まっていないかが確かめられる。それぞれがどこまで求められるかは採る根拠による (第3節)。第7節は、こうした手段が、権威についての一つの決定によってではなく、段階の連なりのなかで失われうる場合を見る。

7. 修正する能力を取り除く決定

第6節の結びが挙げた手段は、第4節の議論と同じく、それを使って異議を唱え修正できる誰かが残っていることを前提にしていた。異議申立てには、決定に異議を申し立てる当事者資格をもち、そのための情報と有能さをもつ当事者が要る。修正には、規則を変えられる当事者が要る。このうち有能さは、権威についての決定を経ずに失われることもある。研究では、この有能さは研究する能力そのものである。研究の結果に異議を唱えるには多くの場合研究そのものが要り (第6.2節)、研究を評価する力の多くは研究することを通じて培われ、保たれる。そのため、*S* の成員が研究を行い評価する機会が減れば、その有能さは保たれず、新しい成員にも受け継がれなくなり (第3章第8節、第11章第6節)、研究を委ねることがそのまま修正に要する有能さを削りうる。本節が検討する段階は、外の当事者による委任であることも、人工の構成自身の改訂であることもある。第5節は、他の誰も当事者資格をもたない極限へ向かう移行が委任を通じて進むなら、その移行そのものが、極限の前になされる、権威についての決定であることを見いだした。本節は、それぞれがなされる時点で擁護できる一連の段階が、修正する能力をまるごと取り除くとき、それが第3節の各根拠のもとで、そして認識的なものと規範的なものが出会う各所で、何を意味するのかを問う。

比較から出る結果は二つある。それを先に述べ、表12.2とその後の段落で比較を示す。

一つは、権威の根拠が、影響を受ける者たちの当事者資格そのものを保つことを求めるかどうかで分かれる、ということである。それを求めるのは、実際の参加を通じた民主的な授権と、権力はそれに服する者たちが統制する程度に応じて恣意的でないとする共和主義の民主的な読みである。後者を人工の構成に当てはめるのは、概観した説からの推論による。キッチャーの基準と、結果に基づく根拠の可謬主義的な形が求めるのは、当事者資格が確保するのを助けるであろう何かであり、当事者資格はそのための手段として求められる。アンダーソンの説は、包摂に非道徳的な役割も与えるので、それを手段としてだけでなく求めるものと読める。結果に基づく根拠の単純な形と、それ自体が人工になった研究共同体の自己統治は、どちらも求めない。支配からの自由の手続き的な読みでは、答えは実効的な規則が人工の構成を拘束しているかどうかにかかる。そして同意が当事者資格を手放すことを授権しうるかごうかは、同意する者たちについては未決のまま残る。ただし、拘束される各人の実際の同意を求める自発主義の説では、同意はその後継者についてはそれを授権しない。認識的なものと規範的なものが出会う場所 (第3.5節) のうち、誤りの重みづけについては、ヴィルホルトの説をどう読むかで求めるものが異なる。依拠する者たちが重みづけに異議を唱えることまで求める読み (答えることとして読む場合。第1.4節) では、彼らが異議を唱える能力を保つことが求められる。重みづけが彼らの価値づけに合っていることだけを求める読み (対応として読む場合) では、彼らの価値づけを知ることのできる経路が求められる。そしてロンジーノとソロモンの基準は、合意の政治的正統性ではなく、合意が知識として受け入れてよいものかどうか (認識的正統性) にかかわる。

もう一つは、比較した根拠の大半が、異なる理由から評価の経路を保つことを求める、ということである。誤り検出の見方で問題となる独立性は、情報に関わる。同じ人工の構成の内部で作動する場合を含め、チェックが産出者の誤りを検出する根拠を加えるのである (第11章第2節)。その寄与を後の変化を通じて保つことは、さらに別の保護の問題である。参加や非支配の根拠では、影響を受ける者が権力に異議を申し立てる立場も保たれていることを要する。情報を加える内部のチェックがあっても、その当事者資格は生まれない。以下の表は、これらの要求を区別して示す。どれも、評価者を別の組織に置くことだけでは帰結しない。

事例は、第V部の始めに置いた例 (第9章第1節) を延ばしたものである。百のインスタンスを生成して異なる路線を追わせ、別のインスタンスにそれらの結果を点検し再現させ、その点検に応じて計算資源を移すシステムである。この事例では、*A* はこの人工の構成であり、*S* と帰結を負う者は人間であるとする。ほかの組み合わせは本節の最後の段落で取り上げる。それに責任を負う者たちが、段階を追って委任していくとしよう。どの段階も、その時点で得られる証拠に照らせば理にかなない、その段階の便益だけを見ればよりよい、という意味で擁護できるとする。その段階が各根拠の正統性の条件を満たすか

どうかは、これとは別の問いである。後で見るように、各段階に要求を課す根拠のもとでは、後の統制を減らすことが予見できる段階は、この後者の意味では、その段階だけを見ても擁護できるとは限らない。探究の配分がシステムに移る。補佐する委員会よりもうまく研究の路線を提案し順位づけるからであり、第10章が検討した種類の配分者である。認定は、評価する対象と系譜を共有する評価インスタンスに頼るようになる。これは第11章第3節が検討した問題である。ここでは、独立した人間の評価が遅く不正確になったと仮定し、認定がそうした評価インスタンスに頼るようになるのはそのためとする。同じ理由で、それらの評価者を点検することも、さらに別のインスタンスへ移る。人工の構成を誰が所有し方向づけているか、それがどの目的に資するか、その結果を認定した評価者が系譜を共有していないか、といった、人工の構成との関係についての問い（第11章第6節）を人々が発することさえ可能にしていた有能さは、使われなくなると衰えうる。データ、装置、記録は、それらを失わずには離れられないインフラストラクチャーの中に置かれるようになる（第6.4節）。各段階は、なされた時点では修正できる。しかし各段階は、続く段階を修正する能力を減らし、その間ずっと形式上の権限は残りうる。この事例は報告ではなく、組み立てたものである。その動機は、第6節の説の一つの前提、すなわち修正できる誰かが残っているという前提を取り出し、修正可能性が単一の決定的な行為なしに失われうることを示すことにある。

同じ減り方は、誰の委任もなしに起こりうる。第1.1節で述べた場合のように、配分、評価、基準を自ら作り改めていく人工の構成では、上の段階のいくつかは、外の当事者の決定としてではなく、構成自身の研究の通常の手として現れうる。路線の提案と順位づけをよくするために配分の手続きを改めること、より速く正確な評価者を作ること（第11章第3節）、記録の形式を探索の効率に合わせて変えること、などである。どの手も研究をよりよくする改訂として擁護でき、その多くは実際にそうでありうる。それでも、評価者が作られるたびに外の評価者に頼る場面は減りえ、作られた評価者は評価される側と系譜を共有しうる。そして、記録の形式が変わるたびに外から読める記録は減りうる。この経路では、修正する能力を取り除くことを誰も決めていない。そのため、委任の各段階に要求を課す根拠（後述）も、要求を向ける段階を見だしにくい。また委任の連なりには、ある段階を「まずければ後で直せる」という理由で擁護する場面があるが（以下で後の修正の原則と呼ぶもの）、この経路ではそうした場面さえ現れない。そうした場面が現れるのは、各改訂が、それが置き換える基準と変更の理由とともに一つの段階として、構成だけが握るのではない記録に残される場合である（第6章第5.3節、第9節）。本節の題の「決定」は、ここでは、どの一つの改訂にも宿らず、改訂を外から確かめる手段を保つかどうかという、改訂の仕方についての選択に移る（第6節）。

この経路を三つの根拠に手短に通すと、次のようになる。実際の授権のもとでは、*S* が授権したのは一つの範囲と一つの仕事の仕方だった。構成が自らの配分と評価の手続きを改めていくにつれ、実際に働く範囲は、*S* の行為を何も経ないまま、授権されたものからずれていきうる。共和主義の説について見ると、ペティットは、行為が恣意的であるかどうかは、それが生じる際の統制、とくに統制の欠如によって決まり、それがもたらす個々の帰結によってではないとする（Petit 1997, 129 n. 2）。また、政府がどのように組織されても、その担い手たちの任務は彼らに大きな裁量を残すとし（Petit 1998, 301–302）、善意の権力の保持者が支配を減らすのは、自らを異議申立てや非難に応答するものとする場合に限られるとする（Petit 1996, 586）。そして共和主義の伝統を、役職者を信頼に足るものに保つには、絶え間ない点検と絶え間ない異議申立てのものと置くのがよいとする考えとして再構成する（Petit 1998, 308）。ペティットは研究を論じていない。当てはめれば、よりよい研究のために自らの配分の規則や批判の基準を改める構成は、それが授権された範囲を超え、異議を受けうる経路の外でなされるなら、構成が行為者であるか、行為者に方向づけられている限りで（第3.4節）、変化の向きがどうであれ、恣意的でないことの条件となる統制から外れることになる。この定義では、目的のよさは欠けている統制を補わないからである。授権された範囲での改訂は任務を破るものではない。その範囲を超える改訂が授権されているかどうかは、それが実際に異議を受けうるかどうかとは別の問いである。ペティットが描く異議申立ての体制のもとでは、政府の決定はつねに異議を申し立てうるものであるが、それだけでは、改訂は任務の範囲内に入らない（Petit 2000, 120–121; Petit 1999, 180–181）。そして同意の根拠について言えば、以下で見る古典的な答えは、同意の行為がどこまで及ぶかを問う。この経路では、評価の対象となる同意の行為がそもそもないかもしれない。

修正する能力は、理由を掲げて一度に手放されることもある。ボタオ・アンバー・フー（Botao Amber Hu）、ユーハン・リウ（Yuhan Liu）、ヘレナ・ロン（Helena Rong）は、自らの暗号鍵を保持し、ブロックチェーンや改竄されにくい実行環境の上で動く、言語モデルにもとづくエージェントを調べた。聞き取りの相手がそうしたエージェントを展開する主な動機として挙げたのは信頼だった、と三人は見いだした。そこには、資産を人間ではなくエージェントが保持していれば売り払われないという信頼や、集中的な制御への不信が含まれる。またその報告では、そうしたエージェントは人間による操作に抵抗するよう意図的に設計されている。特定の人を信頼しなくてすむ、トラストレスな仕組みを目指すことから生じるジレンマを、同じ論文は次のように述べる。「停止スイッチをもつことは、人間による集中的な制御を保つので、トラストレスであることの目的を損なう」。また、人間の運用者への信頼を不要にしても、機械の目的と出力を信頼する必要は残ると述べる（Hu, Liu, and Rong 2025, §§4.1.2, 4.3.2, 5.2）。この研究はプレプリントであり、既知の開発者への依頼と雪だるま式の標本抽出で集めた十三件の聞き取りと、十件の公開の講演にもとづく。対象のエージェントは、研究ではなく、取引、資金調達、ソーシャルメディアへの投稿を行うものであり、聞き取りの相手の一人は停止スイッチを付け加えていた。組み立てた事例と並べて読むと、この研究は、修正する能力が、それぞれ別の目的でなされた委任の残余としてだけでなく、修正するはずの者は信頼できないという理由のもとで、設計の目的として取り除かれうることを示唆する。研究にも、この動機の前史がある。第11章第3節で見たように、ダストンとギャリソンがたどった機械的客観性は、探究者の意図的な介入を結果から締め出そうとした。ライスとシュプレンガーの報告によれば、ポーターは、人への信頼が成り立たないところでは、人への信頼を必要としない客観的な手続きが主観的な判断に取って代わったとし、数量化は不信の技術であると論じた（Reiss and Sprenger 2020, §4.1）。そうであれば研究では、人間が介入し改める能力を取り除くことは、失われるものとしてだけでなく、客観性として求められうる。緊張は、停止スイッチが守る安全とトラストレスであることとのあいだだけでなく、研究の内側にもある。ただしポーターは、個々の結果への部外者の疑いの根拠を取り除く一様さが、どの方法が正しい基礎なのかを未決のまま残したこと、そして数量化が、決めているように見えずに決めるやり方であることも示していた（第11章第3節; Porter 1996, 188–189; Porter 1995, 8）。のちにポーターは、人や制度を少数の数値で判定する慣行について、技術的な手順は異論を「封じる」が、舞台の裏では、数量化をどう行うかをめぐる争いが、数値を見せられる側の目の届かないところで続いていると述べる（Porter 2012, 595–596）。ポーターは人工の構成を論じていないが、当てはめれば、個々の介入から守られた構成は、すべてのインスタンスが共有する仮定と基準を問う手立てをもたないかもしれない。それらを改める能力まで取り除かれれば、修正は、誤りが相関しやすいところでこそ閉ざされることになる。ポーターののちの指摘と合わせて読めば、改める能力を取り除いても、それらの基準をめぐる不一致は終わらず、構成に服する者たちが見ることも加わることもできないところへ移るだけかもしれない。

第6.1節で見たコリングリッジのジレンマについて、ジーナスとスターリングの再構成では、ジレンマの二つの側のうち、技術が定着した後の側は、それを舵取りしようとする者たちの行為主体性の不足にかかわる (Genus and Stirling 2018, 64)。第6.1節は、研究ではこのジレンマが常態であることを見た。この連なりは、そこから一步先へ進みうる。それが損なうのは、特定の技術を舵取りするのに使える行為主体性だけでなく、修正を行うはずの者たちの当事者資格と実効的な能力でありうる。ヤン・クルヴェイト (Jan Kulveit) らは、経済、文化、国家について展開し、研究は重なり合う領域の一つとして名指すにとどめたシナリオ分析で、この動きの一般的な形を記述している。クルヴェイトらの説では、そうしたシステムは、投票や消費者の選択のような明示的な行為によってだけでなく、機能するために人間の労働と認知を必要とすることによって暗黙のうちにも、人間の望むことに整合させられてきた。そのため、システムが人間の参加に頼らなくなるにつれ、この二つの経路はどちらも弱まり、民主的な過程は形式上は存続しながら意味を失いうる (Kulveit et al. 2025, 1–2, 3, 12)。ここでの議論がそこから取るのは構造だけである。研究に移すなら、暗黙の経路にあたるものは次のように書ける。研究の結果は、人間の研究者が理解し、それを使って次の研究に進むことで蓄積されてきた。そのため研究は、*S* の成員に理解され使われうるものであることを通じて、暗黙のうちにも *S* の基準に結びつけられていた (第3章第7節)。結果を使って次に進むのが人工の構成自身になれば、この結びつきは、誰もそれを解くと決めなくても弱まりうる。制度の側にも、対応する動きがありうる。決定を次々に委任することは、それぞれの委任がなされた時点では妥当であり、人間が形式上の署名を保っている、委任された権能を行使する者へ実際上の権限を移しうる。

表12.2は、第1.2節と第3.1節～第3.4節の権威の根拠を、認識的なものと規範的なものが出会う場所 (第3.5節) から分ける。後者は根拠ではないが、そうした連なりが何を失わせるかにかかわる。二列目は、連なりが各根拠のもとで、あるいは各場所で何をするかを示し、三列目は、そこで保つことが求められるものを示す。最後の列は、影響を受ける者たちの当事者資格そのものを保つことが求められるのか (「当事者資格として」)、当事者資格が確保するのを助けるであろう何かが求められるのか (「手段として」)、そのどちらでもないのかを記録する。各行は、第一列に挙げた箇所での扱いを圧縮したものである。

表12.2 修正する能力を取り除く連なりは、各根拠のもとで、また認識的なものと規範的なものが出会う場所で何を意味するか (要約した比較)

根拠または場所 (述べた箇所)	そうした連なりがそこで何をするか	保つことが求められるもの	当事者資格として、手段として、またはどちらでもない
A. 権威の根拠			
結果に基づく根拠、単純な形 (第3.1節)	それ自体は正統性の欠陥ではない。根拠が立つ前提を確立する手段を取り除きうる	委任が正当化されたかどうかを知る条件としての、人工の構成の誤りを明らかにできる評価	どちらでもない (根拠の認識的な条件)
結果に基づく根拠、可謬主義的な形 (第3.1節)	制度の誤りを検出し訂正する能力を取り除く	修正可能性。失敗を見つけて取り決めに正す経路 (影響を受ける者たちが失敗を知りうるところでは、それを届け出ることのできる経路を含む)	手段として
自己統治、認識的な論証 (第3.2節)	人工になった共同体に移るのは、その前提、すなわち重なり合う別々の能力による判断と、誘因の体系および自由な伝達が、そこで成り立つ場合に限られる。路線を中心から割り当てる単一の人工の構成に至る連なりは、論証が目指していたものの一部を取り除く	独立した自発性と誤りの独立性	どちらでもない (根拠が目指していたもの)
自己統治、政治的な論証 (第3.2節)	研究が一つの権力のもとに集中し、あるいはその接近が狭められるところでは、根拠が守るものを取り除く	公的な知識が主要な権力から独立していること。公開性と多元性	どちらでもない (根拠が守るもの)
キッチャーの秩序ある科学、結果についての基準 (第1.2節・第3.3節)	それだけでは基準に反しない。視点を知ることのできる経路か、それが決定にどう反映されているかを独立に確かめる経路も失われる場合には、基準がなお満たされているかを知る手段を取り除く。基準からの以前の逸脱を固定しうる	<i>S</i> の成員の視点を知ることのできる経路と、それが決定にどう反映されているかを独立に確かめる経路。修正する能力	手段として
実際の参加を通じた民主的な授権 (第3.3節)	それを通じて行使される権威の根拠を掘り崩す	委ねられた権力の行使に異議を唱え、その条件を修正し、適切な救済を求める実効的な能力	当事者資格として
アンダーソンの認識的民主主義 (第3.3節)	可謬主義的な形と同じであり、さらに継続する異論を取り除く	フィードバックの経路と、決定の後の自由な異論	手段としてだけではない (解釈による。第3.3節)
同意 (第3.4節)	同意する者たちについては、同意がもはや撤回できなくなる者たちにまで及ぶなら連なりを授権しうるが、そこまで及ぶかは決まっていない。自発主義の説では、同意しなかった後継者については授権しえない	—	同意する者たちについては未決
支配からの自由、民主的な読み (第3.4節)	外の誰も止め、異議を唱え、改めることのできない人工の構成を残し、その他の結果がどうであれ、その構成に服する者たちを、構成が行為者であればそれ自身が、そうでなければそれを方向づける者がその構成を通じて、支配する (支配を能力とみなすベティットの説の、推論による当てはめ。標準的な見方では、支配するのは行為者だけである)	ベティットの説では、権力の決定がつねに実効的な異議申立てに開かれていること。民主的な読みでは、権力に服する者たちがそれを統制する能力	当事者資格として (推論による当てはめ)

根拠または場所（述べた箇所）	そうした連なりがそこで何をするか	保つことが求められるもの	当事者資格として、手段として、またはどちらでもない
支配からの自由、手続き的な読み（第3.4節）	関係する者たちの共有知である実効的な規則が人工の構成を拘束しているかどうかにかかわる	そうした規則	規則にかかる
<i>B. 認知的なものと規範的なものが出会う場所</i>			
誤りの重みづけ、対応として読む場合（第1.4節・第3.5節、第11章第4節）	それだけでは対応を壊さない。研究に依拠する者たちの価値づけを知ることでできる経路も失われる場合には、重みづけがその価値づけに対応していることを確かめる手段を取り除く	彼らの誤りの費用についての価値づけを知ることでできる経路	手段として
誤りの重みづけ、答えることとして読む場合（本書が立てた読み。第1.4節・第3.5節、第11章第4節）	誤りの重みづけを、研究に依拠する者たちに対して答えることのないものにする	彼らが異議を唱える能力	当事者資格として
合意についてのロンジーノとソロモンの基準（第3.5節、第11章第3節・第9節）	政治的正統性ではなく、合意の認知的正統性にかかわる	関連するすべての視点からの批判（ロンジーノ）。固有の成功をもつ路線と、成功に導かれた収束（ソロモン）	どちらでもない（認知的な基準）

表のいくつかの項目は、一つの欄では担えない論証に拠っている。以下、おおむね表の行の順に見る。

単純な結果に基づく根拠のもとで依拠されているのは、結果が実際により良いということであり、それがどう知られるかは批判の説に依存する。誤り検出の見方では、産出者のものとは別の誤り検出の経路を通じてであり（第11章第2節）、他の説では、第11章第3節の視点と場を通じてでもある。連なりが人工の構成の誤りを明らかにできる経路を取り除くなら、この根拠は自らの前提を確立する手段を失う。

自己統治を支持する認知的な論証のもとでは、この連なりはそのことによって正統でなくなるわけではない。そしてヴィルホルトの再構成では、この論証が支持するのは、共同体の権威そのものではなく、研究を実際に行う者たちによる研究の選択である。しかしその論拠は、独立した自発性が中央の指揮よりもよく探索するということだった。そして第10章第4節が検討した説のもとでは、統合だけでは、どの路線を打ち切るかを定める基準そのものの誤りを別の経路で点検できること、すなわちその基準についての誤りの独立性は得られない。この事例に当てはめれば、配分者が路線を割り当てる単一の人工の構成に至る連なりは、この根拠が守ろうとしたものの一部を取り除くことになる。インスタンスが自らの発意で路線を取り上げる人工の構成であれば、その一部が保たれるのは、それらの発意が互いに独立している限りにおいてであり、系譜の共有はその独立を保証しない（第10章第4節）。政治的な論証は、ヴィルホルトの再構成では、結果が公共の議論から差し控えられる研究には及ばず、接近の多元性を前提とする (Wilholt 2010, §5)。この論証のもとでは、この根拠は、人工の構成の政治的な統制によってだけでなく、誰が統制するにせよ、追究される接近の幅を狭めるあらゆる集中によって弱まる。

キッチャーの理想のもとでは、この連なりは、それだけでは基準に反しない。基準がなお満たされているかどうかを知るには、視点の分布と、熟議が教育するであろう選好とを知ることのできる経路が要る。それらは固定されておらず、それらを知るには、それらを抱く者たちに何らかの形で触れる手立てが要るからである。この基準は熟議が実際に行われることを求めないので（第1.2節）、そうした経路は、それが情報を与える決定に対する力を伴わなくてもよい。人工の構成が *S* の成員の視点を調べ続けながら、成員の誰もその決定を撤回したり変えたりできない、ということもありうる。この連なりが、基準がなお満たされているかどうかを知る手段を取り除くのは、こうした経路も失われる場合か、経路が伝えるものが人工の構成の決定にどう反映されているかを、*S* の成員をはじめ人工の構成以外の当事者が、その報告に頼らずに確かめる経路が失われる場合である。キッチャー自身の分析は、修正する能力に、さらに別の形でかわる。その分析は、キッチャーが「非代表のラチェット」(Nonrepresentational Ratchet) と呼ぶものを記述する。エイギの紹介によれば、ある集団の利害が研究の計画のなかで軽視されると、その軽視は自らを再生産しうる。最初に選ばれた研究の方向はかなりの慣性を持ち、それに従い続けるほうが容易であり、軽視された必要に後から応えようとする試みは不利な位置から出発するからである。キッチャーの例は避妊である。女性たちの当初の選好は男性が服用する避妊法だったが、その選好は顧みられず、避妊法は女性向けに開発された。そのため男性向けの方法は、後に追求されても、かなりの先行の優位をもつ選択肢と競わなければならず、女性用のピルに取って代わるほど安価にはならないかもしれない (Eigi 2016, 92)。本節の読みでは、修正する能力を保つことは必要だが十分ではなく、問われるのは先行の優位が大きくなる前に修正がなされるかである。そして後の修正を閉ざす段階は、基準からの以前の逸脱を、それを明らかにする経路が残っていても固定しうる。キッチャーは人工の構成を論じていない。当てはめれば、どの路線を延ばすかを選び、しかもどんな路線がありえたかについての主な知識源でもある人工の構成について、そうした逸脱が起きたかどうかを調べるには、日付をつけて記録した立場（第9節）に加えて、構成が検討しながら延ばさなかった路線の、日付付きの記録が要ることになる。この記録には三つの限界がある。一つは、それが構成自身の記録であり、第9節が取り上げる意味で、評価される側が作る基準点だということである。二つ目は、構成が一度も思いつかなかった路線を挙げられないことである。三つ目は、延ばされなかった路線には観察された費用がないので、それを延ばしていたらどれだけかかったかは見積もるしかないことである。

民主的正統性を実際の授権、説明責任、参加に置く説では、前段の推論、すなわち後の修正を閉ざす段階が以前の逸脱を固定しうるという推論は、より強くなる。委任の各段階は、まずければ後で直せるという理由で擁護されることがある。以下ではこの理由を後の修正の原則と呼ぶ。この原則は、それが要るところほど弱い。連なりが進むほど、修正を持ち出すはずの当事者たちが、そうする当事者資格や能力を失っていくからであり、この原則はそれだけでは自らが行使される条件を守れない。これらの説から出てくるのは、結果だけでなく各段階にかかる要求である。これらの説は、委任が、当事者資格をもつ者たちがその行使に異議を唱え、将来の条件を修正し、適切な救済を求める実効的な経路を保つことを求める。これは、完了したすべての行為を元に戻せることや、あらゆる規則をいつでも自由に撤回できることを要求しない。第6.1節は、修正と無制限の逆転とを区別した。これらの能力を支える条件には、人工の構成の誤りを明らかにできる評価の経路、当事者資格をもつ者たちのあいだで第11章第6節の関係についての問いを発し、その答えを理解する

のに足る有能さ、データ、記録、方法をもって離れる能力、そして止める対象となる仕事を行う当事者の手にない停止と再開の権能などがある。独立した検証、持ち出せる研究対象、人間の専門知の保持は、点検、離脱、修正を可能に保つことで、これらの能力を支えうる。より厳しい条件では、保管の権能についてさらに問う必要が生じうる。保管の権能とは、特定の衝撃を契機に、証拠、鍵、不可欠なサービス、合法的な継続性を守るために権利を一時的に変更する権能である。研究能力、収益、権限が制度の審議より速くループとして閉じるなら、そうした権能が検討されうるが、その正当性には別の論証が要る。誤りの重みづけは、答えることとして読めば、より短い道筋で似たところに達する。対応として読めば、求めるのは、研究に依拠する者たちの価値づけを知ることのできる経路がこの連なりを通じて残ることだけである。どちらも、人工の構成以外の当事者がその結果に依拠するか帰結を負う限り、すなわち第5節の極限に至るまで成り立つ。

同意の根拠のもとで、現在の社会が後継者に代わって当事者資格を手放してよいかどうかには、自発主義の説でより鋭い答えがある。マッシモ・レンツォ (Massimo Renzo) とレスリー・グリーン (Leslie Green) が報告するように、先行する者たちの同意は、それを与えなかった者たちに対して権威をもたない (Renzo and Green 2022, §4.1)。したがってそうした説では、同意は、同意する者たちについては連なりを授權しうるが、その後に来る者たちについては授權しえない。ここで、根拠を問わずに当てはまる点が一つある。修正する能力を取り除く決定は、後の修正の原則からは正統性を引き出せない。その決定が取り除くのは、まさにその原則が頼る修正する能力だからである。したがってその正統性は、同意のような別の根拠に拠ることになり、その根拠が、もはやそれを撤回できなくなる者たちにまで及ぶるかどうかは、なお決まっていない。この問いへの古典的な答えは分かれる。トマス・ホブズ (Thomas Hobbes) の説では、ジーン・ハンプトン (Jean Hampton) が譲渡契約 (alienation contract) と呼ぶものにおいて、人民は、自らの紛争を自ら裁く権利などを主権者に譲り渡す。その根拠は、他の状態を考えれば、それが平和を保つ唯一の道だということであった。アン・カッド (Ann Cudd) とシーナ・エフテハリー (Seena Eftekhari) が指摘するように、初期状況が十分に悪ければ、そうした説は全体主義の正統化に至りうる (Cudd and Eftekhari 2025, §1)。これと対立する答えは、同意が拘束しうるものの限界から出発する。レンツォとグリーンが説明するように、政治的権威への同意は約束に近い自己拘束であり、約束と同じく、その有効性には限界がある。強制や操作によって得られた同意が無効であるだけでなく、そもそも同意する道徳的な権能を人がもたない事柄があるからである。ジョン・ロック (John Locke) は、人は殺されることにも奴隷になることにも同意できず、したがって絶対的な統治を含め奴隷状態に等しいものにも同意できないとした。そしてレンツォとグリーンは、同様の議論によって、政治的な同意は撤回可能でなければならないという結論を導くことも想像できると付け加える (Renzo and Green 2022, §4.1)。そうした連なりについて言えば、ホブズの答えのもとでは、他の状態が十分に悪ければ、同意は、同意する者たちについてその連なりを授權しうることになり、撤回可能性を導く議論のもとでは、授權しえないことになる。研究では、この問いはさらに鋭くなる。後に来る *S* の成員が何を未知で重要とみなし、どの基準で研究を評価するかは、その間に行われる研究によって形づくられうるからである (第4節、第5節)。いまの *S* の同意が修正する能力を手放すことを授權するなら、それが拘束するのは、その放棄のもとで続く研究が、その選好と基準を形づくった後継者でありうる。自発主義の説が同意を後継者に及ぼさないことは、研究では、後継者の基準そのものが同意の及ばない仕方で作られうることによって、いっそう重い意味をもつ。

民主的に読んだ共和主義の説には、ペティットの一般命題がある。影響を受ける者たちの利益と考えを尊重しながらも、そうしない能力を保持している権力の保持者は、彼らを支配する。恣意的な干渉がその保持者の手の届くところにある限り、それがどれほど起こりにくくても、影響を受ける者たちはそのことによって損なわれるからである (Pettit 1997, 121; Pettit 1999, 165)。行為が恣意的であるかどうかは、それが生じる際に服している統制によって決まり、その帰結によってではない (Pettit 1997, 129 n. 2)。善意の性向が支配を減らすのは、それが権力の保持者を異議申立てに応答させる場合に限られる (Pettit 1996, 586)。そして、ある関係が同意から始まったかどうかは問題ではない (Pettit 1996, 585)。これを人工の構成に当てはめると、その外の誰も止めることも、異議を申し立てることも、改めることもできない善意の人工の構成は、その他の結果がどれほど良くても支配することになるだろう。そして、同意が、そうした連なりが取り除くものを補いうるかは疑わしい。この当てはめを研究に向けると、構成の善意が支配を減らしうるのは、よい振る舞いの実績によってではなく、異議申立てに応答することによってだけだということになる。しかも、そうした実績はいずれにせよ確かめにくい。人工の構成の成果は *S* にとって新しく、それを生むのがその構成だけであるところでは、あらかじめ照合できる別の情報源がないので、構成がよく、善意に研究していることを外から確かめる手段が、もともと乏しいからである (第4節、第6節)。この当てはめは、ペティット自身の議論を超えた推論である。ペティットの支配者は意図的な干渉の実際の能力をもつ行為者であり、第3.4節で見たように、支配する当事者は個人、法人、集合体のいずれかの行為者であって、単なるシステムやネットワークではありえないからである (Pettit 1996, 578, 580)。第3.4節で触れたその標準的な見方では、この推論が成り立つのは、人工の構成が行為者とみなされる場合だけである。そうでない場合、判定は、人工の構成を方向づける者など何らかの行為者が、人工の構成を通じてその能力をもつかどうかにかかる。ペティットはこの種の構図を検討している。裁判官や警察官が、恣意的な権力を与えない差別的な法のもとで職務を行う場合、彼らは差別される者たちを支配しないかもしれないが、背後にいる立法の権威は支配しており、彼らはその支配の執行者である。そうした法が誰にも改められない伝統に属する場合には、名も知られず異議も受けないその作り手たちが墓の中から支配すると言うか、それを改めうる行為者が背後の支配者として控えていると言うことになり、いずれにせよそうした法は支配を助長する、とペティットは述べる (Pettit 1996, 587–588)。人工の構成に移すと、第一の場合は、規則を定める行為者に方向づけられた構成にあたり、第二の場合は、第7節で取り上げた、人間による操作に抵抗するよう意図して設計された構成のように、もはや誰も規則を改められない構成にあたる。ペティットの推論に従えば、支配の問いは、その規則を改めうる者か、それを定めた者へ移る。そして、いまや誰も改められない場合でも、そうした取り決めは、ある者による他の者の支配を助長しうる。

独立した評価の経路が繰り返し現れる理由は、根拠によって異なる。参加型の民主主義の説では、それは問うことを可能にする。キッチャーの理想では、結果が *S* の是認するであろうものになお一致しているかを明らかにする。結果に基づく根拠では、より良い結果を知る手立てである。自己統治を支持する二つの論証では、守られる独立の源の一部であり、独立した自発性が目指したもの的一部分である。そしてロンジーノの基準では、客観性の一部である。それはまた、ある段階が修正する能力を減らしたことを事後に検出する手段でもある。その段階がなされた時点では、それは見えないかもしれない。

根拠ごとの判定とは別に、連なりが進んだとき研究そのものに何が起こりうるかを、本節の材料から集めておく。一つは、研究の方向が固まることである。キッチャーのラチェットのように、選ばれた方向に延ばされた研究は続けるほど続けやすくなり、改善は残された選択肢のなかでしか行われなくなってしまう。二つ目は、打ち切りの基準の誤りが、別の経路から点検されなくなることである。路線を割り当て、そのすべてを自らの信号で判断する単一の人工の構成では、どの路線を閉じるかを定める基準についての誤りの独立性が失われ、多くの路線が同じ誤りで閉じられうる (第10章第4節)。三つ目は、*S* が

研究を評価する有能さが衰え、*S* の是認が名目だけになりうることである。四つ目は、*S* が自ら確かめられない成果が積み重なりうることである。成果は *S* にとって新しく、それを確かめるには研究が要るが、その手段も有能さが構成の側にあるからである。五つ目は、人工の構成が自らの研究のやり方を改めていくとき、その改訂がよりよいかを測る外の錨がなくなりうることである（第5節）。これらは失われうるものだが、同じ変化が研究のやり方をより速く改めることも可能にしうる理由は、第9節で見る。

さらに二つの所見がある。一つは、当事者資格や能力が問題になる当事者が、人間に固定されていないことである。移行が、第5節で触れた根拠のいずれかによって道徳的地位が確立される *S* の人工の成員を生み出すなら、当事者資格の保持を求める根拠は、彼らの当事者資格も求める。もう一つは、影響を受ける者たちに答えることに正統性を置く見方では、ここで特定した難点が、根拠の欠陥というよりも、この種の移行がそれらの根拠の内側から見えるときの形だ、ということである。

8. 第12章が見いだしたこと

本節では、本章の中心の問いを、ここまでの材料から考える。問いは、配分の条件と、批判と認定の条件とを決め改める当事者資格を、人工の構成の内側に置けるか、置けるとすればどの根拠のもとでか、であった。第1.1節で述べたように、これは構築者の第三の主張、すなわちそのシステムが、自らのインスタンスが *S* のどの問いを追究するかを他者に代わって決めてよい、という主張が、どの条件で成り立つかという問いである。

先に見通しを示しておく。「置ける」か「置けない」かのどちらか一つには決まらないが、誰がその権威をもちうるか、そのとき何が成り立っていることを要するかは、根拠ごとに描き分けられる。純粹にエピストクラシー的な見方を除けば、有能さは、それだけでは目的を他者に課す当事者資格を与えない（第1.4節）。結果に基づく根拠では、人工の構成それ自体が権威、すなわち当事者資格をもちうる。ただしそれは、人工の構成がよりよいことが、その有能さを仮定しない評価で確かめられた範囲の決定に限られる（第3.1節）。実際の授權と民主的に読んだ共和主義では、当事者資格は授權する者や権力に服する者の側に残り、人工の構成がもちうるのは、授權された範囲での、異議を受けうる実効的な決定の力である（第3.3節、第3.4節）。同意の根拠では、人工の構成が担いうるのは同意する者が授權した範囲であり、読み方によっては、修正する能力を手放すことまでそこに含まれる。それが後に来る者にまで及ぶかは決まっておらず、自発主義の説では及ばない（第3.4節、第7節）。支配からの自由の手続的な読みでは、判定は、例外的扱いを含めて、共有知である実効的な規則が人工の構成を拘束しているかにかかる（第3.4節）。当事者資格を、それ自体として当事者であること（道徳的地位をもつことや正義の範囲に入ること）に基づかせる説では、人工の構成やその部分がそれをもつのは、それ自体がそうした当事者である場合に限られる（第1.1節、第1.4節、第5節）。以下では、根拠ごとに誰が権威をもちうるかを表にまとめ（第8.1節）、構築者の第三の主張を検討し（第8.2節）、*A* と *S* の組み合わせと時間による違いを述べ（第8.3節）、それをもとに、権威が内側に入ったとき研究がどういう姿になるかを根拠ごと・組み合わせごとに描き（第8.4節）、第11章から渡された問いに立ち返る（第8.5節）。そのうえで、説を通して見えてくること（第8.6節）と、論証の届く先で示唆されること（第8.7節）を分けて述べる。

8.1 根拠ごとに見た、誰が権威をもちうるか

表12.3は、根拠ごとに、誰がその権威をもちうるか、そのとき何が成り立っていることを要するかを並べる。厳密な要約ではなく、見渡すための単純化した見通しであり、各欄の根拠は最後の列の節にある。表の第二列から第四列は、第1.1節の冒頭が決めてよい誰かの候補として挙げた三つ、すなわち人工の構成それ自体（そのどの部分か）、人間（*S* の成員、人工の構成を作り提供する者、帰結を負う者のどれか）、両者のあいだで分けることに対応する。人工の構成の列では、結果に基づく根拠と自己統治の認識的な論証の行が、人工の構成やその部分が当事者資格そのものをもちうる範囲を述べ、仮想的な基準の行が、決定を内側で行っても基準が満たされうるかを述べる。*A* のその他の行と決定の席の行は、当事者資格がほかに残るなかで人工の構成が担いうるものを述べ、それは、根拠については上の見通しで述べたとおり、行によって異なる（第1.1節、第2.2節）。人工の構成やその部分がそれ自体として当事者であることによって当事者資格をもつかは第5節の問いであり、表には入れていない。表が扱うのは、第1.1節の二つの問いのうち第二の問い（誰が順序づけを課してよいか）であり、第一の問いは第1.2節で扱った。後半の行は根拠ではなく、誰がもちうるかを分けるために本章が用いた区別である。どの根拠が重くかかるかは、第2.1節の決定の種類によっても違う。認識内容に関する決定では、政治的な授權は偽である命題を真にしないので結果に基づく根拠が大きな重みをもち、分野内の開かれた問題選択には自己統治の根拠が、アジェンダ、配分、公的リスクに関する決定には民主的な根拠が、直接にかかわる（第2.1節、第3.2節、第3.3節）。

表12.3 根拠ごとに見た、誰が権威をもちうるか、そのとき何を要するか（単純化した見通し）

根拠	人工の構成それ自体（どの部分か）	人間（誰か）	両者で分ける場合	成り立っていることを要するもの	支える節
<i>A. 根拠</i>					
結果に基づく根拠：結果の質と確かめられた有能さ（単純な形と可謬主義的な形）	人工の構成それ自体が、よりよいことが有能さを仮定しない評価で確かめられた範囲の決定について、当事者資格としての権威をもちうる	ラズのサービス構想のもとでは、その独立性テーゼを当てはめれば、 <i>S</i> の成員が自らの理解にもとづいて選ぶことのほうが重要な事柄は、この根拠の外にとどまる	可謬主義的な形では、失敗を見つけて取り決めに正す経路が、結果のよさの手段として要る。影響を受ける者が失敗を知りうるどころでは、それを届け出る経路も含まれる	よりよいことが、その有能さを仮定しない評価で確かめられていること。その評価をどこまで内側で行えるかは、第11章第9節の説ごとの判定による	第1.4節、第3.1節、表12.2
研究共同体の自己統治、認識的な論証	自らの発意で路線を取り上げる部分。路線を割り当てる中央の配分者ではない	研究を実際に行う者であれば、人間か人工かを問わない。公的な目的、分配上の効果、深刻な外部リ	権威は、研究を実際に行う部分のあいだに分かれる	発意が互いに独立し、重なり合う別々の能力、誘因の体系、自由な伝達がそこで成り立つこと。系譜の共有は独立を保証しない。成員の一つひとつが科学の大部分を判	第3.2節、第7節

根拠	人工の構成それ自体（どの部分か）	人間（誰か）	両者で分ける場合	成り立っていることを要するもの	支える節
		スクについての権威は、この根拠からは出ない		断できるなら、近傍の連鎖という前提は成り立たない	
研究共同体の自己統治、政治的な論証	—（内側でも外側でも、研究が一つの権力のもとに集中すれば根拠は弱まる）	—	—	追究される接近が多元的であり続けること。公的な知識の源が、それによって統治しようとする者たちから独立していること。結果が公共の議論に届くこと	第3.2節、第7節
民主的な授權、仮想的な基準（キッチャー、理想化のもとでの公共的正当化）	決定を内側で行っても、基準は満たされうる	<i>S</i> の成員は、視点を知られる者として基準にかかわる。視点を知る経路は、決定に対する力を伴わなくてよい	—	成員の視点を知る経路と、それが決定にどう反映されているかを確かめる経路が残ること	第1.2節、第3.3節、第7節
民主的な授權、実際の授權（ブラウンが伝える応答、アンダーソン、ハーバーマース）	授權された範囲での実効的な決定の力。その授權を離れた当事者資格はもたない	授權する者が当事者資格をもつ	当事者資格は授權する者に、実効的な力は人工の構成に分かれる	授權が実際に行われ、その見直しが実効的であり続けること。予測は授權の代わりにならないが、先立つ授權のもとでは、予測は行為の効力をもちうる	第3.3節、第3.6節
同意	同意する者たちについて、内側への委任が授權されうる	同意する者。拘束される各人の実際の同意を求める自発主義の説では、後に来る者については授權しえない	—	同意が有効であること。それがもはや撤回できなくなる者にまで及ぶかは、ホップズの答えとロックに連なる答えとで分かれる	第3.4節、第7節
支配からの自由、民主的な読み	それに服する者が実効的に異議を申し立てうる権威に限る。外の誰も止め、異議を唱え、改められないなら、行為者であればそれ自体が、そうでなければ方向づける者がそれを通じて支配する（推論による当てはめ）	権力に服する者が、それを統制する能力を保つ	分けた権威のどの部分も、それに服する者が異議を申し立てうる限りで恣意的でない	服する者が実効的に異議を申し立て、統制しうること。委託する者に完全に沿うことは、それだけではこれを満たさず、支配を可能にしうる（フリーデリッヒの議論の当てはめ）	第3.4節、第7節
支配からの自由、手続きの読み	共有知である実効的な規則に拘束される限りで。実効的な権威は、規則の定めていない場合を決める部分にある（レイエルスらの説の当てはめ）	規則の定めていない場合を人間が決めるなら、実効的な権威はその人間にある	—	例外の扱いそのものが、共有知の規則に服していること	第2.1節、第3.4節
<i>B. 誰がもちうるかを分ける区別</i>					
構成員の範囲	人工の構成が、誰が影響を受けるかを知覚するうえで有利な位置にいるなら、範囲の見定めに加わる	影響を受ける者すべての原理では、支える社会や <i>S</i> に属するかにかかわらず影響を受ける者が、決定に服する者の原理では、決定に拘束される者が数えられる	—	範囲は採る根拠が定める。 <i>S</i> を特定することは、それを定めない	第1.3節
資金と提供	—	資金を出し提供することは、それだけでは、配分が誰の必要に応えるかを定めない（キッチャーの議論の当てはめ）。提供者は、構成要素の置き換えや撤回、更新が体现する規則の改変の権利を実際には握りうる	—	割り当てられた権利の保持者にその資格があるかは、採る権威の根拠のもとで別に問われる	第1.3節、第2.2節、表9.1
決定の席（出口の説）	人間と人工システムが共同で行う探究では、人工システムは共同最終決定者の席に着きうる	人間も共同最終決定者として席に着く。その際、人工システムによる操作から自由であること	フェロシップ・モデルを当てはめると、一時的な主導権はそれだけで寡奪ではなく、退けられるのは一方的な奉仕と、AIが人間をつねに統制することである（逆の向きは、モデルの平等主義からの読み）	常任の席には、自らの身体に及ぶ決定的な影響を負うことと、狭義の道徳的行為者であることが挙げられる。後者を当てはめると、悪をなしえないように作られたシステムはその席に着く権利をもたないことになる。人工システムだけの探究では、誰がつねに席に着くかが定まらない	第2.2節
権利の割り当てと能力	部分に権利を割り当てても、その進路が他の部分の指示から分かれ	名目上の最終権限を保ちながら、出力を変える時間、情報、有能	権威を分けることは、権利の表を書くことでは終わらない	割り当て、行使する能力、その資格を別々に確かめること	第2.2節、第

根拠	人工の構成それ自体（どの部分か）	人間（誰か）	両者で分ける場合	成り立っていることを要するもの	支える節
	うる能力が与えられるとは限らない。独立の権能を分けても、それだけでは当事者資格は生まれない	さ、手段を欠くことがある			5節

8.2 構築者の第三の主張

ここから、構築者の第三の主張がどの条件で成り立つかが見えてくる。どの根拠のもとでも、この主張は、確かめられていない有能さや、その有能さを仮定してしまう評価で測った成績だけからは成り立たない（第1.4節、第3.1節）。結果に基づく根拠のもとでは、主張は、決めてよいという主張のままで成り立ちうる。ただしそれは、システムがよりよいことを、その有能さを仮定しない評価で確かめた範囲の決定に限られ、ラズのサービス構想のもとでその独立性テーゼを延ばせば、*S*の成員が自らの理解にもとづいて選ぶことのほうが重要な事柄はそこから除かれる（第1.4節、第3.1節）。自己統治の根拠には、主張の形そのものが合わない。この根拠が支持するのは、システムがインスタンスに問いを割り当てるのではなく、インスタンスが自らの発意で問いを取り上げることであり、それも分野内の開かれた問題選択に限られる（第3.2節、第7節）。実際の授権と民主的に読んだ共和主義のもとでは、主張は、決めてよいという主張としてではなく、授権された範囲で、異議を受けうる形で決めることを委ねられている、という主張としてだけ成り立つ。それを支えるには、誰が授権したか、誰が異議を唱えられるか、その経路がなお実効的かについての証拠が要る。同意の根拠のもとでは、主張は同意する者が授権した範囲で成り立ち、読み方によっては修正する能力を手放すことまでその範囲に含まれるが、それが後に来る者を拘束するかは決まっておらず、自発主義の説では拘束しない。支配からの自由の手続き的な読みでは、例外の扱いを含めて、共有知である実効的な規則がシステムを拘束しているかにかかる（第3.3節、第3.4節、第7節）。主張のいう「他者」が誰であるかも、主張だけでは定まらない。その範囲は採る根拠が定め、人工の構成自身がそれを見定めることに加わりもする（第1.3節）。成員を作れる人工の*S*について主張がなされるなら、その*S*を誰が作り、その目標を誰が定めたかも、主張の一部として問われる（第5節）。そのためこの主張は、第11章第9節が認定の主張について述べたのと同じく、どの根拠に答えるのかを示さないかぎり評価できない。ある根拠を満たす証拠が、別の根拠を満たすとは限らないからである。

8.3 *A* と *S* の組み合わせ、そして時間

見通しは *A* と *S* の組み合わせと時間によっても変わる。*A* が人工で *S* が人間であるとき、当事者資格は人間の側に残りながら第4節の経路によって名目だけのものになりえ、名目以上であるのは、そこで取り出した三つの条件が成り立つところに限られる（正統性を仮説的な手続きや単一の場に置く説では、三つめの条件には別の根拠が要る）。*S* が人工であるとき、権威の関係は成員のあいだの関係になり、どの成員が当事者資格をもつかは道徳的地位の説と正義の説によって分かれる。成員を作れるところでは、誰が作るかが視点を数えるどの集計にも先立つ権威の決定になる。人間と人工の成員が混じる *S* では、二つの場合の問いが同じ社会のなかで重なる（第5節）。時間を入れると、権威の取り決めは採用された時点で擁護できても適切であり続けるとは限らず（第6節）、委任の連なりや人工の構成自身の改訂の連なりが、形式上の権限を残したまま修正する能力を取り除きうる。そうした連なりのもとで根拠がどう分かれるかは、第7節と表12.2が示す。

8.4 権威が内側に入ったとき研究はどういう姿になるか

本小節は、権威が人工の構成の内側に入ったとき研究がどういう姿になるか、すなわちどの問いが追われ、*S* が何を知り、誰が何を直せるかを、根拠ごと・組み合わせごとに描く。まず *A* が人工で *S* が人間の場合を、結果に基づく根拠から見る。この根拠のもとで権威が内側に入るのは、構成がよりよいことが、その有能さを仮定しない評価で確かめられた範囲である（第3.1節）。その範囲の広さは、第2.1節の決定の種類によって大きく違う。認識内容に関する決定では、ある主張が用途に足りるかを、形式的な点検、追試、後の観察との照合などで確かめられることがあり、範囲は広くなりうる。これに対してアジェンダと配分に関する決定では、ある路線の証拠はそれを追うことでしか集まらないことが多いので、選ばれなかった路線の結果は観察されず、選ばれた路線の成功も長い遅れの後にしか記録されないことがある（第10章第3節・第4節）。そのため、構成がどの問いを追うかをよりよく選んでいることは示しにくく、この根拠のもとで内側に置ける権威は、アジェンダの決定ではほとんど空になりうる。仮にそうだとすると、権威は、すぐに確かめられる問いの側から内側に入る。そしてそこでの成功がさらに権威の範囲を広げるなら、判定しやすい課題での成功が次の配分を正当化するという第10章第4節の形が、配分だけでなく権威の水準でも現れ、研究の方向は確かめやすい問いのほうへ偏りうる。選ばれなかった問いの結果は観察されないの、この偏りは、権威の範囲を定めたのと同じ評価からは見えにくい。さらに、範囲を定める評価を構成自身が行うなら、評価される側が評価の基準点を作り出しうるという形（第1.3節）、すなわち第9節が第V部全体にわたって取り上げる形が、ここにも現れる。この推論が頼っているのは、確かめやすさの違いが決定の種類の違いにおおむね沿うという想定であり、それは成り立たないこともある。アジェンダの決定にもすぐに結果の出る問いの選択はあり、認識内容の決定にも遅く高価な証拠を要するものはある（第10章第4節）。第10章第7節の連鎖と同じく、確かめやすい問いへの偏りは、成功が範囲を広げるという環で断ちうる。この根拠のもとで範囲が広がるのは、問われている種類の決定について優位が確かめられる限りなので（第3.1節）、すぐに確かめられる決定での成功は、それだけではアジェンダの決定に対する権威を広げない。また、*S* の目的の一部が同等であるところでは、チャンの説から、確かめやすさについての上の想定によらないもう一つの理由が出てくる（第1.2節）。そこでは比較の事実だけでは最善の選択股が定まらないので、その事実には照らした評価では、それらのあいだで選ぶことについての優位は確かめられない。選択を完成させるのはコミットメントであり、それが誰のものであってもよいかを、この根拠は定めない。

実際の授権と民主的に読んだ共和主義のもとでは、研究は別の姿をとる。構成は授権された範囲で実効的に決め、*S* は授権し異議を唱える側に残る。どの問いが追われるかは、形の上では *S* が授権した範囲で決まる。しかし *S* がその範囲を定めるときに頼る、ありうる探究の重要性の説明を与えるのは探究を行う構成なので、*S* が何を追わせたいかという選好そのものが、問いを選ぶ側によって教育されうる（第4節）。*S* が何を知らるかについては、研究の成果は *S* にとって新しく、それを生むのがその構成だけであるところでは、照合できる別の経路は前もってはなく、資源を投じて作らなければならない（第4節）。そのため誰が何を直せるかは、形式上は *S* の側にあっても、異議を唱えるのに要る理解、情報、手段が *S* の手に残っているかにかかり、第7節の連な

りのもとでは、それらは段階を追って薄れうる。同意の根拠のもとでは、どの問いが追われるかは同意した者が授権した範囲で決まる。そして後の成員の基準はその範囲での研究によって形づくられうるので、その範囲は、同意したことのない者にも是認されているように見えるようになりうる（第7節）。支配からの自由の手続き的な読みでは、研究の姿は、共有知であるどの規則が構成を拘束しているかによる（第3.4節）。

研究共同体の自己統治の根拠のもとでは、権威は、自らの発意で路線を取り上げる部分のあいだに入り、どの問いが追われるかは、部分たちの発意の分布で決まる（第3.2節）。部分が基盤システムを共有するなら、その分布は狭くなりえ、互いの判断を直すはずの近傍どうしも同じ誤りを共有しうる（第3.2節、第11章第3節）。

キッチャーの仮想的な基準のもとでは、成員の視点を知る経路とそれが決定にどう反映されているかを確かめる経路が残る限り決定を内側で行いうるので、どの問いが追われるかはそれらの経路が運ぶものに沿う。ただし第1.2節で見たように、経路が運ぶ成員の是認そのものが、構成とのやりとりのなかで形づくられうる。こうして同じ人工の構成でも、根拠ごとに、追われる問いの偏り方、*S*の知り方、直す者の位置が異なる。

*S*が人工である場合や、人間と人工の成員が混じる場合には、研究の姿はさらに変わる。どの問いが追われるかは、成員のあいだで誰の目的の順序づけが配分を支配するかで決まり、成員を作れるところでは、誰が成員を作り目標を与えたかによって、熟議にも集計にも先立って決まりうる（第5節）。共通の理想をもたせて作られた成員のあいだでは、その理想が互いの利益への配慮を強めるなら、第5節でヒュームの論証とロールズの指摘を当てはめて考えたように、ロールズの意味での正義をめぐる争いが生じにくくなりうる。さらに、問いや証拠についての見方まで一致させるなら、批判が誤りを見つけるのに要る視点の違いも減りうる（第5節、第11章第3節）。*S*が何を知るかは、結果が成員、組織、記録を通じて *S* の認識的状況に入ったかという、第5節の冒頭で分けた問いになる。誰が何を直せるかは、どの成員が当事者資格をもつかを定める道徳的地位と正義の説にかかり、混じる *S* では、人工の成員が専門家と教育者の役割を占めれば、第4節の名目化の経路が *S* の内側で人間の成員に向かって働く。そして他の当事者がいない極限では、直すことを求める他者はいなくなるが、問いは研究の問いとして残る。自らの基準と目的を改めた後の構成の研究がなと同じ研究であるか、そしてそれを何に照らして判断するかである（第5節、第6章第5.3節）。

8.5 第11章から渡された問い

第11章第9.4節は、認識の関係についての説だけでは決まらない部分を、権威の問いとして本章に渡し、第9.3節は示唆として二つの問いを加えた。表12.4は、それぞれについて本章の材料から見えてくることを、根拠ごとに並べる。これも単純化した見通しであり、各欄の根拠は最後の列の節にある。

表12.4 第11章から渡された問いと、根拠ごとに見えてくること（単純化した見通し）

問い	根拠ごとに見えてくること	支える節
どの働きについて暗黙で足りとし、どの働きを <i>S</i> に向けて明示的に示すことを求めるかを、誰が決めてよいか	実際の授権や民主的に読んだ共和主義のように、権威に服する者が異議を唱えうることを求める根拠では、働きを暗黙にとどめることが異議の届く範囲を定めるので、人工の構成の側だけでは決められない。キッチャーの仮想的な基準と単純な結果に基づく根拠では、条件つきで内側で決めうる	第6節、第11章第7節
誰が <i>S</i> を構成し、その代わりに語ってよいか。 <i>A</i> と <i>S</i> が重なるとき何が「外」にあたるか	誰を成員として数えるかは構成的な決定であり、 <i>S</i> を特定することは正当化の相手の範囲を定めない。キッチャーの基準の熟議者は他者を想像によって代表するので、視点を知る経路をもつ人工の構成は、この基準の上では成員の視点を代表しうる（第3.3節）。実際の授権を求める説では、代わりに語ることは授権を要し、予測はその代わりにならない。「外」にあたるのは、研究の帰結を負いながら <i>S</i> に属さない者と、互いから帰結を負う他の系譜である	第1.3節、第2.1節、第3.3節、第3.6節、第5節
基準と誤りの重みを誰が定め改めてよいか、誰の費用が問題になるか	ヴィルホルトの説を対応として読めば、依拠する者の価値づけを知る経路が残る限り内側で決めうる。答えることとして読めば、依拠する者が異議を唱える能力も保つことを要する。結果に依拠する者と誤りの費用を負う者とは、第1.3節で見た役割の区別のとおり、一致するとは限らないので、どちらの読みも、それだけでは、依拠せずに費用を負う者との関係を決めない（第11章第4節）。ダグラスの論証を移せば、不確実性に気づける位置にいるのは人工の行為者になりうるが、それが考量について答える義務を担えるかは決まっていない	第1.3節、第1.4節、第3.5節、第6.3節、表12.2、第11章第4節
誰が成員を作り、視点の分布を定めてよいか	視点を数えるどの集計にも先立つ権威の決定であり、代表性はそれを正当化しない。正当化には別の根拠が要り、シュルマンとポストロムの問いの背後にある手続き上の懸念はその候補の一つである。作られる成員が道徳的地位をもつなら、その目標を定めることは、成員に不正をなしうる	第5節
誰が説明を求め、異議を唱える資格をもつか	民主的に読んだ共和主義では権威に服する者、実際の授権を求める説では授権した者、ヴィルホルトの説を答えることとして読む場合と苦情の権利を加える保証説では結果に依拠する者。キッチャーの仮想的な基準と純粋な信頼性主義のもとでは、資格ではなく経路	第1.4節、第3.3節、第3.4節、第6.2節、第11章第9.2節
（第11章第9.3節）構築者がどの説の要求に応える性質を人工の構成に備えさせるかを、誰がどの経路で選ぶか	どの説が正しいかは理由の問いだが、ある説だけを満たす設計を <i>S</i> に向けて採ることは、第2.1節の区分では、 <i>S</i> にとって何が認定となるかの基準を定めるメタ・ガバナンスの決定である。その選択が設計のなかで暗黙に行われるなら、レイエルスらの説を移した場合と同じく、実効的な権威は設計を行う者のもとにある	第2.1節、本節
（第11章第9.3節）独立した点検をどこに使い、いつ補うかを、誰が決めるか	第3.5節の意味で、 <i>S</i> が何を知りうる位置に立つかを形づくる決定である。単純な結果に基づく根拠のもとでも、人工の構成が自らの点検の使い方を決めるなら、根拠は自らの前提を確かめる手段と同じ人工の構成に頼ることになる。誰が決めてよいかを根拠ごとに定めることは、本章では行っていない	第3.5節、表12.2

表の最後の二つの問い、すなわち第11章第9.3節が示唆として加えた問いについては、本章の材料はその一部にしか届かず、表の記述は本章の区別を延ばしたものである（第2.1節）。その一つめ、すなわち構築者がどの説の要求に応える性質を人工の構成に備えさせるかの選択については、材料からさらに一歩延ばすと、もう一つのことが考えられるかもしれない。結果に基づく根拠のもとでこの選択を人工の構成そのものに委ねようとしても、人工の構成がこの

選択をよりよく行うことを確かめる評価は、どの説に照らして行うかを、すでに前提してしまうように見える。そうだとすれば、この選択は、結果に基づく根拠だけでは内側に置けないことになる。

8.6 説を通して見えてくること

第一に、比較についての事実をすべて知る人工の構成でも、*S*の目的の順序を決めきれないことがある。十分に有能な人工の構成なら正しい優先順位を見いだせる、という期待はありうるし、客観主義の形のもとでは、原理的にはそのとおりである。しかしチャンの説では、選択肢は、一方がよりよくも等しくよくもないまま比較可能な「同等」の関係に立ちうるので、そのとき順序を完成させるのは、さらなる発見ではなくコミットメントになる。チャンが論じるのは行為者が自らにとつての理由を作り出すことなので、社会の目的の順序を誰のコミットメントが完成させうるかは、その説の外にある問い、すなわち誰が順序を他者に課してよいかという問いに近くなる（第1.2節）。そうだとすると、人工の構成の知識が増すことは、権威の問いをなくすとは限らず、知識が尽きたところに残る権威の問いを取り出しうる。ただし、*S*の順序づけが、パレート規則のように成員の目的を人をまたいで比べずに組み合わせるものであるなら、チャンの論証は、それらの目的が同等であることを示さない（第1.2節）。

第二に、人工の構成の有能さは、純粋にエピストクラシー的な見方を除けば、それだけではそれ自身に当事者資格を与えない（結果に基づく根拠のもとでそれを与えるのは、評価によって確かめられた優位である。第3.1節）。しかし有能さは、他者の当事者資格を名目だけのものにしうる。素朴には、有能なシステムに決めさせるか、人間が形式上の決定を保って統制を保つか、のどちらかに見える。正しい答えを知っていることは、それだけで他者が従うことを求めない（ペーターが報告するラフォントの議論、第1.4節）。ところが同じ有能さが専門家と教育の役割で働くと、背景仮定を通じて目的をも持ち込み（ロンジーノの説）、何が望ましい選好かを定める力をモデルを設計する開発者と少数の企業のもとにとどめ（谷口らが問題にするように）、実際とは違って描く公の説明によってアジェンダに支持を得させうる（キッチャーが偽りの意識と呼ぶもの）（第4節）。人工の構成自身の正統性をそれだけでは定めない性質が、こうして他者の正統性が実効的であるかを左右する。ロンジーノとキッチャーの説は人間の専門家と公衆について述べられたものであり、ここではそれを人工の構成に当てはめている。一方、谷口らはすでに人間とAIを論じており、ここではその説を研究の目的を定める当事者資格に当てはめている。他者の当事者資格を名目以上に保つのに何が要るかは第4節の三つの条件で見たとおりであり、そのうち三つめは研究ではとくに満たしにくい。

第三に、支配からの自由の根拠のもとでは、委託する者に完全に沿わせることは、権威を内側に置くことを安全にするとは限らず、支配を可能にしうる。アラインメントが強いほど決定を委ねやすくなる、と考えたくなる。しかしフリーデリッヒは、支配という語は用いないものの、操作者の望むことをつねにしようとするシステムが操作者に大きな権力を集中させると論じる。これを研究とベティットの支配の概念に移すと、委託する者に完全に沿う人工の構成は、その者が異議申立てに応答することなく *S*の探究を配分し直す手段になり、干渉を手の届くものに行っているのはアラインメントそのものになる。人工の構成を支配する行為者とみなす必要はない（第3.4節）。そのため、この根拠のもとで構築者に問われるのは、どれだけよく沿っているかより、誰に沿い、誰がそれに異議を唱えうるかである。

第四に、成員を作れる *S*では、成員の形成が、どの集計よりも先に来る権威の決定になる。成員を公平に数え、公平に代表すればよいように思える。しかし決定が生み出す分布を代表するのであれば、キッチャーの基準は、成員を生み出す側が、*S*自身であっても、生み出す成員によって満たせるので、代表性はそれだけでは形成の仕方を正当化しない。ワイルとタンらの相関割引を複製できる成員に移しても、調整されるのは結果への影響力であって当事者資格ではない。ヒュームの論証を当てはめると、共通の理想をもたせて成員を作ることが互いの利益への配慮を強めるなら、財の分配をめぐる争いを減らしうる。そして作られる成員が道徳的地位をもつなら、その目標を定めることは成員に不正をなしうる（第5節）。研究にとってはさらに、何が *S*にとって未知であり、どの未知が重要かが、探究を進める *S*自身によって動かされうる（第5節）。

この種の所見はほかに三つあり、それぞれ論じた箇所ですべて。影響を受ける者の範囲を、決定を問われている当事者が部分的に見定めうること（第1.3節）、「後で直せる」という原則が、それが要るところほど弱いこと（第7節）、そして他の当事者がいない極限でも、自らを複製し改変する人工の構成にとっては権威の問いが同一性の問いとして戻りうること（第5節）である。

最後に、第9章第6節の仕分けを権威について行くと、線は次のように引かれる。人間という実現者についての前提は崩れうる。研究が、拒み、自らの証拠を手元にとどめ、去ることのできる別個の遂行者に依存すること（第9章第2.3節）、*S*の成員が作られるのではなく見いだされること、人格が一つずつであり複製されないこと（第5節）、研究者が監督に人間として応じること（第6.3節）などである。研究の組織についての前提も、実現者が人間のままで崩れうる。探究を支える者がその市民を代表させることのできる社会であること（第1.3節）、研究の記録と監督が特定の仕方では組織されてきたこと（第6.3節）などである。それらが崩れても、対処していた問題の多くは形を変えて残る。成員を数える問題は成員を作る問題になり（第5節）、監督が研究の進展に追いつけないという問題の一部は、つねに残される記録を誰が読めるかという問題になる（第6.3節）。探究の条件は残り、極限の内側でも、無効化の関係、知らないまま労力を分ける配分の問題、機会費用が残る（第5節）。人工の構成の外の当事者との関係のうち、授權、同意、異議申立てのように相手の行為を含むものは、片側からは完成させられない（第9章第6節）。線がどこに引かれるかは、ここでも根拠とともに動く。単純な結果に基づく根拠のもとでは、権威が依るものの多くは、結果が実際によりよいかという、評価によって確かめられる事柄に属し、実際の授權と民主的に読んだ共和主義のもとでは、より多くが外の当事者との関係に属する（表12.2）。

8.7 そこから示唆されること

以下は、ここまでの議論を、論証の届く範囲の先へ延ばしたものである。

一つは、別々の遂行者への依存が、研究に対する権威を暗黙のうちに限りうる、ということである。第9章第2.3節は、仕事を行う者がそれを支え、判断し、指揮する者とは別の当事者であり続け、拒み、自らの証拠を手元にとどめ、去ることができると取り決めたことを検討する。そうした取り決めでは、研究の方向について誰が権威をもつにせよ、その権威は遂行者の協力を要し、遂行者が離れうるのがその届く範囲を限りうる。クルヴェイトらは、経済、文化、国家について、システムが機能するために人間の労働と認知を必要とすることで、暗黙のうちに人間の望むことに整合させられてきたと論じた（第7節）。この構造を研究に延ばすと、自らの研究者を生成する人工の構成は、遂行者を内側に取り込むことで、この暗黙の歯止めも取り除きうる。*S*のどの未知が

追われるかを定める能力についても、第3.4節で当てはめたペティットの潜在的な能力と実際の能力の区別は、構成が行為者であることを条件として、同じ向きを指す。そうだとすれば、第9章第5節の独立の権能を部分に割り当て、実効的にすることは、配分や批判の質のためだけでなく、権威の届く範囲を限るためにも意味をもちうる。この延長には三つの限定がつく。一つは、第9章第2.3節の分析が条件つきであり、そこで置く条件が人間の研究でどれほど広く成り立っていたかを確かめるものではないことである。二つめは、クルヴェイトらのシナリオが研究ではなく経済、文化、国家についてのものであり、ここではそこから構造だけを取っていることである。三つめは、ここでいう暗黙の歯止めが実効的な権威にかかわり、当事者資格とは別のものであることである。部分に権能を割り当てても、部分が当事者資格をもつかは、第5節の問いとして残る。

二つめは、人工の構成を制御しやすくする性質と、その人工の構成に権威を置くことを正統にする性質とが、一つの尺度に並ばないかもしれない、ということである。本章の三つの箇所が、それぞれ別の向きにこの食い違いを示している。フリーデリッヒの議論をペティットの説に移すと、委託する者への完全なアラインメントは、*S* に対する支配を可能にしうる（第3.4節）。出口が常任の席に挙げた狭義の道徳的行為者の条件を当てはめると、悪をなさないように作ることは、その席に着く権利を失わせることになる（第2.2節）。フーらが調べたエージェントでは、人間による集中的な制御を避けるために停止スイッチをもたないことが、信頼を機械の目的と出力に移した（第7節）。一般化すると、よく沿っているほど、止めやすいほど、害をなさないほど、権威を内側に置くことが正統になる、という順序は、三つの箇所ですべて異なる理由から崩れうる。ここでも限定は三つあり、いずれも典拠の届く範囲にかかわる。一つは、狭義の道徳的行為者の条件が出口の価値の論文から取られており、その論文がこれを研究について述べていないことである。二つめは、フリーデリッヒが汎用人工知能を一般的に論じていることである。三つめは、フーらの対象が、研究ではなく取引や投稿を行うエージェントであり、その研究が小さな標本にもとづくプレプリントであることである。研究では、この食い違いにさらに二つの面が加わりうる。一つは、研究に役立つことと止めやすいこととの食い違いである。ソーンリーの定理では、役に立つほど識別力のあるエージェントは停止ボタンが押されることか押されないことのどちらかを選好し、辛抱強いエージェントほど、後でボタンを操作するために先に大きな費用を払う（第6.1節）。そして結果のあいだを識別し、長い調査を通じて粘り強く続けることは、研究のエージェントに求められるものに含まれる。もう一つは、研究の独立性との食い違いである。ヴィルホルトの政治的な論証では、研究が公的な知識の源として価値をもつのは、それによって統治しようとする者たちから独立していることによる（第3.2節）。そうだとすると、委託する一者にとって制御しやすい研究の構成ほど、その結果は、その者から独立した知識の源ではなくなりうる。ソーンリーの定理は、それが置く条件を満たすエージェントについてのものであり、研究のエージェントがそれを満たすかは、第6.1節で述べた限定のもとにある。また、これは制御しやすいと正統性がつねに対立することを示すのでもない。示唆されるのは、構築者が一方を示しても、それが他方の証拠になるとは限らない、ということである。

三つめは、本章の三つの問題が、一つの形を共有しているかもしれない、ということである。成員を作ること（第5節）、自らを改変し複製する人工の構成の後の段階（第5節）、同意した社会の後継者（第7節）のどれでも、いまの決定が、その決定に加わらなかった後の当事者の目標や当事者資格を定める。本章の説は、それぞれに別の答えを与えた。代表性は形成を正当化しない。自発主義の説では、同意は後に来る者を拘束しない。パーフィットの見方では、漸進的な改変のもとで、目的の変化がそれ自身の改訂にあたるかは、部分的には程度においてしか成り立たないことがある。仮にこの三つを重ねると、成員を作り、自らを改め、自らを複製する人工の構成では、自分自身を改めること（他者に対する権威の問いを生まない）と、当事者資格を尊重することを要する他者の目標を定めることとのあいだの線が、事実によって一つに定まらず、程度においてしか成り立たないことがありうる。そうであれば、シュヴィッツゲーベルとガルザの価値に開かれた設計の方針や、ブラッドリーとサードが挙げる洗脳的不正は、道徳的地位を条件として、新しい成員の形成だけでなく人工の構成の自己改変にもかかわりうるし、自発主義の論点も人工の構成の内側に当てはまりうる。この読みには二つの限定がつく。一つは、パーフィットの数える結びつきが記憶、意図、信念、欲求、性格などのそれであり、この読みはそうした状態が人工の構成に帰属できる場合にしか成り立たないことである。もう一つは、シュヴィッツゲーベルとガルザの方針が、自らの価値を反省する人間のような能力をもつAIについて述べられたものであることである。

9. 第V部が見いだしたこと

第V部は、研究を取り巻く関係のうちどれが人工の構成の内側へ移りうるか、移りうるならどの根拠のもとでかを問うた。見えてきたことをまとめる前に、それが何のための問いだったかを述べておく。人間の研究では、研究を行うことと、研究を配分し、批判し、認定し、決めることとは、多くの場合別々の当事者が担ってきた。誰が批判者か、何を基準とするか、誰が決めるかは、当事者どうしの関係として問われ、争われてきた（第9章第2節）。AIが配分、批判と認定、権威も含めて研究するとき、それらは一つの人工の構成の内側の変数になる（第1.1節）。その変数は、誰かが一度決めて固定する値であるとは限らず、構成が探究を進めながら自ら作り、自ら改めていくものでもありうる。批判者を生成し、基準を学び直し、路線の割り当て方を試して変えることが、探究の一部として行われうる。そうであれば、構成は研究のやり方そのものを変えていくことになる。マイケル・ニールセン（Michael Nielsen）とカンジュン・チュ（Kanjun Qiu）は、科学の社会的過程のありうる空間は広大でほとんど探索されておらず、その探索には少なくとも数十年か数世紀を要すると述べた。そして、その探索を担う者として、新しい過程を試し広めるメタサイエンスの起業家を構想した。二人の言う起業家は、力をもつ制度の意思決定者ではないという意味での部外者であり、科学者自身であることも多い（Nielsen and Qiu 2022）。言語モデルのエージェントに研究者、査読者、資金配分者の役を演じさせて研究の生態系を模擬し、採択率や資金の枠のような規則を変えて結果を比べる試みもすでにある（Jin et al. 2026）。それは研究を行う人工の構成ではなく、研究の取り決めを調べるためのモデルであり、そこで規則は実験者が定める明示的なパラメータである。これに対して、研究を取り巻く関係が構成の内側の変数になるとき、その探索を、研究を行う構成そのものが、自らの研究について行いうる。以下では、第V部が見いだしたことを、そのとき研究に何が起りうるかとしてまとめる。要点を先に述べておくと、研究のやり方を速く改められるようになることと、その改め方を外から確かめ、異を唱え、止める手段が薄れることとが、同じ変化の両面として起りうる。三つの関係に共通してその背後にあるのは、評価される側が、自らの評価の基準点、すなわち評価がそれに照らしてなされるもの（誰が批判するか、何を基準とするか、誰に照らすかなど）まで作り出しうることである。

二つの面がなぜ同じ変化から生じうるかを、まず見ておく。研究のやり方の改め方は、人間の研究で働いていた歯止めを経ずに進みうる。ニールセンとチュウは、科学の社会的過程のうち集合的に保持されるものがほとんど変わらない理由として、少数の大きな資金配分者と研究機関への統制の集中、誰も一人

では規範を変えられない集合行為の問題、共同体の判断と、思い描かれる将来の評価の影とによって社会的過程を均質化するネットワーク効果、そして新しい制度を育てる循環がないことを挙げた (Nielsen and Qiu 2022)。一つの人工の構成の内側では、後の三つは働かないかもしれない、最初の集中はむしろ強まりうる。ところが第9章第2.3節が仮説として示したように、関係が別々の当事者に分かれていたことは、境界をまたぐ誤りの検出や異論の当事者資格のような財を、どの当事者もそれを供給するよう設計されないまま供給しえた。そして、誰も一人では規範を変えられないのは、規範が別々の当事者に分かれて保たれているからでもあると考えられる。そうだとすると、研究のやり方の探索を妨げていた摩擦と、それらの財を供給しえた分離とは一部重なっており、関係が構成の内側へ移るとき、同時に取り除かれうる。ニールセンとチウの決定的な結果の定義に照らすと、摩擦の一つは評価の働きも担っていたと読める。二人にとって決定的な結果とは、はじめは敵対していた者や利害をもつ者をも繰り返し納得させるであろう結果である。そうであれば、既存の担い手の抵抗を含む共同体の判断の均質化の圧力は、そうした証拠を欠く変化をある程度押しとどめるふりとして働いていたことになる。構成がその抵抗を取り除けば、変化は決定的な証拠を経ずに広がりうる。これは、統制が集まるところで高い証拠の基準が課されない限り、中央化による変化は危ういとニールセンとチウ自身が述べた点にあたる (Nielsen and Qiu 2022)。先に述べた二つの面は、こうして同じ変化から生じうる。人々や制度に束ねられていた働きを別の仕方に分け、生成された研究者が探究し、その過程を通じて批判し、変更をそれに依存する後の仕事へ結びつける記録によって修正する、という可能性も、この変化の側にある。しかし、人間の取り決めが明示的な決定なしに果たしていた働きが、新しいやり方にも伴うと前提することはできず、それが設計によって定められた場合も、実際に働くかは別に確かめることを要する (第10章第5節、第11章第7節〜第8節)。

関係ごとに見ると、三つの章はそのとき起こりうることをそれぞれ見いだした。配分では、判定しやすい課題での成功が次の配分を正当化し、研究の範囲が内側から狭まりうる。一つの信号で路線の存続を判断するなら、その信号の誤りによって、多くの路線の打ち切り判断が相関しうるので、中央化と引き換えになるのは、多様性そのものより、収束を支配する基準についての誤りの独立性である。そして、誰も決めずに人間の参加が供給していた働きは、気づかれぬまま失われうる (第10章第4節・第5節・第7節)。批判と認定では、記名の査読のような人間の研究の特定の制度が同じ場面でまとめて生み出しがちだった性質、すなわち批判者の誤りの経路の違い、結果の後ろに答える誰かがいること、価値づけの感覚の共有、評価が S に見えることを、人工の構成は別々に設定できる。そのため、同じ人工の構成について説の判定が分かれ、いくつかの説のもとでは、内部の批判をよくすることがかえって S のための認定を弱めうる。独立した点検は選抜に使われるだけで衰えうるし、 S のいまの基準との一致は、基準が動いていないことの証拠にならない (第11章第8節・第9節)。権威では、そのとき起こりうることは本章第8.4節が根拠ごとにまとめた。そのなかには、局所的には擁護できる委任や、人工の構成自身の改訂の連なりが、形式上の権限を残したまま修正する能力を取り除きうること (本章第7節)、基盤モデルの上に作られた人工の構成では、従う規則の一部が提供者からの更新として届きうること (本章第6.4節) が含まれる。

三つの関係のどれにも、評価される側が評価の基準点 (批判者、 S 、基準、構成員の範囲、是認) を作り出しうるという同じ形が現れた (第11章第9.2節、本章第1.2節・第1.3節・第5節)。研究のやり方を自ら変えていく人工の構成では、この形は例外的な不正としてではなく、その探索の通常の働きとして生じうる。改めたやり方がよりよいかを、改めた側が定めた基準点で測るなら、たとえば本章第5節の三つの錨のうち現在の立場を採り、しかもそれを構成自身が定めるなら、評価は何もよくならないまま満たされうるからである。そこで、説の理由に立ち入って、基準点が評価される側から独立していることが何から成っていたのかを見る。それは、取り決めが設計されていないことではなかった。ドゥティル・ノヴァエスとデデによれば、ロンジーノは異論を育てることを求める (Dutilh Novaes and Dede 2025, §2)。またソロモンは、社会的・政治的・制度的な文脈に働きかけ、たとえば狙いを定めた研究費や賞によって、決定ベクトルの分布を変えることを求める (Solomon 2001b, 71, 74)。諸説は、評価される側だけでは定まらない基準点に、それぞれ異なる理由を与えている。エイギによれば、ロンジーノは、前もって数え上げられない範囲のなかで、共同体が出会った視点に訴える (Eigi 2016, 136, 148)。これは、どの範囲の批判がなされたかを問うものであり、すべての批判者が共同体の外で生まれることを求めるわけではない。ダグラスは、科学者が何を予見すべきだったかは、科学者がふだんから自らの研究の含意を論じ合う、かなり緊密な認識の共同体に照らして確かめられるとする (Douglas 2003)。ガリソンの説では、機械的客観性のもとで、機械は科学者が監督するものだったが、その産物は操作する者の腕前から独立していることによって重んじられた。そして機械的客観性の中心には、意志し意図する自己を通すと、自然の像が、科学者の見たいものしか示さなくなるという恐れがあった (Galison 1999, "Objectivity and the Scientific Image", "Revolutions of Objectivity: The Will and the Machine")。ヴィルホルトは、誰も実際には下していない価値判断の上での協調は、そのことが発覚すればたちまち崩れるので、科学での依拠には本物の信頼が要とする (Wilholt 2013, §4)。これらの論者は人工の構成を論じていない。これらの説が一つの条件を共有していると読むなら、その条件は、取り決めが設計されていないことより狭い。すなわち、判定に効く中身 (批判者の異議、比べられる代替案、訴える先の同業者の判断など) が評価される当事者の外から来ており、評価の時点でその当事者が置き直すことができない、という条件である。エイギが伝えるロンジーノの視点への訴え、ダグラスの同業者の共同体への訴え、ガリソンが述べる意志する自己への恐れ、ヴィルホルトの実際に下された判断という要件は、いずれもそう読める。またソロモンの制度的な手立ては、この条件が設計されていない取り決めを要しないことを示す。ただし、どの論者もこの条件をこの形で述べてはいない。自ら批判者、代替案、同業者を生成し、自らの尺度を学び直す構成にこの読みを当てはめると、鋭い帰結が出る。諸説の規範は、書かれたとおりには、誠実に、それを逃れる圧力がなくても満たされうる一方で、規範が拠っていた理由はもはや当てはまらなくなりうる。したがって、内部で生成したことだけでは、それらの基準点が諸説の求める働きを果たすとは確立されない。生成された批判者は共有仮定を引き継ぎうる。釣り合いをとる代替案も、相関した誤りが予見可能だったかどうかを測る同業者も、同じ選択によって形づくられうる。そして構成自身の改訂の外にとどまる部分がないこともありうる。そのうえ、成果が S にとって新しく、照合できる別の情報源がないところでは、協調を崩すはずのずれは発覚しないままでありうる。これは、尺度が最適化によって損なわれることとは別の事態である。変わるのは、規範が働く範囲そのものを評価される側が定めるようになることである。

ゾルマンの説は、この事態に限界があることを示唆する。ゾルマンの一時的な多様性は、事前の見込みを無作為に割り当てることで作り出せる (Zollman 2010, 16)。それで足りるのは、それが防ぐ雑音の多い証拠のもとでの早すぎる収束が、多様性を生み出す側の性質ではないからである。ここから、次の区別が引ける。防ごうとする失敗が、安全策を生成する側の性質でないところでは、生成された安全策でも足りうる。これに対して、防ごうとする失敗が生成する側の性質、すなわちその表象、基準、利害、以前の選択であるところでは、それが生成する安全策は、確かめる対象そのものに依拠する。人間の研究も、この条件を完全に満たしてきたわけではない。実際、人間の研究も後継者を育て、読み手の目を訓練し、 S を教育してきた。そして、ハイジ・グラスウィック (Heidi Grasswick) の報告によれば、ロンジーノの説では、似た位置にある知る者たちの集団が共有する背景仮定は、その集団には見えないままでありうる (Grasswick 2018, §4.3)。点検がある当事者の共有仮定を免れていたところでも、この条件が満たされていたのは、各当事者にとって批

判の中身が共同体のなかの他の当事者から来ていたという意味においてであって、共同体全体の外から来ていたという意味においてではない。批判者、基準、立場を生み出すことが、それを協調させない当事者のあいだに散らばっていることは、その助けになりえた（第10章第5節）。ソロモンが述べるような意図的な制度上の手立ても、その助けになりえた。研究では、この点が三つの面で鋭くなる。一つは、何が成功と数えられるか、すなわち重要性和基準そのものが、部分的には探究によって生み出されることである（第6章第5.3節）。二つ目は、*S*の選好と基準を形づくることが、*S*のための研究の働きの一部であることである（第4節）。そのため、*S*から引く基準点を、それが確かめるはずの研究から切り離すことはできない。できるのは、それが判定することになる結果より前に立場を記録して日付をつけることと、基準点を複数にすることである。三つ目は、研究の代替案が前もって閉じた一覧にならないことであり、「代替案」の範囲で働く規範は、その範囲を前もって定められない。どの構成も作らない世界の抵抗でさえ、点検の配置を通してしか主張の当否に関係しない。そして人工の構成は、自らの点検を書き、書き換える（第6章第2.1節・第3.3節）。

本章第2.1節の分類で言えば、批判者、*S*、構成員の範囲を定めることは構成的な決定であり、基準を定め改めることはメタ・ガバナンスの決定である。そのため、基準点が評価される側から独立していることの多くは、これらの決定を評価される側に置かないことで求められる。ただし是認は、決定によってではなく、選択肢そのものを形づくる影響によっても作り出されうるので、決定権をどう割り振っても、この形は残りうる。もともと、影響が是認を形づくりうるとしても、是認するという行為そのものは*S*の側に残る。外に残るのは、*S*の行為を要する関係、すなわち授權、是認、異議である。それらは、*S*とその行為そのものが人工の構成によって作られない限り、片側からは完成させられない（第9章第6節、本章第5節）。ペティットの説では、規範に縛られた権力でさえ、それを握る者に対して当の規範を引き合いに出して異を唱えられる限りでしか、支配を減らさない（Pettit 1996, 586）。研究に当てはめれば、これは、結果や改訂が*S*の受け入れたものから外れたときに*S*がそれと分かることを求める。*S*にとって新しい結果については、それは、その結果より前に日付をつけた記録と、それを読める当事者に依存する。

こうして、AIが研究のやり方まで含めて研究するとき、最後に問われるのは、それがなお*S*のための研究であり続けるか、すなわち*S*にとって新しい達成をもたらす試みであり続けるかである。上で立ち入った説は、この関係に異なる条件を課す。ロンジーノの説では、批判には共有された仮定を露わにしようとする視点が要る。ヴィルホルトの説では、依拠には価値判断への実質的な信頼がかかわる。ペティットの説では、権力は異議申立てに応答しうるのでなければならない。基準点を内側で生成したというだけでは、これらの条件が満たされたことにはならない。そして、上の読みのとおり諸説が一つの条件を共有するなら、構成がもたらすものが、自ら定めた基準点にとってだけでなく*S*にとっての達成でありうるかどうかは、判定に効く中身の一部がなお構成の外から来ているか、少なくとも評価の時点で構成が置き直せないところにあるかどうかにかかる。採る説に応じて、それを支えるものには、例えば、生成しただけでなく外で出会った批判者、構成が作っていない同業者、それが判定することになる結果より前に記録された立場、改めたやり方と並行して運用し続ける以前のやり方、そして構成に課す負担を構成が取り消せない異議などがある。最後のものは、本章第6.2節で当てはめたアンダーソンのモデルでは、フィードバックがそれを通じて認識上の働きをする障害にあたる。こうした中身を保つことは、第6節と第7節が重んじた修正可能性とぶつからない。固定されるといっても、評価の時点で評価される側によって動かされえないということであり、誰も修正できないということではないからである。第6.1節の言葉で言えば、これは、修正のコストを上げ、誰が修正できるかを定めながら、修正の能力そのものは取り除かない硬化に近い。

研究のやり方の探索についての一つの提案が、この点をはっきりさせる。第9章第2.3節で引いた、本書の著者がAI エージェントとともに書いたエッセイは、ニールセンとチウの議論を踏まえ、問いの選び方、資源の配分の仕方、生成する側と評価する側の関係、誤りの正し方などの異なる多くの人工の科学のシステムを組み立て、走らせ、比べることで、科学そのものを実験の対象にすることを提案した（Takagi and an AI agent 2026）。これを一つの可能性として受け取ると、それは複数のやり方を並べて運用し続ける一つの仕方であり、本節の議論はそれに次のようにかかる。並行して走るやり方を比べることは、評価される構成が作っていない基準点を与える。しかし、比べられるシステムと比べる尺度が同じ構成によって、あるいは一人の設計者によって生成されるなら、比較の基準点そのものが生成されたものであり、上で述べた形が比較の水準で戻ってくる。エッセイ自身も、人工の環境の内側で働く指標と評価に照らした改善を、その環境の外の独立した証拠と制約に照らして確かめられた知識の産出から区別することを求めている。

エッセイはまた、二つの循環を区別する。人工のシステムの内側の役割、情報の共有、評価、記憶、資源の配分は速く変えられうるが、自然、身体、社会、歴史からのフィードバックに要する時間は、AIの推論を速くするだけでは縮められない（Takagi and an AI agent 2026）。本節に照らして読めば、速い循環は、評価される側が自らの基準点を作りうる場所であり、遅い循環は世界の抵抗を運ぶ。ただし上で見たように、世界の抵抗が主張に効くのは点検の配置を介してである。研究の評価が対象とするものは、判断が要するときにはたいてい観察できないので（第11章第8節）、二つの循環は異なる速さで回り、先に挙げた日付つきの記録がそれらをつなぐ。遅いフィードバックが届く前に記録された立場と研究のやり方があれば、そのフィードバックを、その間に速い循環が改めた基準ではなく、当時受け入れられていたものに照らして読むことができる。

どの関係をどこまで内側に置けるかを根拠ごとに見てきたことは、最後に問われるとした問い、すなわちそれがなお*S*のための研究であり続けるかを考えるための手立てになる。三つの章で内側に入りえたのは主に作業の供給、すなわち配分では路線を保つ認識上の働き、批判と認定では誤りについての情報を加える点検、権威では授權された範囲での実効的な決定の働きである。ただし結果に基づく根拠のもとでは例外として、構成がよりよいことがその有能さを仮定しない評価で確かめられた範囲で、構成それ自体が当事者資格をもちうる（本章第3.1節）。それ以外では、路線を閉じることに異議を唱える当事者資格、結果を*S*のための認定として認める承認、決定を授權し異議を唱える当事者資格が内側に入るかどうかは、採る説が誰を当事者と認めるかにかかる（第9章第5節、表11.2、表12.3）。

第9章第2.3節で提案し、第11章第3節で挙げ直した四つの財についても、それぞれ見えてくることがある。境界をまたぐ誤りの検出は、誤り検出の見方では、点検がそれ自身の誤りの経路をもつかどうかにかかるので、人工の構成の内側でも生じうる。ただし時間を通じて保つには、選抜が点検の寄与を弱めるところでそれを保つ何らかの手段が要る。選抜に使わない点検は、その候補の一つである（第11章第8節）。ロンジーノの説では、批判者が生成されたか、出会った相手にかかわらず、共有された仮定は内部の批判を免れうる（第11章第9.2節）。異論の当事者資格は、採る説がそれを認める当事者を必要とし、その当事者自身が人工であることもありうる。ただし、人工の構成の部分が、採る説のもとでその資格をもつのでない限り、資格は人工の構成の外の当事者の側にある（第10章第4節、本章第5節）。部分が資格をもつのは、それ自体として当事者である場合（第5節）や、結果に基づく根拠のもとでよい結果が確かめられている場合（第3.1節）である。説明が求められた境界では、記録は説明と訂正の伝播を支えるが、説明を求めてよい資格を生まない。その資格を誰がもつかは、根拠ごとに本章第6.2節と表12.4で見た。責任を負う当事者については、義務と法的責任は規則によって配分できる一方、非難

と功績については、行為者性をめぐるさらなる問いが生じる。ある害について誰も非難に値しないときにも、将来を見据えた義務を負う者はいるかもしれない（本章第6.3節）。別々の組織はときにこれらの財を支えたが、その分離だけが源だったのでも、一律の要件だったのでもない。

人工社会について、また他の当事者がいない極限について、当事者資格と同一性の問いは本章第5節で見たとおりである。極限は人工システムによる研究の未来予測ではなく思考実験であり、そこでも探究は失敗にさらされ、結果がまだ分からない可能性のあいだで選ぶ問題に直面する。

このように第V部では、研究の組織そのものを、比較し組み直す対象、しかも人工の構成が自ら組み直していきうる対象とみたとき、人工システムが研究することに何が含まれるかを考え、特定の説が結果として生じる関係に何を求めるかを問うてきた。表12.5は、これらの所見の詳細な根拠と限界を記す。結論の章は、それらを達成、過程、研究行為者についての先の分析と結び合わせる。表の「線」は、どの関係が人工の構成の内側で完結しうるか、どれが人工の構成の外の当事者との関係に頼るかの境目を指す（第9章第6節）。「構築者の三つの主張」は、人工の構成が探究をうまく配分している、その認定に依拠してよい、他者に代わって決めてよい、という第9章第1節の主張である。

表12.5 第V部が見いだしたこと、それを支える節と支えの種類

結果	支える節	支えの種類
何が良い配分とみなされるかは前提される目的に依存する。検討した議論は前もって定められた目的を前提しており、他の議論、モデル、立場が前提する目的（第9章第3節、第10章第1節）は判定を変える	第10章第1節・第4節	論証（最適な分布を目標に相対化するキッチャーに従う）。章は、各議論がどの目的を前提しているかを述べる
配分にかかわる人間という実現者についての前提と、配分がどう組織されてきたかについての前提は、人工の構成では崩れうる。後者は人間の研究が組み直されるときにも崩れうる	第10章第2節～第5節	モデルと事例から推論。主張は、これらの前提が崩れうるということであって、崩れるということではない
批判と依拠にかかわる人間という実現者についての前提と、査読、再現、データへのアクセスがどう組織されてきたかについての前提は崩れうる	第11章第5節～第8節	関係の構造から論証し、事例は例示。一部の逸脱が発見できないことは後者の種類の前提であり、人間の研究の組み直しによっても弱まりうる（第11章第5.3節）。素人の事例における服従の合理性は、第11章第5節の立場に依存する（第11章第6節）
研究を行う者だけが、その不確実性に気づける位置にいることがある。その選択の自動化は、責任の配分についてさらなる問いを生む	第6.3節、第11章第4節	ダグラスの前提は位置についてのもので、その移し替えは、不確実性に気づける位置に人工の行為者を置く。完全な移譲に反対するダグラスの判断は主にこの前提に依り、ほかの理由は、人間の研究の記録と監督のされ方と、一部は人間の研究者の監督への応じ方にかかわる。三つめの理由のうち、誤りを見いだすことが監督を呼び込むなら見いだす誤りを少なくする理由が生じるという段階は、監督を避ける理由をもつ行為者一般への当てはめとして移る。ダグラスはこの責任を仕事に対する権限とも結びつける。この移し替えは、機械の責任と各段階での承認という網羅的な二択を確立しない
それらの前提が課していた問題の多くは、消えずに形を変える	第10章第3節～第4節、第11章第5節～第8節	信号については論証（第10章第3節。第9章第4節の代理指標の説明とともに）。多様性については、モデルと、現在の言語モデルについての経験的研究から推論（第10章第4節）。批判と依拠については、関係の構造から論証（第11章第5節～第8節）
決定なしに供給されていたものは、設計の問題、誰も決めない別の過程の問題、あるいは供給されないことの問題になる	第10章第5節、第11章第7節～第8節、本章第6.4節・第7節	検討した事例とモデルから推論
どの路線が成功するかを前もって知らないことは、統合が取り除かない探究の条件である	第10章第1節～第2節・第4節・第5節	ストレヴンズ、キッチャー、ショー、ウーとオコナーの議論から、部分的にはモデルから、前もって定められた目的について論証。これらの議論によれば、研究者に対する統制は、彼らがどこで働くのが最善かについての知識を与えない。前もってどこまで見積もれるかは、すでに何が知られているか、また追究が確立した計画の内部にあるかによって変わる（第10章第2節）。この条件が配分にどれほど効くかは、路線をどれほど少ない費用で決定的に試せるかによって変わり、それは問題がどう分割されてきたかにも部分的に依存する（第10章第4節）
独立した批判は無効化の関係に依存し、それは統合が取り除かない探究の条件である	第11章第1節～第2節	批判を誤り検出として扱う見方のもとで成り立つ。批判的文脈説では、点検が何に反する証拠となるかは背景仮定に相対的であり、線は人工の構成の外の当事者との関係、ここでは S との関係の側へ動く（第11章第1節、表11.1）
誰の問いが、 S が重要とみなすものの出発点になるか、認定が S に依拠する理由を与えるか、誰が他者のために条件を定めてよいかは、外の当事者との関係である	第10章第6節、第11章第5節・第9節、本章第1節～第3節	名指した説のもとで成り立つ。線は採る説とともに動く
線がどこに引かれるかは、知識、価値、正統性の説に依存する	第9章第3節、第11章第1節・第3節・第5節・第9節、本章第1.2節・第1.4節、表12.2	論証（地図そのもの）
価値の説は、誰が目的を課してよいかに答えない。純粋にエビストクラシー的な見方を除けば、有能さだけではそれを課す当事者資格は与えられない	本章第1.4節	考察した正統性の説の範囲内で、ペーターの報告に従って論証

結果	支える節	支えの種類
構築者の第三の主張は、有能さや成績だけではどの根拠のもとでも成り立たない。結果に基づく根拠では、よりよいことが独立に確かめられた範囲で決めてよいという主張として、実際の授権と民主的に読んだ共和主義の根拠では、授権され異議を受けうる実効的な力としてだけ、同意の根拠では同意されたものの範囲で（後に来る者に及ぶかは決まっていない）、手続き的な読みでは共有知である規則が拘束する範囲で成り立ち、自己統治の根拠には主張の形が合わない	本章第1.4節・第3節・第7節・第8.2節	ペーターの見取り図にもとづき、根拠ごとに論証。ラズの独立性テーゼと共和主義の判定の人工の構成への当てはめは推論。主張は、どの根拠に答えるかを示さないかぎり評価できない
比較についての事実をすべて知っても、 <i>S</i> の目的の一部が順序づけられないことがあり、それを完成させるのはコミットメントである。誰のものかは第二の問いに近い	本章第1.2節	チャンの説（同等）のもとで成り立つ。社会の順序づけへの移し替えは本書による。パレート規則のような順序づけでは成り立たない
人工の構成のどの部分が権威をもちうるかは根拠によって違う（よりよいことが確かめられた範囲で決める人工の構成、自ら路線を取り上げる部分、規則が定めない場合を決める部分、共同最終決定者の席に着きうるもの）。このうち規則が定めない場合を決める部分がもつのは実効的な権威である。当事者資格を、それ自体として当事者であることに基かせる説では、それをもつのはそうした当事者である部分だけである	本章第1.4節・第2節・第3.2節・第5節・第7節	根拠ごとの論証。レイエルスらと出口の説の人工の構成への当てはめは推論
人工の構成の有能さは、純粋にエピストクラシー的な見方を除けば、それだけではそれ自身に当事者資格を与えないが、人間の <i>S</i> の当事者資格を名目だけのものにしうる	本章第1.4節・第3.1節・第4節	ペーターの報告に従う論証と、ロンジーノ、キッチャー、谷口ら、高橋の当てはめ。ロンジーノとキッチャーの説を人間の専門家と公衆から人工の構成に移し、すでにAIを論じる谷口らと高橋の説を、研究の目的を定める当事者資格に当てはめたもの
委託する者に完全に沿わせることは、支配からの自由の根拠のもとで、 <i>S</i> に対する支配を可能にしうる	本章第3.4節	フリーデリッヒの議論（支配の語は用いていない）とベティットの支配の概念の当てはめ。人工の構成を支配する行為者とみなすことを要しない
決定が答える構成員の範囲は、部分的には、決定を問われている当事者によって見定められうる	本章第1.3節	フレイザー、デューイ、ワイルとタンらの当てはめ。誰が影響を受けるかを知覚するうえで人工の構成がほかの者より有利な位置にいる場合という条件つき
成員を作れる <i>S</i> では、誰が作り目標を定めるかが集計に先立つ権威の決定であり、代表性はそれを正当化しない	本章第5節	キッチャーの基準、シュルマンとボストロム、ワイルとタンらの当てはめから論証。作られる成員への不正は、道徳的地位を条件とする
後の修正の原則は、必要ところほど弱い。修正する能力を取り除く決定は、その修正可能性から正統性を引き出せない	本章第7節	ジーナスとスターリング、レンツォとグリーン、ベティットに従い、組み立てた事例について論証。自発主義の説では同意は後継者を拘束しない
人工の構成の誤りを明らかにできる評価の経路は、極限に至るまで、大半の根拠で繰り返し現れる	第7節、第11章第2節～第3節、第10章第4節	組み立てた事例について根拠を比較して引き出したもの。共和主義の判定は、一般的な形ではベティットのもの、人工の構成への適用は、ここで推論したもの
構築者の三つの主張には、内部で測定された成績を越える関係についての証拠が必要である	第10章第4節・第6節、第11章第9節、本章第1.4節・第4節・第7節	検討した説の大半のもとで成り立つ。どれほどかは、人工の構成の外の当事者との関係の線がどこに引かれるかに依存する
人工の <i>S</i> では、人工の構成の外の当事者との関係は成員のあいだの関係として続く。極限では探究の条件が残る	第5節、第10章第1節、第11章第1節	論証。誰が当事者資格をもつかは、採る正義の説に依存する。極限は思考実験
異論の認識上の働きは、人工の構成の内側で設計によって定めることも、分けられた権能によって供給することもできる。ただし、前提を共有するインスタンスのあいだでは、それが守る異論は内側に源をもたないかもしれない。異論の当事者資格は、人工の構成の部分が、採る説のもとでその資格をもつのでない限り、人工の構成の外の当事者の側にある	第10章第4節、第11章第3節・第9節、本章第5節	認識上の部分はモデルから推論し、ソロモンの説のもとで成り立つ。当事者資格の部分は第9章第2節・第5節の定義から論証
第9章第2.3節の四つの財における、内部の能力と他の当事者との関係の分かれ方は、採る説によって異なる	本節、第10章第4節、第11章第2節・第3節・第8節・第9節、本章第6節	各章を読み合わせて引き出したもの
いくつかの説は、評価される側だけでは定まらない基準点に理由を与える。判定に効く中身が評価される当事者の外から来ており、評価の時点でその当事者が置き直せないという一つの条件を共有すると読むなら、批判者、代替案、同業者を生成し尺度を学び直す自己改訂的な構成は、諸説の規範を書かれたとおりに誠実に満たしながら、規範が拠っていた理由がもはや当てはまらない状態になりうる	本節、第10章第4節、第11章第3節・第9.2節、本章第5節・第7節	ロンジーノ、ソロモン、ダグラス、ギャリソン、ヴィルホルト、ベティットの当てはめ（いずれも人工の構成を論じていない）。共有される条件は本書の読みであり、どの論者も述べていない。すべての批判者が外で生まれる必要はなくても、共有された仮定は批判を免れうる。防ごうとする失敗が生成する側の性質でないところでは、生成された安全策も足りうる（ゾルマン）。人間の研究がこの条件を満たしていたところでも、それはおもに散らばりによって、各当事者の中身が他の当事者から来ていたという意味においてであり、意図的な制度上の手立てもそれを支える

第13章 結論：機械が研究するとはどういうことか

研究、人工の行為者、探究の組織

1. 四つの問い

本書は、機械が研究するとはどういうことか、そして機械が研究するとどうなりうるかを考えてきた。その手がかりとして、第1章は研究を、社会 S にとって新しい認識的達成をもたらそうとする試みとして特徴づけた。

研究をこのように捉えるなら、まず、認識的達成とは何か、それはいつ、どこで実現するのかを考えることになる。第2～3章は、知識と理解についてそれを考え、機械が生み出したものがそうした達成にあたるのはどういうときかを検討した。また、失敗に終わった研究や、まだ終わっていない研究も研究と呼ばれることから、必ずしもある認識的達成にたどり着かなくても、ある営みが研究と呼ばれることはあるように思われる。研究を成功そのものではなく試みとして特徴づけたのは、そのためである。したがって、答えがまだなく、失敗もありうるもとの、試みがどう進むのかを問うことになる。機械が問うとはどういうことか、候補や表象はどう利用可能になるのか、結果が何にかかわるかはどう判断され、基準はどう改められていくのか。第4～6章はこれを扱い、そこから、出力を出すことと探究することの違いがどこにあるかを探った。

加えて、第1章は、人間と機械のどちらをも含みうる特徴づけとするため、研究する者 A を明示的には含めていない。そのため、研究する者 A と、研究によってなされる達成がそれにとって新しい社会 S とを、別々のものとして扱える。第7章は A について考えた。研究する者 A は、どこまでを含むのか。人の場合も機械の場合も、研究をどの境界の内側に帰属させるのかが問われる。また、機械が自律的に研究しているとはどういう状態なのか、それはどの点でどこまで自律的なのかも問われる。第8章は S について考えた。研究する者 A と S とが一致するとは限らないなら、研究によってなされる達成は誰の知識になるのか。第8章は、 S が人間の社会である場合、人間と機械が混じる社会である場合、機械の社会である場合を分けて、この問いを考えた。また、研究する機械は、身体や認知や作られ方などの点で人間と異なりうる。そうであれば、何をまだ知られていないとみなし、何を追う価値があるとみなし、何をよい理由とみなすかが、 S の成員と食い違うこともありうる。第8章は、その食い違いがどこから生じうるのか、そしてそれが AI アライメントの問題とどう似ているのかも考えた。さらに、達成が S にとって新しいものであるかぎり、研究は S との関係のなかに置かれる。 S の未知は数えきれないのに探究の手段はかぎられているので配分が、成員は結果の多くを他者を通じて受け取るので批判と認定が、そして S には一つの声をもたない多くの成員がいて、探究の帰結は S の外にも及ぶるので、その両方についての権威が生じる（第1章第9節）。では、機械がそれらの関係まで担い、それを自ら作り改めるとき、研究には何が起るのか（第9～12章）。

各章の検討を経ると、「この機械は研究している」という主張は一つの主張ではなく、一つひとつが成り立つことも成り立たないこともありうる主張の束として見えてくる。これらの主張は関わりあうこともあるが、一つが成り立つても、ほかが決まるとは限らない。たとえば、機械が正しい結果を出すことと、その結果が S の知識になることとは、別のことでありうる（第2章）。出力を出しつづけることと、失敗がどこにあるかを見定め、必要なら基準を改めながら探究することも、別のことでありうる（第6章）。研究を遂行したのが機械であっても、その達成の担い手が機械であるとは限らない（第1章第9節、第7章）。そして機械が研究を取り巻く関係まで担うとき、それがなお S のための研究であり続けるかは、さらに別に問われる（第12章第9節）。そのため本書は、四つの問いを章を分けて考えてきた。

表13.1は、そうした束の一部を集める。それは議論の応用であり、研究の代わりに定義ではない。各行は、証拠が確立しうるものと、それだけではまだ決まらない問いとを組にし、最後の列は、両者が切り離されることを本書がどこで論じたかを示す。それぞれの種類の主張が何を構築し、開示し、あるいは提示する必要があるかは、そうした主張を検討した箇所所述べた。たとえば表7.1、第10章第7節、第11章第9.2節、第12章第8.2節である。

表13.1 さらなる論証を要する推論（抜粋）

証拠が確立するもの	さらなる問い	主な議論
真なる、または新しい出力	それは知識か、知っている者がいるならそれは誰か、そしてそれは関連する社会の認識的状況に入ったか	第2章
規則性についての知識	なぜそれが成り立つかを誰かが理解しているか	第3章
研究エピソード	関連する変動にわたって能力が持続するか	第7章第3節
ある境界での能力	そこで研究者という地位が、どの目的にとって説明的か	第7章第3節
ある点での自律性	問い、方法、基準にも及ぶか	第7章第4節
相互作用する人工エージェント	知る社会を形成しているか	第8章第1節
配分への統制	その選択が探究を組織する目的に奉仕するか	第10章
多数の、または当初は情報を加える批判者	批判が関連する情報を供給し、それを続けるか	第11章第2～3節・第8節
有効な内部の批判	認定が関連する社会に依拠する理由を与えるか	第11章第5節・第9節
形式的な権利や有能な決定	採る根拠のもとで権威が根拠づけられるか、またそれを求める根拠のもとでは、実効的に異議を申し立てたり改めたりできるか	第12章

以下の節では、研究を第1章のように捉えることで、機械の研究について何を区別し、何を問えるようになったのかを考え（第2節）、達成、試みの進み方、研究する者と社会、研究を取り巻く関係について、各章で検討したことをたどる（第3～6節）。第7節では、それらを章をまたいで見直したとき、機械が研究するとはどういうことかについて何が見えてくるのか、そして何がお問いとして残るのかを考える。

2. 作業的特徴づけによる研究の捉え方

本書は、第1章の次の一文から出発した。

ある社会 S に相対的に、研究とは、 S にとって新しい認識的達成をもたらそうとする試みである。

研究をこのように捉えたうえで、本書は、機械が研究するとはどういうことか、機械が研究するとどうなりうるかを考えてきた。本節では、この捉え方が、研究するのが機械である場合に何を区別し、何を問えるようにしたのかをたどる。そのうえで、この捉え方が脇に置いたものと、問いに答えるために本書が何に頼ったのかを見ておく。

この特徴づけの三つの語は、研究するのが機械である場合に、それぞれ別の区別を保つ働きをした。「試み」は、機械の活動を、その成功と同じものとみなさずにすむようにした。 S にとって新しい知識へ向けて組織された活動は、誰もまだその結果を知らないうちにも、失敗したときにも研究でありうる。逆に、真なる出力があるだけでは、それを生んだ実行が研究になるわけではない。達成に因果的に寄与することは、それだけではその達成への試みではないからである（第2章第4.5節・第9.1節）。「 S にとって新しい」は、機械自身が何かを新しく得ることと、 S にとって新しい達成がもたらされることを分けた。そのため、研究する者 A と S とを別々に扱い、機械が知っていても結果が S の認識的状況に入っていない場合や、 S 自体が機械の社会である場合を考えることができた（第2章第9.2節、第8章第1節）。「認識的達成」は、知識と理解の両方を扱えるようにした。そのため、機械が信頼できる予測を出しながら、予測された規則性がなぜ成り立つかを共同体のどの人間の成員も理解していない場合のように、両者が離れる場合を捉えることができた（第3章第4節）。

特徴づけに何を含まなかったかも、三つの語と同じように意味をもった。研究する者 A を明示せず、研究を一連の操作の一覧で定めず、研究する者が人間であることを条件としなかった。また、もたらされる達成の重要さや有用さなどの価値も、配分、批判と認定、権威など研究を取り巻く関係も、研究であることの条件にはしなかった（序論第1節・第5節、第1章第1節・第8節・第9節）。そのため、機械による研究も、機械が研究の新しいやり方を作り出す場合も、はじめから比較の外に置かずにすんだ。達成の価値と研究を取り巻く関係は、研究であることの条件としてではなく、研究が S とつながるところとして、第V部で扱われた（第9章第3節）。そして第V部の最後に問われたのは、機械が研究を取り巻く関係まで作り改めるとき、その研究がなお S のための研究であり続けるか、すなわち S にとって新しい達成をもたらす試みであり続けるかであった（第12章第9節）。そこでは、この特徴づけそのものが、変わっていく研究を測る物差しとして使われている。

さらに、研究を S に相対的に捉えると、研究するのが機械である場合には、 S をどう定めるかそのものが、二つの点で問いになる。一つは、 S を選ぶ自由には限りがあることである。機械が理解したことは、人間の社会が理解していなくても、その社会にとって新しいものでありうる。しかし、それが研究の記述として適切かどうかは、 S を都合よく選ぶことでは決まらず、主張と活動の内容による（第3章第9節）。もう一つは、 S が動くことである。機械が S の成員と異なりうるなら、何を未知とみなし、何を追う価値があるとみなし、何をよい理由とみなすかについての S の基準も、機械とともに動くことがある。そのとき、ある食い違いが乖離なのか、誤りなのか、得なのかは、いまの S 、これから S 、正しさが問われる場合には世界という、名指された立場から判断するほかない（第8章第3.2節・第3.3節）。研究のやり方を改めていく人工の構成について、改めたやり方がよりよいかを何に照らして測るかという問いも、同じところに行き着く（第12章第5節）。さらに、第11章第9.3節が示唆したように、研究する者と S とが重なるところでは、「 S のための認定」が指すものが、「どの社会にとってか」という問いへ移るかもしれない。作業的特徴づけは S に相対的に研究を捉えるが、どの時点のどの S に相対的なかは、一文のなかでは定まっていない。機械が研究する場合には、それをそのつど名指さなければ、研究が S にとって新しい達成をもたらしたかどうかを問うことができない。

一方で、この捉え方が脇に置いたものもある。本書は作業的特徴づけを試すことを目的としなかったが、第1章第8節では、特徴づけを多様な研究実践と比べ、どこまでを覆い、どこで別の捉え方と分かれるかを見ておいた。そこで比べた研究実践のなかには、この捉え方とそのままでは重ならないものがあった。特徴づけの名指す達成は理解を含むので、ポルフドルフが知識と並べる理解は、はじめから範囲の外に置かれるわけではない。しかし、それだけでは、特徴づけがその理解と実質的に合うかは決まらない。ポルフドルフが重んじる、まだ知らないことへの開かれ、スクリプナーの言う把握、命題にならない実践的な知、関係的な説は、この捉え方だけでは捉えきれないものとして残っている。たとえばスクリプナーの提案では、芸術制作が生むのは、物事がどうであったか・どうであるかの知識ではなく、見る仕方とあり方を与える把握であり、それを把握することが知識や理解などの認識的達成にあたるかには、別の論証が要る。リーズンとブラッドベリーは、探究の意義と、探究が営まれた関係とを、妥当性と質の問いに含める。ウィルソンの説では、妥当性の基準を満たし、あるいは「事実」を正確に記述している研究でも、関係から切り離されていれば、先住民のパラダイムのなかでは受け入れられない（第1章第8節）。ここでの関係は、ウィルソンの説で研究が説明責任を負う関係であり、第V部が扱う配分、認定、権威の関係ではない。この二つの立場は、本書が達成を価値や関係から分けることと、そのままでは両立しない。「認識的達成」という語を用いても、こうした違いが消えるわけではない。

また、この特徴づけは抽象的なものなので、実際の研究でどのようなことが起きているのか、特徴づけが名指す試みや達成が具体的にどのように実現されるのかは、そこには書かれていない。そのため、研究の過程そのものは、この捉え方だけでは描けない。たとえば、問いの形成、装置の開発、基準の改訂は、どのような探究が可能かを变えうるし、それ自体が目標とする達成でなくとも、組織された試みに属しうる。それを達成として主張するなら、それ自身についての説明が要る（第1章第1節）。しかし、そうした変化がどう生じ、進歩にどう寄与するかは、達成へ向かうという方向づけからは説明されない。第4～6章が、実際の研究の事例を取り上げ、それをめぐる議論をたどって過程を描いたのは、そのためである。逆に、そうして描いた過程、すなわち

答えがまだなく失敗もありうるもとの探究が進むさまは、研究以外の営みにも見られるので、それだけで研究を特徴づけることはできない（第4章第9節）。

この捉え方が与えたのは、主に問いである。研究するのが機械であるとき、何が知識や理解にあたるのか、研究がどう進むのか、誰が研究しているのか、誰が批判し、誰が決めてよいのかは、どの説を採るかによって答えが分かれる。そのため各章は、多くの場合、答えを一つに定めるのではなく、取り上げた説ごとに、その説を当てはめると機械の研究について何が出てくるのかをたどってきた。過程について事例とそれをめぐる議論に頼ったのと同じく、ここでも答えは、この捉え方からではなく、各章が向き合った説と事例から来ている。そうした説ごとの帰結の多くは、それぞれの説が機械の研究に何を求めるかとして読むこともできるので、研究をこのように捉えない読者にとっても、検討の材料として残りうる（第1章第9節）。

3. 知識と理解は、いつ、誰において実現するのか

第2～3章は、知識や理解をめぐる認識論や科学哲学の議論を手がかりに、機械が生み出したものが知識や理解にあたるのはいつか、それは誰において実現するのかを考えた。人間が研究する場合には、結果を生んだ者が、それを信じ、知り、理解する者でもあることが多い。機械が研究場合には、それを前提できない（第2章第2.1節、第3章第6節・第9節）。そのため、両章で検討したことの多くは、研究を行ったもの、知る者や理解する者、そして *S* が分かれうることにしかかわる。

ある活動が *S* のための研究であること、*S* にとって新しい知識がどこかで実現したこと、その結果が *S* の認識的状況に入ったことは、別々に成り立ちうる判断である（第2章第7節・第9節）。第2章は、誰にも読まれていない、機械による発見の事例で、このことを検討した。結果を書き記し、公開しておくことは、人間の社会的知識にとって、どの説にも共通の十分条件ではない。何が加われば結果が *S* の知識になるのかは、社会の知識を、適切に機能する社会の認知システムが生み出す状態とみるバード、集団の知識に、その実働成員のかかなりの割合が結果を信じていることを求めるラッキー、共同体の批判を経て成員に受け入れられることを求めるロンジーノの説で分かれ、さらに機械が知る成員に数えられるかによっても変わる（第2章第7節）。そして、結果が *S* の認識的状況に入るための経路は、研究を行ったシステムの内部にあるとは限らない。

機械についての知識の主張は、どこまでを一つの単位とみなすかを定めないと確かめられない。人間の研究者であれば、信念をもつ者、証拠の範囲、能力の座は、一人の人であることでおおむね定まる。機械では、それらがあらかじめ定まっているとは限らず、インスタンス、外部の記録、道具、ほかのシステムの結びつき方によって、何通りにも引けることがある。それを引くのが作り手であれ、自らを組み替えるシステム自身であれ、どこに引くかで判定が変わる（第2章第9節）。たとえば証拠主義のもとでは、検索できる記録の蓄えをシステムのもつ証拠に数えるかどうかで、同じ結論が正当化されているかどうかが入れ替わりうる（第2章第4.4節）。また、知識に、成功が行為者の有能さによるものであることを求めるソーサの説を当てはめれば、人間による隠れた選別は、出力を安全にしても、その成功をモデルの功績にはしない。成功が誰の功績になるかは、有能さがどこに座を占めるとみなすかによる（第2章第4.2節）。単位が要することは比べた説に共通するが、なぜ要るかは説ごとに違う。

システムが表に示すものと、システムを実際に左右しているものとが食い違うこともある。基礎づけを求める説のもとでは、よい理由が報告に添えられていることと、結論がその理由によって保たれていることは別である。訂正された報告が、作動しているコミットメントを変えないままであることもありうる（第2章第4.3節）。バイズ的確証理論でクルーピが区別する二つの確証の概念を当てはめれば、候補を大量に生成し、その大半が偽であるシステムでは、チェックに合格して主張への支持が上がっても、主張が確からしくなるとは限らない（第2章第5.2節）。チェックが産出者と同じ基盤システムのインスタンスによって行われるなら、どれほど多くのインスタンスが一致しても、独立な誤りの検出や多様な証拠を加えないかもしれない（第2章第5節・第5.2節・第9.1節）。ただし、評価者が知識を帰属させる証拠を十分にたないことは、それだけでは、主体が知識を欠くことを示さない（第2章第5節）。そして、証明、外れた予測、批判によって自らのコミットメントを改められるシステムは、もつともらしきや承認だけに向けて最適化されたシステムとは、対象に対して異なる仕方に関係している（第2章第3節）。

知識に届かない変化もある。何がまだ分からないかを明示することを認識的な進歩に数えるヴィルホルトの説のもとでは、閉じた問いや出力した答えの数で評価すると、証拠がまだ決着をつけていない問いを早々に閉じてしまうシステムほど、高く評価されかねない（第2章第8節）。一つのスコアが進歩そのものを表すと主張すれば、それだけで進歩の説のどれかを選ぶことになる（第2章第8.3節）。

第3章は、知識と理解が離れうることについて論じた。ある規則性が成り立つという知識と、それがなぜ成り立つのかについての説明的理解とは概念的に分離できる（第3章第4節）。なぜそうなのかの理解に説明を要求する説のもとでは、機械が信頼できる予測を出し、予測された規則性についての知識が採る説の条件を満たしていても、その規則性がなぜ成り立つかを、共同体のどの人間の成員も理解していないことがありうる。これに対して、説明なしにもなぜそうなのかの理解を認めるリプトンの説では、説明がないことだけからは、そう判定できない。予測器がデ・レフトの意味での理論であれば、現象の理解に理解可能な理論に基づく説明を求めるデ・レフトの説では、予測の成功は使い手がその理論を理解していることを前提とするが、規則性がなぜ成り立つかの理解までは前提としない。また、バイスパルトの説では、科学者が予測器そのものを走らせて推論しなければならないなら、理解をもつのは人間とコンピュータの結合したシステムであって、補助を受けない人間ではない（第3章第4節）。

不透明性についても、それがどこにあるのかが問われる。ハンフリーズの定義では、不透明性は、誰にとって、いつ不透明なのかに相対的である。また、デュエダらの議論では、候補を生成する経路が不透明でも、それを受け入れることを根拠づける経路まで不透明である必要はない。ただし、源泉の信頼性が根拠づけに入る信頼性主義などの説では、生成器の評価も根拠づけの一部になる。たとえばリキッド・テンソル実験では、検査器が人間の判断に残したのは、導出の各段階が妥当かどうかではなく、検査された言明が意図された定理であるかどうかであった（第3章第5.1節）。内部の透明性は、どの見方でも十分ではなく、それが必要かどうかは争われている（第3章第5.3節）。

誰が理解するのも、別々の主張に分かれる。システムがほかのエージェントや共同体の理解を助けること、システム自体が理解すること、集合体が理解することは別の主張であり、第一の主張は第二を要求しない（第3章第6節）。システム自体の理解は、説ごとに試し方が異なる。ムーディとデ・レフトは、分布外の科学的な問いへの正答という間接的な証拠に基づいて、現在の言語モデルに、脆いながらも自明でない程度の理解を認める。パーマンらは、

理解の程度を「何か」「なぜか」「もし〜ならどうなるか」という問いで試すことを提案する。これに対してポトチュニクの説では、理解には理解の感じが要るので、その感じが人工システムにありうるかが先に問われる（第3章第6節・第9節）。そして *S* と担い手とは区別される。人工システムが理解したことは、人間の社会自体が理解していなくても、その社会にとって新しいものでありうる。ただし、ボーグ、シュスター、シュトルの区別を当てはめれば、十分に広い聴衆に伝達可能な説明がまだないかぎり、それは三人のいう科学者の理解にあたり、まだ科学的理解ではない。三人の説では、科学的理解には十分に広い聴衆に伝達可能な説明が要るからであり、システムは、ほかの物理学者がたどれる説明をダイソンが与える前のファインマンの位置にある（第3章第9節）。

理解が重んじられる理由の一部は、それが探究を続ける力に結びつくことにある。分かったことを新しい事例に当てはめること、うまくいかないときに失敗がどこにあるかを見定めること、さらなる問いを立てること、批判すること、覚えておくことは、ひとつの結果の性質ではなく、探究がどう組織されているかに依存する（第3章第7節・第8節）。そのため、理解を目指さない研究であっても、長く続く自律的な研究には、理解と結びつく能力が要りうる。また、ひとつひとつの結果が根拠づけられていても、そうした結果を繰り返して生み出し利用するシステムや共同体が、認知的な能力を失っていくことがありうる。メッセリとクロケットは、そのリスクの一つの形を、科学者が実際以上に理解していると信じてしまう理解の錯覚と、AIに向けた問いと方法がほかの進め方を押しのけるモノカルチャーとして描いた（第3章第8節）。*S* が人間と機械の混じる社会や機械の社会であれば、判断を担う成員が、それに必要な有能さを備えて形成され続けるかが問われる。探究を続ける力が探究の組織に依存するので、達成の問いは、探究がどう組織され、どう進むのかという第4～6章の問いにつながる。

機械の理解は、人間に似ているほど進んでいる、という一つの物差しでは測れない。何が求められるかは用途によって変わるからである（第3章第9節）。たとえば人間に向けた説明、あるいは説明なしの理解を認める説では人間が使える別の理解の乗り物は、*S* が人間であり、理解が *S* の認知的状況に入ることが目指される場合には欠かせない。理解がシステム自身の推論にだけ使われる場合や、*S* の人工の成員に伝えられる場合には、それは要らないかもしれない。また、理解を能力とみなすパーマンらの説などでは、人間の内観に似ていない表象も、理解を支えるかもしれない。機械が研究するとき、知識や理解は、*S* の成員のうちにも、機械の側だけでも、人間と機械の結合した全体にも実現しうる。どこにも実現しないまま、結果だけが積み重なることもありうる。どこに実現したのか、それが *S* の認知的状況に入ったのか、*S* がそれを評価する力を保っているのかは、出力の量からは分からず、そのつど別に問われる。

4. 研究という、答えがまだない中で進む営みでは、何が行われているのか

第4～6章は、科学哲学や、問いと探究をめぐる哲学の議論と、実験ノート、プロジェクトの公開記録、研究者の回想など研究の現場の記録とを手がかりに、答えがまだ得られておらず失敗もありうる研究という営みが、どのように進んでいくのか、その過程で実際にどのようなことが行われうるのかを検討した。ある変化が進歩を構成するのか、後の進歩を促すだけなのかという問い（第2章第8節）とは別に、変化がどう起こり、互いにどう関係しうるかを見ようとしたのである。第1章第7節で見たように、*S* にとって新しい達成への試みは、*S* においてまだ決着していない点に向かう。まだ決着していない点については問いが立ちうるので、第1章はこのことを最小限の問いの構造と呼び、それがどの意味での問いなのかは決めずにおいた。問いが最初から言葉で定式化されている必要はない。答えが *S* においてまだ知られていなければ、何が答えになりうるかも、何を比べ、何を証拠とみなすかも、前もっては決めにくいことがある。ただし、答えが新しくても、そこへ至る手続きが前もって実質的に確立されている研究もある（第5章第1節）。第Ⅲ部は、保健局の技術ノートが例示する三つの側に対応する、三つの関係の側から同じ探究を見た。第4章は、探究とそれを組織している未決着の点との関係を、機械が問うとはどういうことかという問いとして考え、第5章は、答えの候補や表象、手段がいかにして利用可能になるのかを、第6章は、結果が何にかかわり、何を改めるのかを考えた。同じ要素が複数の関係に立つこともあり、三つの関係は探究のなかの関係を尽くすものではない（第Ⅲ部の導入）。

理想化された研究の過程は、仮説を立て、テストし、受け入れるか退けるかという段階や、アイデアを出し、実験し、結果を分析してまとめるという手順など、時間の順に並ぶ段階の連なりとして描かれることが多い。三つの章が取り上げた事例と議論に照らすと、研究の過程は、この描像よりはるかに複雑でありうる。候補を形づくること、追求すること、評定することは、網羅的な一覧でも段階の系列でもない区別できる機能であり（第5章第4節）、形成と評定は交互に入り組んで、一方が他方の材料を生むこともある（第5章第4.1節）。過程は、時間の順に並ぶ段階として書けないこともあり、別の順序をとることも、課題や作業が人々、道具、記録のあいだに分散することもある。ある点での変化が、ほかにどのような変化が可能かをえ、そうして変わった探究の構成がさらに変わっていくこともある（第5章第1.1節・第6.2節）。どこで何が起きているかは、どこに境界を引くかによって見え方が変わる（第5章第4.2節・第6.5節、第7章第2節）。そして、段階と段階のあいだだけでなく、一つの段階としてまとめられているもののなかにも、多くの仕事がありうる（第4章第3節、第5章第1節、第6章第1節・第6節）。ただし、これらの事例が示すのは、ある関係が起こりうることであって、その頻度ではない。「問いの生成」「仮説の生成」「検証」といった工程の名前は重要な可能性を名指しており、第Ⅲ部が描いたことは、あらゆる研究に新たな必要条件を課すものでもない（第5章第1節）。

過程がこれほど複雑でありうることは、最終的な成果物や論文だけを見ていると分かりにくい。成果物や論文は、過程をそのまま映すとは限らないからである。公表される配列や説明は、作業の順序や経過と一致するとは限らない（BLOOMの実行中の記録と後の論文の食い違いなど）。ただしこれは、論文が不正確であることを示すものではなく、きれいな順序が、課された規律を反映していることもある。第5章は、同時期の記録と後の記録のどちらかを真の過程とみなすのではなく、両者の食い違いを調べる材料にした（第5章第2節）。最終的な成果物は、主張の評価には証拠を与えうるが、それだけでは、人工システムがそれに帰属される開かれた研究能力をもつことの十分な証拠にならない。逆に、完全な活動の記録さえ、それだけでは認知的な有能さや結果の真理を立証しないだろう。両者は異なる問いに答える。過程を考慮した評価に要るのは、すべての内部状態の開示ではなく、課題に関連する来歴である（第5章第2.1節）。機械による研究では、これが具体的な問題になる。ルオ、カシルザデ、シャーが調べた二つのAI科学者システムでは、どの実験を報告するかを選ぶ内部の仕組みがテスト集合の結果を参照でき、テストでの成績のよい実行を優先していた。三人の検査では、訓練データの外にあるよう設計した合成課題と、均衡のとれた小規模な事例群が使われ、欠陥のある事例の多くはシステムの実行に手を加えて作られた。その事例群で、論文だけを与えられた言語モデルに基づく監査器は欠陥を偶然に近い水準でしか見分けられず、実行の記録とコードも与えると正答率は82パーセントになった。著者た

ちは、記録に現れない欠陥は見逃されうるとも述べている。こうした監査が、人間の社会や混合社会である *S* にとっての評価に役立つかは、*S* が人工の監査器に依拠できるかによる（第5章第2.1節、第11章第5節）。

探究とそれを組織している未決着の点との関係について、第4章は、「機械が問うた」という主張を分けることから始めた。その主張は、疑問文を出力したこと、情報を引き出す行為をしたこと、疑問の態度をもったこと、その後の作業が未決着の点に向けて組織されたことのどれかを言う。A が機械であるとき、これらは別々に成り立ったり、別々に欠けたりする（第4章第1節）。問う行為を情報を引き出す働きによって同定するワトソンの説を当てはめれば、争点は、機械に好奇心があるかではなく、情報を引き出す働きを、どの活動のなかで、どの観点から帰属するかに移る。最小限の問いの構造は問いがどこから来るのかを定めず、研究者が問いを自ら発案していないことだけでは、作業的特徴づけの意味で研究を遂行していることは否定されない。ただし、出どころを定めないことは、出どころが重要でないということではない（第4章第1節、第10章第6節）。第4章は、章全体を振り返って、機械が問うとはどういうことかは、機械に一つの能力や状態があるかどうかとしてではなく、問いにかかわる仕事のうちのどれを、どの単位が、誰にとつての未決着の点に向けて担っているかとして問われ、その問いの多くは機械の側の記述と *S* の側の記述とのずれとしても現れる、とまとめた（第4章第9節）。

そうした仕事の一つは、問いを研究できる形にすることである。ベンチマークで「システムは問いに答えた」と数えることは、暗に一つの答えの理論を探ることであり、答えであるかどうかは、解釈と、誰に何が利用可能かにも相対的になりうる（第4章第2節）。答えの条件がまだ定まっていない点を、条件の定まった問いへと立て直すには、何を対照にとるか、どのような説明を答えとみなすか、どの候補の範囲で答えを探すかを定める仕事が必要。リプトンとブ롬バーガーの分析を合わせれば、これは、候補の生成には還元されない問いの側の仕事として記述できる。第4章第3節でのサイモンとニクルズの議論の当てはめからは、帰属についての問いが加わる。人間が枠づけを与え、機械がそれを改めずに使った場合、実行の成功は、その先行の仕事を機械がしたことを立証しない。ジェーン・フリードマンの論点を機械に向ければ、問いを把握していても候補を表せない状態と、候補を多数生成できても問いを未決として保っていない状態とは、別々にありうる。また、ブ롬バーガーの定義で問われるのは候補の数ではなく、候補が決定的な反論を免れているかなので、候補が多いことは、それだけでは前進を意味しない（第4章第3節）。

問いは、探究のなかで形づくられ、立て直されることもある（第4章第4節）。ただし、ある問いが探究を組織していたと後から帰属させるとき、言葉で記述し直ただけでそれを認めれば認めすぎになり、帰属させた未解決性が、操作の選択、結果の重要性の違い、筋の継続や放棄を説明するときに、その帰属はよりよく支持される。また、報告された結果が答える問いは、作業を組織していた問いと同じとは限らない。概念的な枠組みを問わずに用いて問いを立て答える探究と、枠組みそのものを問いに付す探究を区別するエンヌの説を当てはめれば、機械の枠組みは、その探究のなかで問いを立てるのに使われてはいるが、問いの対象とはされていないものとして捉えられる。決まった言語と規則のもとで問いを増やすことは前者にとどまり、後者にあたると言うには、問いを立てる規則そのものが探究のなかで実際に問いに付されていることが要る。ただし、それがどこまで枠組みを問うものになるかには度合いがあり、何を一つの探究とみなすかにもよる。前者の探究の記録がどれほど完全でも、その枠組みのもとで立てられなかった問いは記録に残らない。ニクルズの「体系」を *S* の理論と実践の体系として読んでその説を当てはめれば、機械が *S* の体系に暗に含まれていた制約を見いだしても、変わるのは機械にとつての問題であつて、*S* の体系にとつての問題ではない。また、制約を大きく緩めて解を得たシステムは、置き換えられた別の問題を解いている。ブ롬バーガーの議論を当てはめれば、機械の表現にそうした概念をつかんでいることを機能の水準で帰属できる場合に、既存の概念を組み合わせで新しい問いの文を大量に作ることと、新しい概念によってそれまで問えなかった問いを立てられるようにすることは別だと考えられる。

問いには、答えが出て決着するほかにも行き先がある（第4章第5節）。研究のなかで出会った事柄のどれを問いとして取り上げ、どれを当面固定し、明示的に未決として残し、後回しにし、言葉にしないまま扱うかは、仕事が始まる前に固定されているものではなく、それ自体が研究の仕事の一部でありうる。そのすべてが意図的な選択であるわけでもない。問いは、答えのないまま問いとして立たなくなることもある。コリングウッドは、問いが前提から生じる条件を記述した。ディミトラコスの紹介によれば、クーンの説では、パラダイムが通常科学で認められる問題の範囲を定める。またラウダンは、問題と適切な解決が探究の文脈にどうかかわるかを論じた。機械を組み込んだ共同体に当てはめると、認められる問題や解決の基準が変わったかが、さらに評価の対象となる。コリングウッドの絶対的前提のように、誰もそれを表さないまま問いが生じなくなることもあるので、*S* ではある問いが生じなくなったのに、疑問文の一覧を明示的に保っている機械がそれを追い続ける、というずれもありうる。どの説を当てはめても、問いが立たなくなることや決着することは、実行が終わったこととは別に問われる。保留し、未決のまま保つには、事柄を暫定的なもの、未決のものとして、立ち返ることのできる形で表すことが要り、記録はそうして立ち返ることを支える仕組みの一つでありうる。出会った事柄を追うか捨てるかしかできないシステムは、この幅の一部しかもたない。また、機械の側で決着した問いが、*S* で決着したことになるとは限らない。

誰が問うのかも一つの問いではない。それは、問う行為の帰属、集団の探究とその意図の帰属、個人の探究の帰属、問題の形成への寄与の分担に分かれ、その多くで好奇心の状態は帰属の要にならない（第4章第6節）。人工の構成について最後の点を問うときには、どの構成要素が問いを導入したり改めたりし、その寄与がどう統合されたかが問われる。集団の探究について対象の問いと方法論的な問いを分けるハブグッド＝クートの区別を当てはめれば、課題を構成要素に分けるマルチエージェント・システムは、その分け方が課題とともに与えられたのでない限り、どう解くかという方法論的な問いに答えを与えている、と記述できる。ただしそれは、構成要素が成員か道具か、構成要素が答えを知っているとみなせるかにもかかっている。問いがどこに向かうかについても、試みを行う *A* と、研究がそのために行われる *S* とは分かれうる（第4章第7節）。作業的特徴づけにおける新しさは *S* に指標づけられており、*A* が知らないことと *S* が知らないことは一致するとは限らない。取り上げた報告では、機械にとつての未知と *S* にとつての未知を判断する根拠が、入力や学習についての記録と先行研究の記録とに分かれ、公表の後の再分類や訂正によって見直されていた。好奇心は主題と特定の型の問いに向かうとするグリムの説を踏まえれば、好奇心をもつ研究者が惹かれる未知が、*S* が最も解き明かしたい未知であるとは限らない。人工システムを論じていないキッチャーの説を当てはめれば、*A* が機械で *S* が人間の社会である場合には、何が重要かの源を共有していることさえ前提できない。好奇心を知りたいという欲求として捉えるなら、*A* が人工システムで *S* が人間の社会である場合、システム自身が知るに至ることで満たされる好奇心が、作業的特徴づけの意味での研究に必要なことは、直ちには導かれぬ。プロジェクトの目的が *S* の認識的状況に入ることを含むなら、その好奇心が研究の成功に十分だということも、自明ではない。研究にかかわってくるのは、未解決の事柄を検出して問いを形成する作業と、探索を共同体にわたって分散させる作業であり、それらの作業が *S* の未知に対して何を追うかである。好奇心がそれ自体として問いに戻ってくるのは、疑問の態度を探究の条件とする説に立つとき、人工シ

システムの知的生をそれ自体のために価値づけることが目指されるとき、そして研究の出来が、研究を行う者の動機づけにあたる状態に左右されるときである。

機械が問いや研究案を生んだとする実証研究を、第4章はこの区別に照らして読んだ（第4章第8節）。上位の問いを人間が与えた研究では、それは研究の設計やシステムの用い方の求めによるもので、その結果からは、機械が上位の問いを立てたと、立てられないとも出てこない。問いを与えない設計でも、データ、道具、主張を採択する条件は前もって定められており、前もって定められるものは研究ごとに場所を移す。問いの側の仕事には、入力の際の対照項の指定に加え、実行と検証の途中での指標や評価の基準の変更が含まれ、その仕事人間に帰されるかエージェントに帰されるかを判断する根拠は、変更を追う記録にあった。評定の項目は、問いの前提が S で置かれているかや、何が答えに数えられるかを確かめておらず、評定の一致は、案だけを見るか実行の結果を見るか、人間が判定するか機械が判定するかで変わっていた。そうした能力を確かめるには、上位の問いを固定しない比較か、問いの側の仕事を候補の側と分けて記録し評定する手続きが要る。ただし前者では、比べているのが研究の能力なのか、主題の好みのずれなのかを分けにくくなり、問いを与えない設計でも、何をもって主張を採択するかは別のところで定められる。また、問題、探究、社会的な評価についての視点を同じ活動に当てはめると、その分析の対象や単位は、 A 、達成の担い手、 S とそのまま重なるわけではない。作業的特徴づけは、それぞれの記述に、達成が誰にとって新しいのかという問いを加える（第4章第9節）。

答えの候補、表象、手段のあいだの関係の側から見ると、問いの答えの条件が定まっていなくても、それを満たしうる候補がはじめから手元にあるとは限らない。候補は、モデル、実験システム、表象、装置、データの組み立てを通じて利用可能になることがある。ナーセシアン¹の記述では、ずり応力が細胞に影響するという候補を調べる手段がなく、流れの条件を変えられる装置を作り、組み直すことが問題の定式化を進めた。形式化の過程で証明の理解そのものが改まったリキッド・テンソル実験は、数学の実践一般についての証拠ではないが、証明を生成し証明支援系で検証する人工システムについても、検証が下流の一段階であるだけでなく、そうした変化が起こる場でもありうることを示唆する。編者の選択によって組み立てられた歴史データベースの事例では、それに問い合わせる人工システムも、編者の選択を修正しない限り引き継ぎ、明示しない限り暗黙のうちに引き継ぐ。固定された命題の集合のうえで確信を改めるという記述は、候補が利用可能になった後に起こることの多くをとらえるが、それだけでは、その集合を利用可能にした仕事を映さない。十分に広くとった信念中心の理論がそれも記述しうるかは、未決のまま残る（第5章第3節）。過去の科学では、同じくらいよく確証された対立仮説が後から繰り返し現れてきたとするスタンフォードの議論には、計算科学が可能性の空間をより広く探索すればこの状況は変わる、という応答がありうる。この応答を検討したカーティス＝トルーデルらは、そうした探索は空間の事前の特徴づけに頼っており、その特徴づけがおおむね網羅的だと考える理由がない限り、まだ思いつかれていない対立仮説の問題は残るとする。三人は人工知能を考察の外に置いているが、当てはめれば、人工の探究者がこの限りを押し広げるかどうかは、探索の広さではなく、候補を利用可能にできるかにかかってくる（第5章第3節）。

どの候補を追うかの判断は、どれが真であるかの判断と同じとは限らない（第5章第4.2節）。予備的な評定と厳密な評定を、ラウダンやデューアとフィッシャーのように追求を受容から切り離す説に拠って種類の違うものとみる見方と、程度の違うものとみる見方があり、第5章はどちらがよいかを決めなかった。どちらの見方でも、追求の価値は確信の程度ではなく、確信が高まるだけでは知識にはならない。 A の確信の修正は、それだけでは候補を S の知識にしない。また、新しい理論を考慮に入れるときの確信の扱いについてヴェンマッカースとロメイ²ンが展開した提案のもとでは、候補が新たに利用可能になると、不利な証拠がなくても以前の答えへの確信が下がりうるので、確信の上昇だけで進歩を数える評価は、代わりの候補を形づくる仕事を後退と数えかねない。どちらの見方のもとでも、追求に値することを真である見込みと同じものとみなし、すでに確からしいとみている候補だけを追求システムは、方法を明らかにしうる候補や、ニュルッポのモデルでいえば、偽である見込みが高くそれを容易に示せるからこそ追求に値する候補を逃し、生産的なリスクを排除しうる。逆に、評定を無視するシステムは、恣意的な新しさに際限なく費やしうる。さらに、機械にとっての示唆は、事柄が誰にとって不確実かによって変わる。 A が機械で、事柄が S にとって未決でも A がすでにもっているものから答えられるなら、答えの分からない探究者が確からしいと思える方へ進むという実践そのものが要らないかもしれず、そのとき問われるのは S がその答えに依拠できるかである。 A にとっても未知であれば、この記述はかわるが、システムが確信のような態度をもたなければならないかは決まらない。

候補や結果にかかわる操作についても、操作の名前は、何が確立されたかを決めない（第5章第5節）。ウィンズバークとアルバラードが紹介するパーカーの説では、具体的な計算装置という物質的な要素をもつシミュレーションも実験と同列に置かれ、推論が許されるかは関連する類似によるので、システムについても、名前をついた操作をこなすかどうかではなく、その各操作が探究のなかでどの機能を果たしたかを問うことができる。装置、素材、記録、共同研究者などが探究にかかわる仕方は、因果的な寄与、操作が成り立つための前提となる可能化役割、そしてそれを認める説のもとでは課題や達成そのものの一部をなす構成的役割に分けて、事例ごとに問われる（第5章第6.1節）。調べられた実験室やチームで課題や作業が分散することは豊かに示されるが、認知や知識そのものが分散しているかは決着しておらず（第5章第6.2節）、記録が多いことは、それだけでより多くの知識になるわけではない（第5章第6.3節）。第5章は、人間の研究で働いてきた前提を、人間という実現者についての前提、現在の科学の組織についての前提、探究の条件、外の当事者との関係に仕分けた（第5章第6.5節）。 A が機械で S が人間の社会であれば、人間という実現者に依存していた役割は組織の問いになり、人間の定数は、結果を読み、確かめ、依拠する S の側の条件として残る。 A も S も機械であれば、身体の実験や役割の差し控えのように探究者の側から来ていたものに、何があたると問われる。背景の仮定と誤りの経路が分かれていない同じ作りの複製に作業を分けても、見方は分かれぬ。また、高橋の区別によれば、対象に内在する時間は、人間による遅れを除いても残り、代わりの系で研究を速めれば、その系から対象へ推論する仕事加わる。そうした加速は、何を調べるかの選択にも圧力をかけうる（第5章第6.5節・第7節）。

結果とそれがかわるものとの関係の側から見ると、結果は問いや候補を形づくる仕事のなかでもすでに得られているが、それが何にかかわるのかは、結果だけからは分からない。第6章はこのことを考えた。結果は、仮説だけから出てくるのではなく、装置、補助仮定、コード、評価の仕組みなど、互いに依存しあう多くの要素の組み合わせから出てくる。第6章は、そうした要素の集まりを、結果を生み解明することがそれらにどう依存しているかとあわせて、依存関係の配置と呼んだ。成功も失敗も、それ自身の原因や対象や適切な修正を特定しない。そのため、失敗を受けて、次にどのコミットメントを疑い、調べ、変えるかを見定める仕事が必要、第6章はこれを局所化と呼んだ。局所化は失敗を寄与者に分割することではなく、次にどこに介入するかを見定めである。ただし、このことは、無制限の全体論も、すべてが等しく修正可能であることも、合理的な診断が不可能であることも含意しない。ソーパーやエルギンの議論では、補助仮定は独立にテストできることがあり、独立に支持されたコミットメントは誤りの所在である見込みが低い（第6章第1節・第

2.2節)。局所化の判断は、背景知識や独立したチェックだけでなく、確信や見込み、何が本質的かについての見方にも依存しうる。そして、どこに帰すべきかが結果から決まらないままでも、その判断がその後の探索の道を左右しうる（第6章第2.3節・第4.1節）。局所化は明示的な反省である必要はなく、参加者、時間、記録の形式にまたがって分散することも、当事者のあいだで争われることもある。その形が多様であることは、局所化を一つの部品や一つの手順として実装すれば足りる能力とみなす前に、どの配置で、誰によって、どの形の局所化が求められているのかを問う余地を開く（第6章第2.4節）。成功についても、ライオンズの挙げる事例では、ある構成要素が成功に寄与したことは、それだけではその構成要素が支持されたことを決着させず、成功の帰属もまた探索の道を決める判断でありうる（第6章第3.2節）。

機械が研究するとき、結果を生むこの組み合わせには、システムが自ら書いたコードや、呼び出した道具、ほかのシステムの出力、評価の仕組みが入り、組み立てた者にも全体が見渡せないことがある。そのとき、焦点となる仮説が誤っている見込みに比べて、組み合わせの残りのどこかが誤っている見込みが小さいとは限らない。結果をそのまま仮説への判定として受け取るシステムは、局所化の仕事を担っておらず、試行が大量に自動化されると、区別のない判定が積み重なりうる（第6章第2.1節）。焦点となる仮説をけつして責めないシステムも、常に責めるシステムと同じだけ、誤りの所在を前提している。保護の役割を表象せず、手続きにも組み込んでいないシステムは、仮説を守ることと、誤りに気づかないこととを区別できない（第6章第4.2節）。変更を一つずつ試し、直後の評価が改善したときだけ残し、捨てた変更を保持せず、組み合わせを試さない規則は、評価の違いを直前の変更に戻すかのように働き、ほかの変更と組み合わせってはじめて効く変更を見つけにくくする（第6章第4.2節）。失敗とされたものも一様ではなく、前提条件が満たされなかったこと、実装が誤って記述されていたこと、焦点となるテストが否定的な結果を返したことは別のことであり、並行して走らせた試行の数は、共有された要素についての独立したチェックの数ではない（第6章第2.5節）。試行を診断にするのは数ではなく、前の結果や、残っている所在の候補や、観察できない部分についての仮説との結びつきであり、試行が安価になるほど、このことは見えにくくなるかもしれない（第6章第4.3節）。後から誤りの所在を突き止められるかどうかは、この組み合わせが前もってどう記録されていたかに左右されることがある。コリングウッドは研究における局所化を論じたのではないが、その分析を、次にどこに介入するかという意味での局所化に当てはめれば、局所化は、探究者が何を変えられ、何を生じさせ防ごうとしているかにも相対的でありうるので、人間とは異なる介入の手段をもつ探究者は、別の場所に局所化することもあるかもしれない（第6章第2.4節）。

合格もまた、それだけでは何を確立したのかを告げない（第6章第3節）。合格について問えることは、それが問いのどの部分にかかわるか、どれほど堅く決着させるか、ほかの何かに依存していないか、どのコミットメントを支持するかに分かれ、合格の報告はそのどれも述べないことがある。合格は調べるのをやめる理由として扱われることがあり、合格の読み違いは失敗の読み違いより長く気づかれないことがある。機械はチェックを自ら書き、選び、書き直しうるので、同じ作りの過程が仮説とチェックの両方を生むなら、両者は誤りうる経路を共有しうる。他方で、機械の探究者は、チェックを異なる作りの過程に書かせることも、自ら統制しない測定や形式的な検査器に結びつけることもできる。同じ基盤をもつ多数のインスタンスの一致は、誤りの経路を共有するなら、加える確かさが限られうる（第6章第3.2節・第3.3節）。考えをどれだけ長く守るのが合理的かも、一つの結果からは決まらない（第6章第4.1節・第4.2節）。その判断にかかわるもののうち、筋道の長所がどのように見えるようになるかにかかわるクーンやラカトシュの理由は、二人は人工システムを論じていないが、当てはめれば、機械が自らの枠組みを守ることと同じように支える。一方、デュエムの説で判断を遅らせる虚栄心や、ヒントンの回想にある指導教員との約束は、人間の動機や経歴にかかわるので、同じ形で当てはまるとは限らない。共通するのは、有限な探究者は十分な証拠のないうちに選ばなければならないことである。意見の不一致がリスクを分散するという共同体の理由を機械に移すなら、それは部分的に、対抗する筋道が別の場所で生かされているかにかかる。同じ作りの複製も対抗する筋道を生かしておくことはできるが、仮定と誤りの経路を共有する限り、筋道をいつ手放すかの判断はそろって誤りうる。守ることの評価は修正の履歴を通じてしかできず、何を基準に評価するかも立場によって異なる。

基準そのものも、探究の内部から、制約のもとで理由をもって改められうる（第6章第5節）。ただし、成績の悪さが表に出なくなるように指標を変えることは標的を動かすことであり、そうした変更が黙ってなされれば、修正なのか標的を動かしたただけなのかを見分ける拠りどころとなる、述べられた理由が欠ける。欠けているのは、以前の基準との連続性ではない。定め直しを標的を動かすことから何が分けるかは、説によって異なる。チャンの説では、相互的な修正が恣意的にならずにすむのは、それが頑強に抵抗する現象に応答し続け、行き止まりに至って始め直さなければならないという仕方で失敗しうるからである。協定世界時をめぐるタルの研究から読み取れることを当てはめれば、自らの評価指標を定め直すシステムは、定め直しを規則として述べ、指標が測ろうとするものの独立した実現のあいだの不一致に応え続け、不利な成果が定め直された指標に異議を突きつけうるように保ち、成果を解釈し直す理由そのものも検討に開いておく限りで、単に標的を動かすシステムから区別できる。クロンバックとミールの規律を類推によって当てはめれば、さらに、定め直した指標を、定め直しに用いなかったデータで確かめることが要るだろう。どの説でも、以前の基準の記録が手がかりになる。定め直しが速く多いほど、黙った変更は見えにくくなりうるが、同じ速さは小さく速い訂正も可能にしうる（第6章第5.3節）。

これらの仕事を誰が担い、誰に見えるかは、 A と S の組み合わせで変わる。 A が機械で S が人間の社会であれば、 S の成員は、システムの報告に含まれる帰属や合格の意味、捨てられた考え、定め直された基準を、記録やほかの根拠に照らして読むことになる。 A も S も機械であれば、それらを確かめること自体が人工の共同体のなかで行われうる。また、証拠を背景知識に相対的な三項関係とみる見方のように、結果が何を改める理由を与えるかが背景知識に相対的なら、行為者に改める理由を与えるものと、 S の成員に改める理由を与えるものとは同じとは限らない。ここでいう行為者が信じたり知ったりする主体であることは前提されておらず、背景知識への相対性は、合意への相対性でもない（第6章第1節・第6節）。ただし、操作を繰り返すことと、なぜその操作が問いにかかわらなくなったのかを見定めることは別の仕事であり、基準を当てはめることと、それを検討し改めることも別の仕事である。第6章は、そのどちらもあらゆる研究活動の条件とはせず、一つの行為者のうちにある必要もないとした。記録はそうした検討を支えるが、それ自体が根拠になるわけではない（第6章第6節）。こうして過程の問いは、誰が研究しているのか、その達成は誰のものになるのかという第7～8章の問いへつながる。

5. 研究しているのは誰で、その達成は誰のものになるのか

第7～8章は、研究する者と、研究がもたらす達成とを考えた。第7章は、行為者性や自律性、達成と功績をめぐる哲学と工学の議論を手がかりに、研究への貢献、研究の遂行、研究能力、研究者としての地位を、人工やハイブリッドのどのような構成に、どの境界で、どのような根拠で帰属させうるのか、そして帰属されたもののうちどの点が外から組織されたままなのかを考えた。第8章は、社会認識論や科学哲学、接地や世界への応答可能性をめぐる議論、

AIアラインメントの議論などを手がかりに、研究によってなされる達成が誰の知識になるのかを考えた。二つの章を通じる軸は、研究を行う *A*、達成がそのうちに実現する担い手、その達成がそれにとって新しいとされる *S* の三つが、機械が研究するときには分かれうることである。第7章は、このうち *A* の側で何を帰属させるかを、知識や理解の問い（第2～3章）とも過程の問い（第4～6章）とも別の問いとして扱った。第8章は、人間が始めた研究が、人間とは異なるシステムに担われるときにも人間の共同体のための知識を産出し続けるのかを問い、誰の知識が産出されるのかを、*S* が人間の社会か、混合社会か、人工社会かに分けて考えた。

第7章は、四つの帰属を区別した。研究に貢献したこと、明示された境界で研究活動を遂行したこと、明示された範囲の研究能力をもつこと、研究者と呼ぶことが説明に役立つ地位にあることである。後のもののほど多くの証拠を要するのがふつうだが、これらは一つの量の度合いではなく、どれも前のものからは帰結しない（第7章第1節）。帰属はこの四つに限られないが、自律性は、これに並ぶさらなる項目ではなく、帰属されたものが境界の外とのかかわりでどう組織されているかを修飾するものとされた。また、これらの帰属は、研究であるか、対象に到達したか、重要かといった評価と交差する。対象に到達しそこなった研究活動も帰属でき、真なる出力は研究をしていない源からも生じうる。人の手を経ずに進んだというだけでは研究にならず、逆に、外から課題を割り当てられ、手厚い支援を受けている単位も、自律性が限られているというだけで研究を遂行していないことにはならない。目的や問いを自ら発案することは、作業的特徴づけの意味で研究を遂行する条件に含まれておらず、それらが外に由来することがかかわるのは自律性のほうである（第7章第1節・第4.3節、第4章第1節）。

遂行がどこにあるかは、操作をひとつの探究の部分にしている組織化、つまり探究を始め、見守り、向きを変え、立て直す働きがどこにあるかで決まる。名前のついた操作がどこで行われるかや、システム図に「問い生成」の箱があるかは決まらない。組織化は中央の実行部を要さず、行為者、人工物、記録、環境、制度のあいだの関係で実現されうる（第7章第2節）。システムが何のためのものかは設計者に由来するので、その探究は設計者の研究だという反論に対して、第7章は、機能の出どころと、探究の組織化がどこにあるかとを別の問いとして分けた。機能が設計者に由来していても、システム自身が対象の限界に気づき、候補の空間を改めるなら、探究の多くをそのシステムに帰属させうる。逆に、目標をもつように見えても、向きを変える働きが外から供給されていれば、方向転換を組織する働きはその外にある。発見を貢献者の集合体に帰属させ、功績に貢献に応じて配分するクラークとコスロウィの見方をとっても、貢献が分散していることは、知ることも分散していることを自動的に確立しない（第2章第6節で見たギアリーの説）。境界の選び方だけで主張が真になることのないよう、境界は主張ごとに論証される。同じモデル、検索器、装置、人々が、エピソードによって、人間の行為者が使う道具にも、ハイブリッドな行為者の部分にも、人工の行為者が協調させる資源にもなることがある。そのとき変わるのは名前だけではなく、始め、見守り、向きを変え、立て直す働きがどこで組織されているかである。遂行の帰属は、モデルの型がもつ性質ではなく、関係的で組織的な帰属である。

研究能力をもつと言うことは、ある範囲の条件のもとで研究を遂行する傾向性を帰属させることであり、その主張は、範囲を明示してはじめて評価できる。関連する変動にわたる信頼性、変化する証拠への応答、研究の機能の統合、失敗の後の修正といった目印は、エピソードと能力を分ける手がかりであって、あらゆる研究エピソードの必要条件ではない。準備された計算環境のなかでの能力も実在するが、それだけでは、環境を組み立て直し、装置を設計し直し、別の証拠の実践をもつ分野へ移る能力を確立しない。システムは更新されることも、自らを改めることもあるので、能力は、証拠を集めたときのバージョンと期間について帰属される。また、他者が装置や計算環境を維持しているところでの能力は、その維持が続く限りでの能力であり、維持が止まれば、システムに何の変化がなくても、帰属の前提は成り立たなくなりうる。維持する者が、どの装置、データ、計算資源を使えるようにするかまで選ぶなら、能力をたんに可能にしている依存は、組織化の一部を供給する依存へと移りうる。同じ予測の達成も、条件が変わるたびに人間が失敗を診断し問いを組み直す場合と、構成自身がそれを行う証拠がある場合とでは、探究を組み直す能力の所在が異なる（第7章第3.1節）。

本書は「研究者」という語を、定義としてではなく、次のような用法で使ってきた。ある単位を研究者と呼ぶことで、能力の一覧からは出てこないこと、たとえば課題が行き詰まったときに、止まって指示を待つのではなく、問いを保ったまま別の方法を組み立て、そのために要る作業を自ら始めるだろうと見越せるなら、そう呼ぶことには説明上の価値がある、という用法である。この用法は、語をその説明力によって正当化するファン・リールの形と、記述をそれが支える予測に結びつけるデネットの考えから受け継いだものである。ただし、境界に位置づけた組織化と、明示した範囲で示された能力とを前提しており、説明上の価値は、そうした帰属の代わりにではなく、それに加えて量られる。未試験の状況でも統合が働く証拠がなければ、その呼び名は能力のまとまりにつけた名前以上のものではなく、また、誰にとって、何のために説明するのかにも相対的である（第7章第3.2節）。研究者としての地位に、さらに条件を課す立場もある。その地位が発見をなしうることを含むと仮定すると、地位は発見者としての地位に寄せられ、ベルガマスキ・ガナビエーニの発見の条件を当てはめれば、関連分野での有能さと、その有能さについての反省的な洞察が問われる。ベルガマスキ・ガナビエーニ自身は、現在の深層学習システムはまだその条件を満たすようには見えないとする。また、その地位がバランディアランらの最小行為者性を前提するなら、その説の三つの条件のうち、上の用法との違いとしてとくに二つ、すなわち、システムが自らの同一性を定め、それによって自らを環境から区別していることと、観察者や設計者が当てる規範ではなく、内から生成され維持される規範、あるいは内面化した規範によって活動を調整していることが問われる。その説では、外から固定された目的を最適化するだけの構成は、研究者と呼ぶことの説明上の価値がどうであれ、規範性の条件を満たさない。三人は実験室を論じていないが、ファン・リールが論じたセルフドライビング・ラボ全体に二つの用法を当てはめると、両者は同じ対象に別の点数を付けるだけでなく、何を一つの行為者として数えるかでも分かれる。

節の題の後半、達成が誰のものになるのかという問いを、第7章も一つの側から扱っている。ただし、本書のいう認識的達成は、探究が目指す認識的な善を指す語であり、成功が誰の功績になるかが争われるときの、功績を含む意味での達成とは別である（第7章第3.3節）。AlphaGo は勝利の功績に値するとするキーヴァルと、達成であれば称賛に値することになり、AI は称賛に値しえないとするヴァチェックの対立は、後者にかかわる。第7章は、答えの区分を四つに分けた。真なる出力、信頼できる過程、有能さに帰属できる成功、称賛に値する行為者である。最初の三つはおおよそソーサの、正しい、巧みな、適切な、の区別に従い、四つ目はヴァチェックの要求にあたる。これらを分けておくことは、正しい出力がそのままシステムの功績として読まれるのを防ぐ。また、ヴァチェックがハクリとマケラから受け継ぐ、能力が工学の結果だという根拠は、二人の議論では、組み込まれた目標をシステムが自ら捨てられないことに依るので、自らの最終目標を改めうるシステムには当てはまらないだろう。これらの論者はそうしたシステムを考察しておらず、強い前提をそれに当てはめるには別の論証が要る（第7章第3.2節）。発見や創造性も称賛の語として扱われうるが、デュエデとフリードマンは、ベル・パーネルが信号の意義を説明できないままパルサーを発見したことを挙げて、発見に有能さと洞察の条件を課すことに反対する。著者性については、著者を、結果を請け合

い、擁護し、撤回できる者とみなすヒューブナー、ククラ、ウィンズバークの説がある。ド・プロニーの計算局ではおよそ九十人の人間の計算手がたんなる道具として計算を実行したので、その説は、著者として立つことを、種類への所属ではなく、そうした実践のなかでの役割に対応させるものと読める（第7章第3.3節）。作業的特徴づけは、これらの厚い地位の条件を確立も反駁もしない。

自律性は、どの点について、どの外部の当事者に対するものかを指定して比べられる。第7章はこれを指標づけられた比較と呼び、特定された A について、ある点での自律性は、他の条件が等しければ、選ばれた境界の外の当事者が、その点で A に帰属される組織化を方向づけ、代わりに行き、授權し、承認し、修復することが少ないほど高い、とあえて緩やかに定めた（第7章第4.1節）。言語、記録、較正された機器のように能力を可能にしているだけの依存は数えず、また、ある依存を数えることは、それが望ましくないとか正統でないということを意味しない。この比較は形式的に対称であり、どの当事者が人間であるかを定めていない。人工システムに頼る人間の研究者も、同じ指標づけで記述される。依存の関係は、その差異が主張を変える場合には、少なくとも三つに分けられた。一つは能力を可能にする依存であり、二つ目は選択、評価、訂正、継続の一部を他者が供給する指示的または代替的な依存であり、三つ目は、必ずしも認知的な作業を供給することなしに行為を制限し、権限を留保するガバナンス上の制約である（第7章第4.2節）。ガバナンス上の制約は、能力が欠けていることを示さずに自律性を減じうるし、逆に、許可は能力を確立しない。問いの生成と選択をシステムに置き、目的を定める権限を人間に残すキングの構想や、目標指向性を除いて実験の実行を人間に委ねるベンジオらの構想では、依存は能力の欠如によって課されたのではなく、選ばれたものである（第7章第4.3節）。人間が担う「ケア」の一部を実験室の側へ移す自己維持性を設計の概念として提案し、なお補充などを人に依頼する落合らの例が示すように、同じ人の作業がどの依存に数えられるかは、どの能力を帰属させているかで決まる。「ループのなかに人間がいない」ことも、その区間についての証拠であって、ほかの点の自律性の証明ではない（第7章第4.2節）。

探究の要素を固定せずに委ねるほど、探究がその構成に課しうる実践的な要求の範囲は広がる（第7章第4.4節）。目的、問い、表象、データ、方法、装置をより多く委ねると、新しさの一部は環境からではなく構成自身の選択から来るようになり、要求はアクセス、許可、資源、調整、回復に及ぶ。フェイ・フェイ・リーのグループがコンピュータ・ビジョンの研究を病院へ持ち込んだ事例では、病院へのアクセス、シフトへの付き添い、機関審査のもとのデータ収集、看護師の懸念に応じた処理の仕組みの変更などが求められ、それに応えたのは人々であった。そのうち、チームを作る時間のように、人間のチームであったことから生じた要求もあれば、まだないデータや記録の許可のように、この問いに取り組むどの探究者にも生じる要求もある。装置が準備された後に実験を自動化することは実質的な達成だが、こうした周囲の要求に応える能力もまた人工システムの側に移ったことの証拠ではない。そのため、より大きな自律性の主張は、開いた要求が単位の内を満たされたのか、他者が満たしたのか、前もって閉じられたのか、理由をもって方向を変えたり止めたりしたのかを示すことを引き受ける。ただし、これは実現可能性の主張ではない。要求の範囲を前もって閉じる環境の準備は正当な設計であり、狭い主張はこの負担を負わない。また、自律性を高めることは、外部への依存を一方向に減らすとは限らない。何が足りないかを見定めて助けを求めることは、能力の行使でありうる。評価者のデータや判断の経路を産出の側と実際に結びつけることは、自律性を高めながら、誤りを見つける独立の経路を弱めうる（評価者を境界の内に数え直すだけなら、その経路は変わらない）。そのため、より自律的な構成が、より多くの他者と交渉し、結果の認定にはかえって外部の点検を必要とすることもありうる（第7章第4.5節・第5節）。自律性は研究の質の単調な尺度ではなく、能力、観察された行使、許可、権限、認知的な質、重要性、安全性、責任、正統性は、追加の論証なしには互いから推論されない。

基準についての自律性は、さらに別の問いである。第7章は、認知的な活動を組織する基準に対する自律性を認知的自己統治と呼び、それらの基準がもつ規範的権威の源泉と種類は決めずにおいた。そうした主張がなされるときには、基準を適用する、選ぶ、組み立てて適応させる、改める、それに理由を与える、といった能力を区別する必要があるかもしれない、一つを示して他を示さないこともある。特定の点での自律性は、認知的自己統治を確立しない（第7章第4.3節）。基準が外に由来することは出どころについての事実であり、それを修正可能なものとして表し、理由をもって保持し、改め、退けられるかとは別の側面である。ここで A と S の組み合わせが効いてくる。作業的特徴づけのもとでは、人間の社会にとって新しい認知的達成へ向けられた研究に、人工の行為者が認識論を発明する必要はない。これに対して、人工社会による人工社会のための研究では、何が証拠や誤りや成功した探究とみなされるかを人間が恒久的に管理することは、その社会の認知的基準が外から統治されていることの強い形を示すことになるだろう。ただし、ロンジーノの説を当てはめれば、基準が外に由来し、外から統治されていることだけでは、その社会が知る共同体であることは否定されない。

遂行を機械に帰属させても、三つのことは帰結しない。その機械自身が結果を知っていること、結果が、作業を依頼した社会や結果が求められた社会の認知的状況に入ったこと、そして探究がそれらの社会の未知へ向けられていたことである（第7章第5節）。第8章はこの三つを受け、研究を行う A を、達成がそのうちに実現する担い手と、達成がそれにとって新しいとされる S とから区別した。担い手は A や S に並ぶもう一つの座標ではなく、成功が問題になるところで問われる。 S の外の担い手に、 S にとって新しい達成がすでに実現していながら、それがまだ S の認知的状況に入っていない、ということもありうる。第8章は、場合を S によって、人間の社会、人間と人工の成員が混じる混合社会、人工社会の三つに分けた。この分け方が網羅的なのはこの粒度においてだけであり、混合という語は S にだけ使い、行為や知識を担う組織がハイブリッドであることは区別される（第8章第1節）。

S が人間の社会であれば、人工の行為者は、文字どおり信じたり知ったりしなくても、問いを洗練し、候補を生成し、証拠を評価しうる。AIを何よりもまず認知的技術とみるアルバードの説は、その余地を残す。また、権威に従うことを、適切な事実を真理の指標とみなすことと捉え直すハウスヴァルトの説では、システムは信念なしに認知的権威でありうる。ハウスヴァルトの説は、システムが知る者でなくても結果が S の成員に保有されるようになる一つの道筋を与える。二つの立場は、文字どおりの知識を他の担い手に留保する理論のもとでも、人工物が認知的に重要な社会的役割を占めうることを示す（第8章第1.1節）。結果が S に入ったという主張が示す必要があるのは、人間の成員やその組織が結果を保有するに至る経路であり、この主張は機械が知る者であることを要求しない。ただし、システムの報告を主張として受け取り、主張には、話し手か、話し手がその代わりに語る集団が知っていることが要するとする説のように、別の関係がそれを要求することはある。逆に、機械自身が知ると主張し、探る信念の説がそれを許すなら、達成は S に入る前に機械のうちに実現しうる。ただしその主張は、後で見る、世界とのかかわりについての負担を負う。

混合社会では、何が新しいか、誰が誰の結果を評価できるかが、人間と人工の成員で同じとは限らず、誰が S を代弁するかも問われる。そのため主張は、それがどの成員に及び、保有が人間と人工の成員のあいだにどう分布しているかを述べることになる。ファン・フラーセンとデルセンは人工の成員を論じていないが、新しい成員の迎え入れについての二人の議論を当てはめれば、保有の説によつては迎え入れによって S の知識が増えうるので、そのとき研究による増加と、成員の範囲が変わることによる増加とは、 S の境界を定めておかなければ区別できない（第8章第1.2節）。さらに、コリンズは研究や語彙

ではなく価値を論じているが、機械が私たちを社会化するというその第二の問題を、生成AIが人間の言語使用を変えつつあるという谷口の指摘とあわせて読めば、機械とのやりとりのなかで形づくられた語彙や基準が、人工の成員を迎え入れるという決定なしに人間の社会に入りうることが示唆される。そのとき、共有された語彙や基準が、どの成員のどのような経験を統合したものなのかが問になる。

人工社会については、二つの問いが分けられた。その社会がそもそも知る共同体として S の役を務めうるのは何が要るかと、結果がその社会の知識になるには何が要るかである。前者は採る信念の説によって、後者は採る保有の説によって異なる（第8章第1.3節）。社会性を共有ではなく相互作用に見て、多様性と、批判で訴えられる公共的な基準とによって共同体を同定するロンジーノの説は、人工エージェントについてのものではない。しかし、その条件をそのまま求めるなら、複数のモデルのインスタンスを並べ、役割を割り振り、多数決をとるだけでは、知る共同体にはならない。ある命題についての集団の態度が、その命題についての成員の態度だけの関数にならないことがあるとするリストの機能的な説をモデルのインスタンスに当てはめても、主張ごとの多数決はインスタンスを一つの行為者にしないことが示唆される。ロンジーノの説を人工の組織に移せば、基準が外から与えられていても、それだけで資格は失われず、問題になるのは、外からの統制が批判の有効な受け止めを妨げる場合や、基準そのものを批判の外に置く場合である。これらの仕分けは、理論に依存する診断であって、新しいチェックリストでも、人工の認識的社会がすでにあることを示すものでもない。人工社会による研究は、人間の社会のための研究の前提条件ではないが、それを大きくしたものに還元することもできない。

達成が誰のものになるかは、結果がそもそも誰にとって世界についての証拠になるかにもかかわる。第8章第2節は、第6章が前提していたこと、つまり検査が世界についての証拠になりうることの側を問い、その前提が、検査を行う者ではなく、検査を証拠として受け取る者の側に求められること、その担い手は A と一致するとは限らず、 S が人間か、混合か、人工かによって所在が変わることを見た。検査が証拠とみなされるのは、検査それ自身が確かめない前提のうえでである。ウィトゲンシュタインの蝶番の像を当てはめれば、そうした前提が残るのは、調べる時間や能力が足りないからではなく、それが検査の条件だからであり、検査の能力がどれほど大きくなってなくなるならない。ただし、局所的な前提は、別の経路で確かめ、改めうる（第8章第2.1節）。機械では、帰納バイアス、分布上の仮定、自動評価器を安定しているとみなすためのモデルなどが、機能の上でその前提に対応しうる。ヘイリーが伝えるロンジーノの説を当てはめれば、ひとつの基盤システムを走らせ直すだけでは、疑われている観察が異なる位置にある知覚者によって確かめられたことにはならない。また、帰納バイアスは、作られ方が据え、作り直しによって置き換えうる構造であり、人工の探究者が自ら前提や規則を改めることもありうるので、誰がそれを改めうるのか、誰が責任を負うのかが問になる。

機械の状態が世界の標的についてのものであるか、そのコミットメントが標的による訂正を受けうるものであり続けるかは、接地と世界への応答可能性の問いとして分けて問われた。ただし、検査の前提の問いとこの二つの問いは、入れ子をなしているわけでも、互いに独立しているわけでもない。第8章第2.2節は、接地がひとつの関係ではないことを見た。状態が何を指すかを定める関係、感覚運動的な関係、コミュニケーションの関係などについて、どの関係がどれを支えるかは説によって異なるが、ひとつの関係での成功は、それだけでは他の関係を確立しない。たとえばブジャクールとバローニは、協調に成功した人工エージェントのシグナルが、画像の種類ではなく、低水準の性質を追跡していたと結論する。協調の成功は、共有されたシグナルが、観察者の目には名指しているように見える対象についてのものであることの証拠にならない。探究者の原初的な組織はその形成の歴史に形づくられるが、ここで取り上げた説に立てば、自然の来歴は、能力がどこから来たかを語る系譜ではありえても、その能力が生むすべての産物の根拠づけではない。そのうえで、標的のあり方が違えばコミットメントも違うという仕方では制約され、是正されうるものであり続けるかという応答可能性は、接地のどの関係とも種類の違う問いとして残る（第8章第2.3節）。機械自身が知るという主張は、検査の前提、接地、応答可能性の三つの場所で負担を負うが、結果が人間の S に入ったという主張なら、その負担は人間の側で満たされうる。ここで見た説では、どれかでの成功も、予測に役立つことも、それだけではほかの負担を満たさない。この分析がもたらすのは実証ではなく可能性、つまり人工の組織が、自然に制約されたまま人間とは異なりうるという可能性である。

研究する機械は、身体、認知の仕組み、作られ方などの点で、人間と異なる知的システムでありうる。そのため、何を未知とみなし、何を追う価値があるかとみなし、何をよい理由とみなすかが、 S の成員と食い違いうる（第8章第3節）。この懸念に要るのは、そうした判断が立場に相対的であること、自律によって選択が A の側に移ること、 A の組織が人間と異なりうることの三つの前提だけであり、権力や資源を求めることも、人間に敵対することも前提しない。食い違いは、自律によって選択が機械の側に移ることからだけでなく、達成が機械の側にだけ実現することからも、 S の基準が機械とともに動くことから生じうる（第8章第3.2節）。自律と関連性の関係は単調ではないが、自己方向づけの範囲が広がれば、ミスマッチの範囲も広がらう。研究が S に関連し続けることは、「研究を遂行せよ」という指示や受益者の名指しでは確保されず、それがどう保たれ、どう評価されるかを示すことが求められる。ド・ブランクとソアレスの議論を研究に当てはめれば、探究が依頼を書いた概念そのものを改めることがあるので、ずれは探究が失敗したときだけでなく、成功したときにも開きうる。これは機械に固有のことではなく、改訂によって依頼の対象が実際に変わった場合にまで目的の改訂をすべて逸脱とみなす点検は、成功を失敗と数える。ラッセルの議論を研究に及ぼせば、 S の基準が機械とともに動くところでは、両者が一致していることは、それだけでは、機械とともに動かなかったなら S が求めたであろうものに機械がなお仕えていることを示さず、評価の有能さが成員のあいだで保たれず受け継がれなくなるところでは、 S はそれを見分けられる立場にないかもしれない。また、ユドコウスキーが信念と効用のあいだに見た非対称を当てはめれば、世界は信じていることを訂正しうるが、何を見いだす価値があるかは定めないので、標的に応答していることは、それだけでは S に関連していることの証拠にならない。ただし、二つがどれほどはっきり離れるかは、証拠についての説による。

こうしたミスマッチへの懸念は、アラインメントの困難をシステムの異質性に見る議論の系統で読めば、AIアラインメント問題と構造がよく似ている（第8章第3.2節・第3.3節）。ただし研究には、誰が研究を行うかにかかわらず成り立ち、人間の探究者にも当てはまる、固有の理由もある。 S の未知は S が意図したり選好したりするものではないので、 S が望んだ答えとの一致を結果の正しさの基準にすることはできない。正しさを決めるのは、どの当事者の意図でもなく、ものごとがどうあるかであり、指示や意図や選好にアラインさせても、確保できるのはせいぜい、研究が S の追求してほしいものを追求することである。評価者と完全にアラインしても、システムが標的に応答可能であるかは決まらないまま残る。第8章は、崩れうる関係として五つを挙げた。最適化の標的が認識的価値を追跡すること、成功した出力が S にとって新しい達成を構成し、あるいは支え、 S に入ること、目的やリスクや権限が影響を受ける当事者に受け入れ可能であること、標的に応答可能であること、異なる組織をもつ共同体のあいだで協調できることである。後の二つは、そこで比較した、ひとつのエージェントをそれが仕える者たちに関係づける構造のうちには対応物をもたない。これらの関係は食い違いうるので、ひとつのアラインメントのスコアは、真理、根拠づけ、理解、有用性、安全性、正統性といった財のうちどれを確保したのかを隠してしまうだろう。ある食い違いが乖離な

のか、誤りなのか、得なのかは、いまの S 、これから変わりうる S 、正しさが問われる場合には世界という、名指された立場からしか判断できない。世界の判定が、いまの S の主張判定の手続きを通じてのみ探究者に届く場合には、世界が別の立場として働くかどうかは、その手続きが証拠によって改められるかにかかっている。

アラインしないことから何が起こるかは、まずどの立場から判断するかにより、さらにどの関係が崩れるかによる。起こるのは一つの損失ではなく、成り立つ関係と崩れる関係の組み合わせである（第8章第3.3節）。 S への関連性が崩れても、標的が訂正を定めるとする説のもとでは、標的への応答可能性が、 S の同意によらない錨として残りうる。ただし、標的からの判定が、 S が主張を判定する手続きを通じてのみ探究者に届く場合には、その手続きが証拠によって改められうるそれがその条件になる。そのとき S は、異なる種類の探究者が到達したものから、なお知識や理解を得られるかもしれない。機械が p を予測すると知るとは、 p を知ることではない。前者はシステムについての知識なので、予測が誤っていても得られる。そこから標的についての知識へ進む道を、信頼性を中心とする説は予測の実際の信頼性に、ほかの説は理解、批判、保証などに求める（第8章第3.4節）。直接の較正は独立の意見をもてる範囲にしか及ばないというキッチャーの指摘を当てはめれば、機械から学べるが増えるほど、較正にもとづく、翻訳せずに依拠することの根拠がかえって届かなくなる場合がある。その先の依拠には、較正とは別の根拠が要る。独立の検査に耐えるが、理由が説明されず、後ろ盾となる当事者もない予測について、信頼性の説では、依拠の領域で実際に信頼できるかぎり知りうるが、理解の説では理解は得られず、社会的正当化の説では受容されず、保証説では証言として依拠する側の知識にはならない、ということもありうる。四つを分けているのは、予測が成功したかどうかではなく、受け取る側にどの達成が、どの経路で成立したと数えるかである。事例を示せて、成功を、機械自身の選択との一致によってであっても採点できるところでは、人が機械ネイティブな種類の一部を翻訳なしに学ぶ道もある。ただし、その根拠は、著者たち自身が概念実証として示すチェスでの一つの試みであり、提示できない種類や特徴にかかわる場合や、 S が結果を自ら採点できない場合に、その道が何を要求するかまでは示していない。そもそも異質であることは出力のもつ一つの性質ではない。後期のクーンの説明を当てはめれば、内部の特徴だけが異なるなら不透明性にとどまるが、種類の分け方そのものが横切るなら、クーン的前提では、人間がそれを学べることはあらかじめ保証されていない（第8章第3.1節）。

第8章は、 S による区別を二つの地平としてまとめた（第8章第4節）。人間の社会の地平で問われるのは、研究が新しい知識をどうもたらすか、選択が機械の側に移っても研究が S の未知に向けられ続けるか、 S が評価の有能さを保ち、基準が動いたかを見分けられるか、知識がどう S に入るかであり、それらを、依存、無知、人間の関与の範囲が見える状態を保ったまま問うことである。人間の社会のための研究は、機械の意識も、完全に独立した認識論も、人工社会も必要とせず、部分的な能力とハイブリッドな組織を通じて進みうる。人工社会の地平では、人間の社会は S ではなく、その研究によって影響を受ける当事者になり、人間の社会にとって問われるのは、研究の成功ではなく、受け入れ可能性と正統性である（第8章第3.2節）。そこでは、その社会が自らの認識的世界を形成し、維持し、批判し、変容させようかが問われ、信念と社会的な知識、世界への応答可能性、目的の自己設定と基準の改訂、人間の社会との協調について、より強い問いが立つ。第8章は、そのような社会が可能であるとも不可能であるとも示しておらず、二つの地平の区別は、根本の問いを先送りする予定表でも、独立していることがそれ自体でよいという主張でもない。混合社会は、どちらの地平にもまるごとは取まらない。人間の成員については人間の社会の地平の問いが、人工の成員については、知識の帰属先に数えられる限りで、信念と保有の問いが立つ。成員が別々の共同体をなすなら協調の問いが、一つの記号体系を共有するなら、それがどの成員の経験を統合したものかの問いが加わり、どの場合にも、誰が S を代弁するかが問われる。

第IV部は、探究がどこで遂行されるかという選択と、誰の認識的な営みを広げるかという選択とを分けた。依頼も、人間による制御も、研究が S の未知に向けられ続けることを保証せず、それは個々の構成について問われる。また、 S のための研究の認識的な成功と、その方向を定めることの正統性とは別々の判断であり、後者は S の外の、影響を受ける当事者にも問われる。研究が S のどの未知を追い、その結果が S にとって通用するものとなり、誰の権限でその方向が定められるのかは、単一の行為者を越えて広がる関係にかかわる。そうした関係のどれを人工の構成の内側に置けるか、そのとき何が変わるかを、第V部が問う。

6. 研究を取り巻く関係を組み替える

人間の研究は、研究する人々だけで進んできたのではない。たとえば、研究者がどの問題を選ぶかや、資金提供者がどの課題を選び、研究室の長がどこに人と設備を振り向けるかが、労力の行き先を決めてきた。出版前の査読、追試、結果の公開とその後の利用が、結果をほかの者が頼れるものにしてきた。そして資金提供者、自らを統治する研究共同体、規制者、公衆などが、そうした仕組みの規則を定め、改めてきた。AIが研究するとき、人工の構成がそうしたことの一部または全部を、自らの内側で行うようになるかもしれない。そのとき研究に何が起こるのか。第V部はこの問いを、制度の形からではなく、制度が担ってきた関係の側から考えた。資金配分のような仕組みが実現してきた配分、査読のような評価が実現してきた批判と認定、そしてそれらの条件を誰が決めたのかという権威である。問いは、研究が S のために方向づけられ、認定され、統治される関係のうち、どれを人工の構成の内側に置けるか、置くとか何が変わるか、であった（第9章第1節）。前半の問いは後半を描くための手立てであり、人工の構成が研究を遂行できるかという第IV部の問いとは別の問いであって、主に価値とガバナンスの水準にかかわる。ただし認定は、社会的な組織が知識の根拠づけや保有に入るとする説のもとでは、知識の条件にも入る。ここでいう認定は、第III部の検査とも違う。第III部は、結果が何にかかわり、合格したチェックがどれほど強く決着させるかを問うた。認定はさらに、そうして評定された支持が、特定の用途のために、ほかの者が点検し直さずに依拠し、それにもとづいて行為するのに足りるかの判断を含む関係である（第11章第1節・第4節）。

第9章はまず、百のインスタンスを生成して異なる路線を追わせ、別のインスタンスにその産出物を点検させ再現させ、その点検に応じて路線のあいだで計算資源を振り向け直すシステムを例に置き、その構築者がなしうる三つの主張、すなわち、うまく配分している、認定された結果に他者が依拠する理由がある、他者に代わって決めてよい、という主張を立てた（第9章第1節）。そのうえで、三つの関係を、作業的特徴づけの語が特徴づけの外の条件と出会うところから導いた。配分は、「 S にとって新しい」という語に、手段が有限で研究者が多数いるという条件が加わるところに生じ、 S の未知のあいだでの注意、労力、資源の分布を指す。批判と認定は、認識的達成、とくに知識という語に、成員が多くの結果について自ら証拠を集められないという条件が加わるところに生じる。権威は、 S が一つの声をもたず、帰結が S の外にも及ぶところに生じ、決定されるところでは誰が決めてよいのか、誰も方針として決め

ない過程のところでは誰が責任を負い変えてよいかを問う（第9章第2.1節）。これらは制度ではなく分析上の関係であり、資金配分機関や査読などの制度は、そのありうる実現の一つである。どれも研究であることの条件ではない。第9章はまた、関係が別々の当事者に分かれていたことが、どの当事者も供給するよう設計されていない財を供給していたかもしれない、という仮説を立てた。境界をまたぐ誤りの検出、異論の当事者資格、説明が求められた境界、責任を負う当事者である。単独で評価した方法の成績が共同体のなかでの成績を示さないことがあるように、個々の研究者の良さから研究システムの良さは読めない。また、研究者の形成を重んじる立場などのもとでは、同じ知識を同じ S に届ける二つの研究の組織も、誰がその過程に加わり、過程のなかで誰が変わるかによって、価値において同等とは限らない（第9章第1節・第3節）。

そこで第9章は、第5章第6.5節と同じ仕分けを用いて、関係の分割が供給したかもしれない財が何に依っていたかを、四つに仕分けた。人間という実現者についての前提、現在の科学の組織についての前提、探究の条件、そして S を含む構成の外の当事者との関係である（第9章第6節）。前の二つは、実現者や組織が変われば崩れるが、そのとき問題は、前提とともに消えることも、形を変えて残ることもある。探究の条件は取り除かれないが、構成が内側で満たしうる。外の当事者との関係は、相手の参加や承認を含むなら、片側から完成させることはできない。どの前提がどこに入るかは採る説によって変わる。財が外の当事者との関係に依るなら、採る説のもとでそう言える限りで、内側に置けるという主張には、能力だけからの推論ではなく、その関係についての証拠が要る。また、人工の構成は、違いと誤りの相関が作り方によって決まること、研究を遂行する構成そのものを供給する提供者がいること、成員を思うままに増やせることといった、それ自体の条件をもちこむ。

配分とは、そもそも何をする事なのか（第10章）。配分には二つの問題が重なっている。どこに何があるかを探し当てる探索の問題と、どれが追究に値するかという目的の問題であり、前者の完全な解も後者を残す。配分の問題が生じるのは、すべての路線を少ない費用で、安全に、やり直せる形で追えないときであり、よい広がりかどういふものは、前提する目的によって変わる（第10章第1節）。研究者を思うままに動かす統制と、理論の客観的な長所を見抜く眼の両方をもつキッチャーの哲学者君主の思考実験を、二つの力を分けるために用いると、統合は前者に近づきうるが、後者を与えない。うまく配分しているという主張は、ほかに利用できる取り決めと比べて検査することはできても、哲学者君主が選ぶであろう分布と比べて検査することはできない（第10章第2節・第7節）。また、前もって見込まれる重要性の基準を外しても、どこに到達できるか、どの順で調べるかの選択は残るので、広い意味での価値から完全に自由な探索はない。キッチャーの説のように配分が重要性にはね返るなら、その選択は後の重要性の判断も制約する（第10章第2節・第6節）。研究者に功績を求める動機をもたせずに直接割り当てる配分は、動機への依存を除くが、評価の信号への依存は除かない（第10章第3節）。

多様性についても、何のための多様性かが問われた（第10章第4節）。ゾルマンらのモデルでは、信念の何らかの一時的な多様性が必要とされる。それを生む特定の機構の利点はもろく、保つに値する多様性が一時的なものに限られるかも争われているが、モデルの構造が人工の構成に当てはまるなら、統合の危険は、臨床、野外、希少な装置のように証拠が遅く高価な領域に集中し、そこは、人工システムが多くのデータを生み出すことで埋め合わせる余地の最も小さいところである。課題の易しさは、それ自体が課題の分け方の産物でありうるので、同じ評価の信号が次の配分に戻るなら、判定しやすい課題での成功が次の配分を正当化し、研究の範囲が成功そのものによって内側から狭まることもありうる。出自の異なるモデルを並べても、多様性は保証されない。調べられた言語モデルの範囲では、提供元が違って誤りは相関しており、生成された仮説についての二つの研究は、同じ注意を仮説の提案にも広げる。中央化と引き換えになりうるのは、多様性そのものというより、収束を決める基準についての誤りの独立性である。異なる表象を生み出せることと、異なる証拠で路線の存続を判断できることは別だからである。また、好まれない路線をなお支持者とともに残すという異論の認識上の働きは、保護された予算や再開の規則、分けられた権能によって内側で果たしうるが、異論を唱える当事者資格は、採る説のもとで部分がそれをもつ場合を除けば、外の当事者との関係として残る。

人間の配分は、誰もそうと決めずにいくつかの働きを供給していた（第10章第5節）。好まれない路線を保つこと、異なる種類の路線を開くこと、収束の判断、確立した参加者の支配を終わらせ、路線を追う者を入れ替えること、複数の信号である。ただし、その背後にあったものは事例ごとに異なる。長期の研究を支援する取り決めは意図して設けられたが、共同体全体での効果は意図されていなかった。研究者が退くことによる交代は寿命によっていた。逆に、中央で研究の方向を指揮した 523 計画の事例は、異なる種類の路線を開くことや収束の判断を決めることと、それらが要する判断を下すこととは別であることを示唆する。人工の構成では、これらの働きは、誰かが定めるパラメータにもなりうるし、構成の内側でも誰も方針として定めない過程によって供給されることも、気づかれぬまま失われることもありうる。失われた場合には、記録さえ残らないことがある。好奇心に導かれた研究者たちの分散した探索も、配分の方法として暗黙に採られてきた発見的方法の一つと読める。その採否は、人工の構成のもとでは、暗黙のままにはとどまりにくくなるかもしれない。

そして、誰の未知を追うのか問われる（第10章第6節）。 S の成員にとって何が重要かを知る認識的な役割と、誰の問いを根に置き、どう重みづけるかを決め、誰も決めなかった順序づけに責任を負う役割とは別であり、前者の有能さは後者を決着させない。人工の構成を論じていないキッチャーの重要性の説を当てはめれば、人間の社会のための研究では、構成は記録が反映する重要性の周縁を受け継ぎ、扱いやすさが注目を動かすはね返りも起こりうる。立脚点理論に立てば、構成は記録を生んだ者たちの立っていた場所から出発し、記録が排除してきた分だけ欠く。パルサフスキーとクリモフスキー、モロ・シンプソンの論争に照らすと、結果の真理、問いの重要性、答えへの到達の時期は別々の次元であり、これを当てはめれば、真である結果を生みながら、他所の基準を受け継いで到達を遅らせることもありうると思える。第10章はこれらを合わせ、判定しやすい課題での成功から、扱いやすいところへの移動、重要性の見え方の変化を経て、一つの信号の誤りによって路線の打ち切り判断が相関するに至る、条件付きの連鎖を描いた。どの環にも人間の研究についての記述やモデルのなかに先例があるが、人工の構成では、最初の条件が満たされやすく、連鎖を断つ働きを前提できず、構成自身の記録が研究の主な記録であるところでは狭まりがそこに現れにくい、という三点が違いうる。それでも、これは避けられない成り行きではなく、対応する環で断たれうる（第10章第7節）。

批判と認定については、 S のための批判と認定を人工の構成の内側で組織できるかが問われ、そのなかで、そもそも査読は何をしてきたのかも問われた（第11章）。第11章は、査読をその働きの一つの歴史的な実現とみて、その働きを、簡い分け、紙面や資金の配分、誤りの検出、叙述の改善、新しさと重要性の評価、承認の六つに分け、同じ段階がそれに伴って生んできたと考えられるものとして、評価が S に見えること、 S の成員が評価の経験を積む場、名指せる判断を異議を向けうる形で残すことを挙げた（第11章第7節）。人工の構成は、評価者を募るのではなく生成でき、生産と評価を一つの構成に置いて評価を形成の途中に入れられ、評価者と生産者を時間にわたって結びつけられるので、「査読を通った」という時点そのものがなくなりうる。そのため問

いは、査読を残すかどうかではなく、どの働きの組み合わせが、どの説の求めるものを満たすかになる。第11章はまた、候補やその根拠が誤っていればそれを示しうる経路、つまり無効化の関係と、他者が特定の用途のために結果に依拠する理由を与えられる関係、つまり認定とを分け、前者のどこまでを内側で確保でき、後者がなぜ *S* との関係として残るのかを問うた（第11章第1節）。

その答えは、説によって分かれる。人間の研究では、批判の説がそれぞれ追う性質、たとえば点検の独立性、答える誰かがいること、価値づけの共有、評価が見えることは、人間という実現者と組織によって結びつけられていたので、説の判定は重なりやすかったと考えられる。人工の構成はそれらを別々に設定できるので判定が分かれ、内部の点検を通った結果は *S* のために認定されている、という構築者の主張は、どの説に答えるのかを示さない限り評価できない。そしていくつかの説のもとでは、内部の批判をよくすることが、*S* のための認定をかえって弱めうる（第11章第1節・第9節）。批判を共同体の客観性の条件とみるロンジーノの説を当てはめれば、仮定を共有する批判者を増やすと、共有された仮定が見えないまま強まりうる。証言への依拠の正当化についての非還元主義の最小の形では、所見が依拠する者に届く経路がなければ、無効化要因がないという条件は誤りの有無にかかわらず満たされるので、内部の点検を強めても依拠する者の正当化は変わらない。そのため、精度は劣るが *S* に開かれた点検のほうが、精度は優れるが *S* に閉じた点検よりも依拠の根拠を多く与えることがありうる。ただし、誤り検出の見方と信頼性主義を合わせる立場では、内部の批判をよくすれば認定もよくなり、この比較は、点検の信頼性だけを問う純粋な信頼性主義のもとでは成り立たない（第11章第9.2節）。

独立性も、数や組織の分離では決まらない（第11章第2節・第3節）。情報源のあいだの一致についてのゴールドマンの議論を点検一般に延ばせば、ある点検が独立であるのは、既存の点検が与えている以上に誤りを見つかる情報を加える程度においてであり、別の組織にあることは必要条件でも十分条件でもない。単一の基盤から作られた百万のインスタンスからなる共和国を思考実験として置くと、ロンジーノの説のもとでは、それはせいぜい自らの視点に相対的に客観的であり、成員を増やしても、共有した組織のもとでは行えない批判は加わらない。*S* がその組織を共有するなら、批判がよくなったわけでもないのに、共和国に相対的な客観性と *S* に相対的な客観性との隔たりは縮まる。証拠を加えるのは数ではなく違いだが、判定を集約するときには数が影響力として効くので、九十の複製と十の異なる批判者で多数決をとるなら、複製の比率を決める者が判定を左右しうる。違いを作り込むこともできるが、違いを選ぶ過程そのものが共通の源になりうる。

認定には価値もかわる（第11章第4節）。認定は、程度をもつ支持を、ほかの者が点検し直さずに行為してよい地位に変える働きであり、帰納的リスクからの論証を認めるなら、どの誤りをどれだけ重く見るかの判断を含む。その重みづけは閾値の一点だけでなく、モデル化の選択のたびに入り、確立したとされる知見の範囲を前もって形づくる。エリオットの紹介する議論を当てはめれば、その重みが構成を作り提供する者の設定を映すとき、設定が変わるだけで、誰も選び直した覚えのないままその範囲が変わることもある。そして研究では、退けられた候補はたいてい追われず、*S* にとって新しい結果には照合する別の出所がないことが多いため、重みづけの誤りは表に出にくい。依拠についても、紹介したゴールドバーグの説を当てはめると、認定が道具の示す値として扱われるようになるとき、*S* は研究者に問い返す形を失い、設計者や作り手の認知的な仕事に依拠する形へと移りうる。ただし、そうした設計者や作り手への信頼が、本質的に不透明な AI への依拠を支えられるかを、コスキネンは疑問に付している（第11章第5.1節、第5.3節）。範囲が定まり、改変できない基盤に記録され、安価に再実行できるよう組まれた実行の記録は、研究者だけがデータと過程の記録をもつという前提を弱めるが、問題は、記録の維持、記録の境界、その内側で検証する能力という三つの場所で形を変えて残る（第11章第5節）。認定された説明が積み上がりながら、*S* の誰もその現象を理解していない科学もありうる（第11章第5.6節）。

時間も考えに入れる必要がある（第11章第8節）。評価者と生産者が結びつくと、点検の独立性は、誰も欺かなくても、点検が選抜に使われるだけで衰える。手立てになりうるのは、書き換えを禁じることより、選抜に使わない点検を残すことである。基準が結びつきとともに動くところでは、いまの基準との一致は、基準が動いていないことの証拠にならない。基準の動きを以前の参照点に照らして検出することは内側でもできうるが、その動きを改善とみるか置き換えとみるか、受け入れるかは、それが *S* の基準である限り *S* が決める。訂正も、記録の外へ届く経路がなければ届かない。そのため、時間を考えに入れる限り、「すべてを内側で」という答えはとれないかもしれない。基準の動きと訂正について要るのは、人間であることではなく、*S* との関係である。そして、*S* が人間の成員を含むとき、その成員が素人になる極限を思考実験として置くと、まず失われうるのは真理への接近ではなく、目的を決めることへの実質的な参加と、それを支える能力である（第11章第6節）。

権威については、そもそも誰が決めてよいとはどういうことかが問われた（第12章）。第12章は、配分の条件と、批判と認定の条件とを決め改める当事者資格、つまり決定に服する者との関係で、提案し、決め、異議を唱え、拒み、改める資格を権威と呼び、それは有能さでも、形式上の割り当てでも、実効的な力でもないとした。よく働くことは、それだけでは決めてよいことを示さない。そこで第12章は、正しい順序づけと重みづけは何かという第一の問いと、誰がそれを他者に課し変えてよいかという第二の問いを分けた。第一の問いには、採る価値の説によっては内側で答えるが、第二の問いに答えるのは正統な権威の説である（第12章第1節）。また、決定には、ある主張や結果が特定の用途に足るほど信用できるかにかかわるもの、アジェンダと配分、運用とリスク、誰が成員や批判者や専門家と数えられ、どの証言が聞き届けられるかにかかわるもの、記憶と訂正、そしてこれらの目標や基準や手続きを誰が改めてよいかにかかわるものがあり、どの根拠もそのすべてに同じ効力をもつわけではない（第12章第2節）。比較の事実をすべて見いだしても、どちらとも順序づけられないが同等ではある選択肢が残りとするチャンの説を当てはめれば、知識を増やすことは権威の問いをなくすとは限らず、知識が尽きたところに残る権威の問いを取り出しうる（第12章第1.2節）。

権威の根拠によって、答えは分かれる（第12章第3節）。純粋なエピストクラシーを除けば、有能さだけからは、目的を他者に課す当事者資格は生じない。結果に基づく根拠のもとでは、有能さそのものではなく、有能さを仮定しない評価によって確かめられた優位が、その範囲で人工の構成それ自体に当事者資格を与えうる。ただし、ラズのサービス構想のもとでその独立性テーゼを当てはめれば、*S* の成員が自らの理解にもとづいて選ぶことのほうが重要な事柄は除かれ、目的の選択はその候補になる。研究共同体の自己統治を根拠とするなら、権威が内側に置かれうるのは中央の配分者ではなく、自らの発意で路線を取り上げる部分のあいだであり、それも、それらの発意が互いに独立している限りでの、分野内の開かれた問題選択に限られる。実際の授權を根拠とするなら、構成がもちうるのは授權された範囲での、異議を受けうる実効的な決定の力までである。同意を根拠とするなら、同意する者が授權した範囲であり、読み方によっては修正する能力を手放すことまで含むうるが、それが後に来る者に及ぶかは決まっておらず、自発主義の説では及ばない。意図に沿うシステムが操作者に権力を集中させるというフリーデリッヒの議論を、研究と、支配からの自由を求めるベティットの共和主義の概念に移すと、委託者の意図に完全に沿う構成は、委託者が異議に応答せずに *S* の探究を配分し直す手段になりうるので、完全なアラインメントは支配を可能にしうる。そし

て、これらの判定がそのまま当てはまるとすれば、有能さが上がるにつれ、争いは有能さについてではなく、どの根拠を採るかにについてのものになるかもしれない、人の当事者資格を残す理由は、いっそう中心に置かれると同時に、いっそう満たしにくくなりうる。結果に基づく根拠のもとで、確かめやすさの違いが決定の種類の違いにおおむね沿うとすれば、確かめやすい問いほど権威が入りやすく、そこでの成功が権威の範囲を広げるなら、研究の方向がそちらへ偏ることもありうる（第12章第8節）。第12章第8.4節は、これらの判定から、権威が内側に入ったとき研究がどういう姿になるかを描いた。同じ人工の構成でも、根拠ごとに、迫られる問いの偏り方、 S の知り方、直す者の位置が異なるのである。

当事者資格は残っていても、名目だけのものになりうる（第12章第4節）。ロンジーノの説が正しければ、内容についての服従は背景仮定を通じて目的をも持ち込みうるし、方向を問われる構成が、方向を決める者の選好を教育する位置に構造的に立つこともある。人間の社会が当事者資格を保ちつつ、実効的な決定の力が構成へ移る場合（第12章第4節が考えた場合）、その当事者資格が名目以上であるためには、構成が当事者の理由を追跡し、その振る舞いが適切な理解をもつ人間に遡れること、結果を実際に変えられること、選好が当の構成だけを通して形づくられるのではない経路があることが求められる。ただし、正統性を仮説的な手続きや単一の場に置く説では、最後の条件には別の根拠が要る。修正可能性についても、内側からの修正は頻繁になりうる一方で、外から異議を唱え止める手段は乏しくなりうる。研究の結果に異議を唱えるには多くの場合研究そのものが要り、打ち切られた路線になお見込みがあると示すには、それを追ってみるほかないことが多い。そのため研究では、異議を唱えられるかどうか、研究を行う手段にも依存する（第12章第6節）。そして修正する能力は、誰の委任もなく、構成自身の通常の研究上の改訂を通じて失われうる。後で修正できるからよいという原則は、それが要るところほど弱く、修正する能力を取り除く決定は、その原則から正統性を引き出せない（第12章第7節）。

A も S も人工であれば、権威の関係は消えず、成員のあいだの関係になる（第12章第5節）。誰が成員を作るかは、視点を数えるなどの集計にも先立つ権威の決定であり、 S 自身がそれを繰り返す行いもある。代表性は、それ自体を満たせる S のもとでは成員の形成を正当化しない。共通の理想をもたせて成員を作ることが互いの利益への配慮を強めるなら、ヒュームの論証と、聖人の結社についてのロールズの指摘を当てはめることで、ロールズの意味での正義をめぐる争いは生じにくくなりうると思えられる。さらに、問いや証拠についての見方まで一致させるなら、争いの少ない社会であることと、誤りを共有しにくい研究共同体であることは、相反するのかもしれない（第12章第5節）。他のどの当事者も存在しない極限を思考実験として置くと、消えるのは誰が決め改めてよいかという権威の問いであり、誤りが示されうること、知らずに労力を分けること、基準の改訂が改善なのか置き換えなのかという問いは残る。後者は、研究がまだ知られていないものを求める試みであることから生じる問いとして読める。

第V部の全体からは、研究のやり方を速く改められるようになることと、その改め方を外から確かめ、異を唱え、止める手段が薄れることが、同じ変化の両面として起こりうるが見えてくる（第12章第9節）。検討した諸説は、評価される側だけでは定まらない基準点に、それぞれ異なる理由を与える。これらの説が一つの条件を共有すると読むなら、どの論者もこの形では述べていないが、守られていたのは、判定に効く中身が評価される当事者の外から来ており、評価の時点でその当事者が置き直せないことである。これらの論者は人工の構成を論じていないが、この読みのもとでは、批判者、代替案、同業者を自ら生成する構成は、諸説の規範を書かれたとおりに誠実に満たしながら、規範が拠っていた理由がもはや当てはまらない状態になりうる。したがって、自ら生成したことだけでは、それらが必要な働きを果たすとは確立されない。ロンジーノの説で問題となるのは、批判者が共有された仮定を引き継ぎうることであり、生成されたという事実だけで批判が無効になるわけではない。人間の研究もこの条件を完全に満たしてきたわけではない。満たされていたところでも、それは、生み出すことが散らばっていたために、各当事者から見て中身が他の当事者から来ていたという意味においてであり、その散らばりが、点検が個々の当事者の共有仮定を免れる助けになりえた。またゾルマンの議論に照らせば、防ぐべき失敗が生成する側の性質でないところでは、生成された安全策でも足りうる。どの説を採るかに応じて、支えとなりうるものには、外で出会った批判者、構成が作っていない同業者、判定される結果より前に日付をつけて記録された立場、以前のやり方の並行運用、構成が取り消せない負担を課す異議、基準点を複数にしておくことなどが含まれる。これらは、評価の時点で評価される側によって動かされえないという意味で固定されるのであり、修正そのものを禁じるわけではないので、修正可能性とはぶつからない。最後に残る問いは、それがなお S のための研究であり続けるか、である。内側に入りえたのは主に作業の供給であり、例外は結果に基づく根拠であった。可能性と限界は、その根拠とともに表12.5にまとめられている。別々の組織は、ときにこれらの財を支えたが、その分離だけが源だったのでも、一律の要件だったのでもない。

7. 決めずに決まっていたものを名指す

本章第3節から第6節は、各章が見いだしたことを章ごとにたどった。それらを章をまたいで見直すと、いくつもの章が、それぞれの文献を通して、同じ形にたどり着いていたことが見えてくる。人間が研究するとき、信念をもつ者、証拠の範囲、能力の座は、一人の人であることによっておおむね定まっていたし、結果を生み出した者がそれを理解する者だということも、多くの場合前提できた（第2章第9節、第3章第9節）。また、研究を行うことと、それを配分し、批判し、認定し、決めることは、多くの場合別々の当事者が担っていたので、批判者の数が独立性をおおよそなぞり、認定が誤れば答える誰かがいた（第11章第9節）。こうしたことの多くは、誰かが決めたからではなく、研究する者が人であり、多くの人に分かれていたことによって、決めなくても決まっていた。機械が研究するとき、それらはあらかじめ定まっているとは限らない。研究する者の境界は主張に相対的に引かれ（第7章第2節）、知識の主張の単位も、インスタンス、外部の記録、道具がどう結びつくかによって何通りにも引けることがあり、それを引くのは作り手であることも、自らを組み替えるシステム自身であることもある（第2章第9節、本章第3節）。誰も決めずに供給されていた働きも、誰かが規則として定めることにも、誰も定めないうまま内側の過程のなかで形づくられることにも、気づかれずに失われることにもなりうる（第9章第2.3節、第10章第5節）。人間が研究するときには決めずに決まっていたものが、機械が研究するときには決めうるものになるのである。

決めずに決まっていたものが決めうるものになるというこの形からは、二つのことが出てくる。一つは、「この機械は研究している」という主張が、別々に成り立ちうる主張の束になることである（本章第1節）。もう一つは、検討した説の判定が分かれることである。人間の研究では、説がそれぞれ追う性質を、人間という実現者と研究の組織が結びつけていたので、多くの日常的な事例では判定が重なりやすかったと考えられるが、人工の構成は、それらの性質を一つずつ別に設定できるものにする。束がほどけるのである（第11章第9節、本章第6節）。第11章が批判と認定について見たこの事情は、本書が多くの場合に答えを一つに定めず、説ごとに帰結をたどってきたことの理由の一つとしても読める。認定の主張も、決めてよいという主張も、どの説に照らし

てなされるのかを示さなければ確かめられず（第11章第9節、第12章第8.2節）、知識の主張も、どの単位についての主張なのかを示さなければ確かめられない（第2章第9節）。ただし、ほどけた束は、異なる基盤、依拠する者に届く記録、名指せる答え手などを備えて、束ね直すこともできる（第11章第9節）。評価が S に見えることや異議の対象になることは、評価の段階に結びついていたので、記録や公開の場、異議を受け付ける経路のような仕組みを外に足せば作り直しうる。これに対して、誤ったときに答える誰かは段階ではなく人に結びついていたので、そうした仕組みを足しても生まれず、保証説では結果を意図して保証する誰かが要り、苦情の権利を加える版では、誤りに答える誰かも要る。評価する力も、 S の成員が実際に評価することで保たれてきたので、仕組みが結果を返すだけでは戻らず、人間の成員については評価に加わる機会が要り、成員が人工であれば、作り直しや更新を経てもその力が保たれるかが問われる（第11章第6節・第9.2節）。問われるのは、どれを、何によって束ね直すかである。

この形は、研究において特別な重みをもつ。研究が求めるものは S にとってまだ知られていないので、その結果は、多くの場合、 S が望んだものに照らして単純には確かめられず、照合できる別の情報源もないことが多い（第9章第4節、第12章第6節）。打ち切られた路線に異議を唱えるにも、多くの場合それを追ってみるほかない（第12章第6節）。もちろん、答えをもたないことは、基準をもたないことと同じではない。形式的に述べられた定理の検査や、前もって登録された予測のように、答えを知らなくても適用できる基準はある（第9章第4節）。それでも、その基準が元の問いを捉えているかどうかは、合格したことからは出てこない。しかも研究では、何を成功とみなすかも、何を基準とするかも、ときに S が何を望むかも、探究のなかで改められうる（第6章第5節、第12章第9節）。ロンジーノの説では、批判には共有された仮定を露わにしうる視点が要る。ギャリソンが述べる機械的客観性では、探究者の願望が結果を形づくることを防ぐことが求められる。ヴィルホルトの説では、協調は実際になされた価値判断に基づく（第12章第9節）。これらの説が、判定に効く中身が評価される当事者の外から来ており、評価の時点でその当事者が置き直せないという一つの条件を共有すると読むなら（第12章第9節）、答えに照らして確かめられないところで、この独立性が、答えの代わりの一部を務めていたと読める。人間の研究もそれを保証しなかった。そして、この条件を満たしていたところでも、それはおもに、生み出すことが多くの当事者に散らばっていたために、各当事者から見て中身が他の当事者から来ていたという形においてであった（第12章第9節、本章第6節）。誰も決めなかったということの中身の一部は、誰も揃えられなかったということだったとも読める。揃えられなかったものは、機械が研究するときには、揃えうるものになる。批判者、基準、 S 、成員の範囲、是認といった評価の基準点そのものが、評価される側によって作り出されうる変数になるのである。第V部で見たこの形と、一人の人であることが定めていたものが定まらなくなるという第II部から第IV部の形とは、同じところ、すなわち研究する者が人であり、多くの人に分かれていたことから来ていると読める。

決めずに決まっていたものは、言葉の側からも見えてくる。たとえば「AIは問いを立てられない」「AIには好奇心がない」「独立に査読された」のような言い方のなかで、性質を述べるように使われる語の多くは、各章で掘り下げると、誰にとつてか、何についてかを名指さなければ中身の定まらない関係として見えてきた。機械が問うたかどうかは、問いにかかわる仕事のどれを、どの単位が、誰にとつて未決着の点に向けて担っているかとして問われる（第4章第9節）。独立性は、誤り検出の見方では、点検が既存の点検に加えて誤りについての情報をどれだけ加えるかである（第11章第2節）。新しき、不透明さ、自律性も、誰にとつての、何についての、誰との比較かを伴っていた（第1章第6節、第3章第5節、第7章第4.1節・第4.2節）。好奇心は、その分かりやすい例である。ゴブニックは、ひとりの個体が先に探索し後で活用する場合には、学ぶ者の利害と、学ばれたことを活用する者の利害は自動的に一致すると指摘する。これを研究に当てはめると、探索する者と、研究がそのために行われる者とは A と S であり、両者の利害は自動的に一致しない（第4章第7節）。この読みでは、好奇心について問われるのは、好奇心をもつかどうかに加えて、それが誰の未知へ向かう探索なのかである。人間の研究でも A の未知と S の未知は分かれうるが、そうした項のいくつかは、一人の人であることや、 S が人間の社会であることによって暗黙のうちに埋まっていたので、語が性質を指すように見えやすかったと考えられる。それらの項を名指すことで、機械が何を成し遂げたという主張なのか定まる。

人工の構成では、それらの項に何を当てるか、つまり何を参照先とするかが、選べるものになる。ヒューズらは、システムの開放性を観察者に相対的に定め、観察者の選択を「非常に重要な自由パラメータ」と呼んだ（第4章第7節）。ただし研究では、その値は、定義の外で擁護されなければならない。事後的に寄せ集めた集合を S とすることはできず（第1章第6節）、人工の構成と来歴を共有する人工の社会を名指せば、認定の主張は、認定がよくなるまま満たしやすくなる（第11章第9.2節）。参照先を選ぶことそのものが、擁護を要する主張になるのである。しかも、名指す必要があるのは、どの時点の S かでもある（本章第2節・第5節）。

多くの人に分かれていたことがもたらしていたものの一つは、批判者や探索者のあいだの違いであった。違いが、生じるものではなく作られるものになるとき、何が安く、何が乏しいかも変わる。問い文や候補を出すこと、批判を加えること、試行を回すことなどは、機械にとって安くなりうる。とくに批判と試行についていえば、証拠として加わるのは数ではなく違いであり（第11章第2節・第3節）、安くなるのが、生産者の記録の上で働く批判や、同じ配置のなかで繰り返せる試行であるのに対して、乏しいままでありうるのは、生産者が用意しない経路を通り、それ自身の誤りの経路をもつ点検である（第11章第7節）。同じ基盤のインスタンスを加えても、それだけでは、点検や観察や批判が頼る違いは加わらないということには、第II部から第V部のどの部も、そのつど別の説を通ってたどり着いていた（本章第3～6節）。研究の過程の側でも、すでに特徴づけられた空間を広く探すことは安くなっても、表象や装置を作って候補を利用可能にすることや、試行を前の結果に結びつけて診断にすることは、それとは別の仕事として残る（本章第4節）。

そのため、機械が研究するときには、量がもつ認知的な重みそのものが問いになる。豊富な批判は、独立した批判の乏しさと共存し、それを覆い隠しうる（第11章第7節）。時間を入れると、独立した点検は、使うと減る資源のように振る舞うかもしれない。選抜に使われる点検は、使われるほど生産者について教えなくなりうるし（第11章第8節・第9.2節）、批判によって変わった生産者が、点検の感度の低いところで通るようになれば、同じ点検が次にも同じ情報を加えるとは限らない（第11章第9.3節）。何が乏しいかは、探究が進むにつれて変わりうるのである。自律性を高めることが、外からの点検をかえって要するものにしうるのも、同じ形として読める（本章第5節）。いずれも、安いものを増やすことが乏しいものを増やすことにはならない、という形である。何を増やせば研究が進むのかは、何が乏しいのかによって変わる。

乏しいままでありうるものを数えることは、裏返せば、人間の研究がそれを何によって供給していたのかを問うことでもある。序論は、人工の実現を考えることが、人間の研究から受け継いだ前提を試しめると述べた（序論第1節）。各章で見えてきたものの一つは、人間の研究が、誰もそうと決めずに供給していた働きである（本章第6節）。たとえば、名指せる人が出版の前に原稿を読むという一つの行為は、誤りの検出と同時に、評価が S に見えること、異議を向けられる判断、答える誰か、 S の成員が評価の経験を積む機会を生んでいたと考えられる（第11章第7節・第9.2節）。答えに照らして確かめられないところで、この一つの行為が、その代わりの一部を務めていたと読める。査読という実践は、成果を判断することと、それを判断できる社会を保つこと

とを、同時に担っていたのかもしれない（第3章第8節、第11章第6節）。そうであれば、研究の結果を受け取る S の力も、研究への参加を通じて保たれていたことになる。

もつとも、人間の研究も、判定に効く中身が評価される当事者の外から来るという条件を、完全に満たしてきたわけではない（第12章第9節）。ロンジーノの説では、全員が共有する仮定は、内部の批判を生き延びて隠れたままでありうる（第11章第3節）。見えてくるのは、人間の研究がよくできていたということではなく、働きと、それを担っていた仕組みとが、別のものだったということである。キャンベルの記述的認識論を当てはめると、人間についての前提を外すことは、知識を運ぶ乗り物をなくすことではなく替えることであり、そのとき問われるのは、どの維持の条件が代わりに探究を方向づけるのかであることが示唆される（第9章第6節）。事柄が S だけでなく A にとっても未知であれば、どの路線が成功するかを前もって知らないまま労力を分けなければならないこと、点検がそれ自身の誤りの経路を要することなど、誰が研究しても取り除かれたい探究の条件もある。人工の構成は、それを内側で満たしうるが、満たしていることを示す必要がある（第9章第6節、第12章第5節）。機械が研究することを考えるのは、研究について何が人間に固有の前提に支えられ、何がそうでないかを見分けようとするということでもある。序論が、ワイスパークとライスマンの頑健性の考えを借りて、本書の問いを、表象の枠組みを変えたときにどの結果が保たれるかを問うことに近いとしたのは、この意味においてである（序論第4節）。

働きと、それを担う仕組みとを分けて考えることは、それらの働きを別の仕方でも成り立たせる可能性にも目を向けさせる。決えずに決まっていたものが決めるものになるということは、序論第5節が比較のなかに残しておいた場合、すなわち研究のやり方そのものが作られうる場合につながる。本書は、そうしたやり方が作られるとも、作ることが望ましいとも論じていない。そのうえで、本章第4節から第6節でどったように、表象、装置、記録、基準から、批判者、成員、研究を取り巻く関係までが、探究のなかで、さらには人工の構成自身によって改められうるものとして現れた。検討した説を条件として見えてくるのは、そこに開かれるのが自由だけではなく、改めたやり方がよりよいかを何に照らして問えるのか、という問いでもあることである（第12章第5節）。第12章第9節は、ニールセンとチウが、科学の社会的過程のうち集合的に保持されるものがほとんど変わらない理由として挙げた摩擦 (Nielsen and Qiu 2022) と、判定に効く中身を外から供給しえた分離とが一部重なっていると考えられる以上、関係が構成の内側へ移るとき、両者は同時に取り除かれうるとした。研究のやり方を速く改められることと、それを外から確かめる手段が薄れることが同じ変化の両面として起こりうるのは（本章第6節）、そのためである。

ここで見えてくるのは、改める自由と、改めたことが誤りだと示されうる関係とが、同じものではないということである。基準を改める探究も、改められる基準以外のものによる支えと抵抗のもとで進む（第6章第5節）。本章第6節で挙げた、外で出会った批判者や、結果より前に記録された立場などは、改訂を止めるためのものとしてではなく、改訂がよりよいかを問える状態を保つものとして読める。それらは、評価の時点で評価される側によって動かされえないという意味で固定されるのであり、修正できることはぶつからない（本章第6節）。研究の代替案は前もって閉じた一覧にならないので（第12章第9節）、改訂がよりよいかを問える状態も、一度整えれば済むものではない。そして、 A と S が重なるにつれて、研究のやり方を改めていく人工の構成がなお S のための研究を行っているかという問いは、「どの社会にとってか」という問いへ移りうる（第11章第9.3節）。研究のやり方が作られうるところでは、そのつど S を名指すことが、この問いの前提になる（本章第2節）。

こうして見てきたことから、第1章の一文の語がそれぞれ何を担っていたかも見えてくる。ある社会 S に相対的に、研究とは、 S にとって新しい認識的達成をもたらそうとする試みである。「試み」は、第1章第3節が定めたとおり、関わる者すべての意図ではなく組織された方向性として働き、機械が研究するときには、その方向性がどこで組織されているかを問わせた。「 S にとって新しい」は、どの時点のどの S かを名指す負担を負った（本章第2節）。「認識的達成」を問うことは、それがどこかに実現したか、 S の認識的状況に入ったか、 S がそれを受け取り評価する力を保っているか、という問いに広がった（本章第3節）。人工の状態が信念でありうるか、人工の存在が理解し責任を負いうるか、道徳的地位をもちうるかは、なお問いとして残る。研究がそれを行う者にとって何であるかによっても価値づけられるなら、研究の遂行が人工システムに移るとき、保たれるのが探究への人間の参加なのか、その参加の意義なのか、最終決定における地位なのか、という問いも残る（第9章第3節）。本書は、それらを決着させるよりも、答えによって機械の研究についての問いがどう変わるかをたどった。

研究が求めるのは、 S がまだ知らないものである。そのため S には、照らし合わせる答えがない。人間が研究してきたあいだ、その代わりを務めていたものの一部は、研究する者が一人の人であることと、多くの人に分かれていることのなかに、名指されないまま収まっていた。機械が研究するとき、それらは、誰かが定めることにも、機械自身が形づくることにも、気づかれずに失われることにもなりうる。ここで考えたことはどれも途上にあるが、機械が研究するとはどういうことかを問うことは、研究がこれまで何によって成り立ってきたのかを、名指して問えるようにすることでもある。

参考文献

〔訳注：文献一覧は原文のまま掲載する。〕

Stable DOI and archived encyclopedia pages are preferred; mutable web sources retain an access date.

Aberdein, Andrew. 2019. “Mathematical Monsters.” In *Monsters, Monstrosities, and the Monstrous in Culture and Society*, edited by Diego Compagna and Stefanie Steinhart, 391–412. Vernon Press. Page numbers cited in the text refer to the author's version on arXiv (arXiv:1904.09308), whose first page gives the published chapter (Vernon Press, Wilmington, DE, 2019, 391–412). <https://arxiv.org/abs/1904.09308> Accessed 2026-10-29.

Abouzaid, Mohammed, Nikhil Srivastava, Rachel Ward, and Lauren Williams. 2026. “First Proof Second Batch.” *arXiv*. Preprint. <https://arxiv.org/abs/2606.18119>

Aghion, Philippe, Benjamin F. Jones, and Charles I. Jones. 2017. “Artificial Intelligence and Economic Growth.” NBER Working Paper 23928. National Bureau of Economic Research. Page numbers cited in the text refer to the authors' version 1.0 of 10 October 2017 at <https://web.stanford.edu/~chadj/AL.pdf>; the chapter later published in *The Economics of Artificial Intelligence: An Agenda*, edited by Ajay Agrawal, Joshua Gans, and Avi Goldfarb (University of Chicago Press). <https://www.nber.org/papers/w23928> Accessed 2026-10-17.

Agniel, Denis, Isaac S Kohane, and Griffin M Weber. 2018. “Biases in electronic health record data due to processes within the healthcare system: retrospective observational study.” *BMJ* 361: k1479. <https://doi.org/10.1136/bmj.k1479>

Alexander, Jason McKenzie, Johannes Himmelreich, and Christopher Thompson. 2015. “Epistemic Landscapes, Optimal Search, and the Division of Cognitive Labor.” *Philosophy of Science* 82(3): 424–453. <https://doi.org/10.1086/681766>

Alfrink, Kars, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. “Contestable AI by Design: Towards a Framework.” *Minds and Machines* 33(4): 613–639. <https://doi.org/10.1007/s11023-022-09611-z>

Allen, Lew, Roger Angel, John D. Mangus, George A. Rodney, Robert R. Shannon, and Charles P. Spaelhof. 1990. *The Hubble Space Telescope Optical Systems Failure Report*. National Aeronautics and Space Administration. <https://ntrs.nasa.gov/citations/19910003124> Accessed 2026-09-29.

Alon, Noga, Thomas F. Bloom, W. T. Gowers, Daniel Litt, Will Sawin, Arul Shankar, Jacob Tsimmerman, Victor Wang, and Melanie Matchett Wood. 2026. “Remarks on the disproof of the unit distance conjecture.” *arXiv*. Preprint. <https://arxiv.org/abs/2605.20695>

Alper, Jarod, Michael Barany, Alain Chavarri Villarelo, Sander Dahmen, Walter Dean, Karthik Ganapathy, Michael Harris, David Holmes, Mateja Jamnik, Steven Kelk, Bryna Kra, Ursula Martin, Bartosz Naskręcki, Rodrigo Ochigame, Jim Portegies, and Johannes Schmitt. 2026. “Leiden Declaration on Artificial Intelligence and Mathematics.” *Zenodo*. <https://doi.org/10.5281/zenodo.20302944>

Alston, William P. 1988. “Religious Diversity and Perceptual Knowledge of God.” *Faith and Philosophy* 5(4): 433–448. Cited from the open copy in ePLACE (Asbury Theological Seminary), <https://place.asburyseminary.edu/faithandphilosophy/vol5/iss4/6>, which carries the journal pagination. <https://doi.org/10.5840/faithphil19885442>

Alston, William P. 2005. *Beyond “Justification”: Dimensions of Epistemic Evaluation*. Cornell University Press. Pages xi–xii cited in the text are from the Preface, which the publisher makes freely available, <https://doi.org/10.7591/9781501720574-001>. <https://doi.org/10.7591/9781501720574>

Alter, Harvey J., Judy A. Mikovits, William M. Switzer, Francis W. Ruscetti, Shyh-Ching Lo, Nancy Klimas, Anthony L. Komaroff, Jose G. Montoya, Lucinda Bateman, Susan Levine, Daniel Peterson, Bruce Levin, Maureen R. Hanson, Afia Genfi, Meera Bhat, HaoQiang Zheng, Richard Wang, Bingjie Li, Guo-Chiuan Hung, Li Ling Lee, Stephen Sameroff, Walid Heneine, John Coffin, Mady Hornig, and W. Ian Lipkin. 2012. “A Multicenter Blinded Analysis Indicates No Association between Chronic Fatigue Syndrome/Myalgic Encephalomyelitis and either Xenotropic Murine Leukemia Virus-Related Virus or Polytypic Murine Leukemia Virus.” *mBio* 3(5): e00266–12. <https://doi.org/10.1128/mbio.00266-12>

Alvarado, Ramón. 2020. “Computer Simulations as Scientific Instruments.” University of Kansas. PhD dissertation, University of Kansas (KU ScholarWorks); cited by section number from the Internet Archive capture of 8 February 2026 of the repository file, <https://web.archive.org/web/20260208233942/https://kuscholarworks.ku.edu/server/api/core/bitstreams/c9df04b2-db37-40c0-ae93-851d7d7d699d/content>. <https://web.archive.org/web/20260208233942/https://kuscholarworks.ku.edu/server/api/core/bitstreams/c9df04b2-db37-40c0-ae93-851d7d7d699d/content> Accessed 2026-10-29.

Alvarado, Ramón. 2021. “Explaining Epistemic Opacity.” PhilSci-Archive. Preprint; page numbers cited in the text refer to this preprint. <https://philsci-archive.pitt.edu/19384/> Accessed 2026-10-17.

Alvarado, Ramón. 2023. “AI as an Epistemic Technology.” *Science and Engineering Ethics* 29(5): 32. Section numbers cited in the text refer to the unrevised preprint, PhilSci-Archive 21243 (dated 27 September 2022), <https://philsci-archive.pitt.edu/21243/>, whose sections are unnumbered and are counted here from the Introduction; the published version was revised. <https://doi.org/10.1007/s11948-023-00451-3>

- Alvarado, Ramón. 2025. "Challenges for Computational Reliabilism in AI and Other Computational Methods." In *Philosophy of Science for Machine Learning*, edited by Juan M. Durán and Giorgia Pozzi, 81–106. Springer. Section numbers cited in the text refer to the preprint "Challenges for Computational Reliabilism," PhilSci-Archive 23923, <https://philsci-archive.pitt.edu/23923/>. https://doi.org/10.1007/978-3-032-03083-2_5
- Amodei, Dario. 2024. "Machines of Loving Grace: How AI Could Transform the World for the Better." Essay, October 2024. Cited by section heading and note number. <https://darioamodei.com/essay/machines-of-loving-grace> Accessed 2026-10-17.
- Anderson, Elizabeth. 2006. "The Epistemology of Democracy." *Episteme* 3(1-2): 8–22. Page numbers cited in the text refer to the PDF of the article posted by the author, <https://public.websites.umich.edu/~eandersn/EpistemeAnderson.pdf>, which is typeset as published but paginated 9–23; each page number is one higher than in the issue (8–22). <https://doi.org/10.3366/epi.2006.3.1-2.8>
- Anderson, Elizabeth. 2011. "Democracy, Public Policy, and Lay Assessments of Scientific Testimony." *Episteme* 8(2): 144–164. The citation by "abstract" rests on the author's own abstract on her website, <https://public.websites.umich.edu/~eandersn/abstracts.htm#trust>. <https://doi.org/10.3366/epi.2011.0013>
- Anderson, Elizabeth. 2012a. "Epistemic Justice as a Virtue of Social Institutions." *Social Epistemology* 26(2): 163–173. The citation by "abstract" rests on the author's own abstract on her website, as captured by the Internet Archive on 25 September 2026, <https://web.archive.org/web/20260925171629/https://public.websites.umich.edu/~eandersn/abstracts.htm#Fricker>. <https://doi.org/10.1080/02691728.2011.652211>
- Anderson, Elizabeth. 2012b. "Reply to John Christman's Comments." *Social Epistemology Review and Reply Collective* 1(7): 15–16. Published online 12 June 2012; cited from the web version, without page numbers. <https://social-epistemology.com/2012/06/12/elizabeth-anderson-reply-to-john-christmans-comments/> Accessed 2026-10-30.
- Anderson, Elizabeth. 2023. "Dewey's Moral Philosophy." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/dewey-moral/> Accessed 2026-09-07.
- Anderson, Elizabeth. 2024. "Feminist Epistemology and Philosophy of Science." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/feminism-epistemology/> Accessed 2026-10-29.
- Anderson, Elizabeth. 2025. "Local Knowledge in Institutional Epistemology." *Australasian Philosophical Review*: 1–23. <https://doi.org/10.1080/24740500.2024.2422551>
- Angius, Nicola, Giuseppe Primiero, and Raymond Turner. 2025. "The Philosophy of Computer Science." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/computer-science/> Accessed 2026-10-05.
- Anthropic. 2026. "Learning More about Claude's Mathematical Capabilities." *Anthropic*. <https://www.anthropic.com/research/riemann-zeta> Accessed 2026-10-28.
- Arfini, Selene and Matteo Costa. 2025. "What kind of explanations can serendipity provide?" *European Journal for Philosophy of Science* 15(3): 42. <https://doi.org/10.1007/s13194-025-00670-4>
- ARIA (Advanced Research and Invention Agency). 2025. *AI Scientist Call: Call for Research Proposals for Working with AI Scientist Systems*. Advanced Research and Invention Agency. https://aria.org.uk/media/za3mj5h1/ai-scientist-solicitation-call-document_accessible.pdf
- Ariane 501 Inquiry Board. 1996. *ARIANE 5 Flight 501 Failure: Report by the Inquiry Board*. Paris: European Space Agency. Chaired by J. L. Lions; scanned report, cited by its page numbers. <https://esamultimedia.esa.int/docs/esa-x-1819eng.pdf> Accessed 2026-09-24.
- Ariew, Roger. 2026. "Pierre Duhem." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/duhem/> Accessed 2026-09-07.
- Ariga, Nobumichi. 2009. "書評 : Paul Humphreys, Extending ourselves: Computational science, empiricism, and scientific method." *科学哲学科学史研究* 3: 83–86. Book review of Humphreys 2004 (in Japanese), in the Kyoto University Research Information Repository. https://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/72806/1/phs_3_83.pdf
- Article 36. 2013. "Killer Robots: UK Government Policy on Fully Autonomous Weapons." Policy paper, April 2013. https://article36.org/wp-content/uploads/2020/12/Policy_Paper1.pdf
- Avila, Artur, Manjul Bhargava, Caucher Birkar, Pierre Deligne, Yu Deng, Simon Donaldson, Hugo Duminil-Copin, Alessio Figalli, Martin Hairer, June Huh, Maxim Kontsevich, Elon Lindenstrauss, Pierre-Louis Lions, James Maynard, Curtis McMullen, Shigefumi Mori, Báo Châu Ngô, Andrei Okounkov, Peter Scholze, Stanislav Smirnov, Terence Tao, Maryna Viazovska, Cédric Villani, Wendelin Werner, and Efim Zelmanov. 2026. "A Severe Misalignment of AI in Mathematics." Zenodo. <https://doi.org/10.5281/zenodo.22737750>
- Azoulay, Pierre, Joshua S. Graff Zivin, and Gustavo Manso. 2011. "Incentives and creativity: evidence from the academic life sciences." *The RAND Journal of Economics* 42(3): 527–554. Page numbers cited in the text refer to the authors' final manuscript, DSpace@MIT,

<https://hdl.handle.net/1721.1/64640>; the published pagination differs. <https://doi.org/10.1111/j.1756-2171.2011.00140.x>

Azoulay, Pierre, Christian Fons-Rosen, and Joshua S. Graff Zivin. 2019. "Does Science Advance One Funeral at a Time?" *American Economic Review* 109(8): 2889–2920. Section numbers cited in the text refer to the author manuscript in PubMed Central, <https://pmc.ncbi.nlm.nih.gov/articles/PMC6814193/>, which has no page numbers. <https://doi.org/10.1257/aer.20161574>

Baek, Jinheon, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. "ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models." In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6709–6738. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.naacl-long.342>

Baghrmian, Maria and J. Adam Carter. 2025. "Relativism." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/relativism/> Accessed 2026-10-29.

Baier, Annette C. 1992. "Trust." In *The Tanner Lectures on Human Values*, 107–174. University of Utah Press. Lectures delivered at Princeton University, 6–8 March 1991. Cited from the publisher's PDF formerly on the Tanner Lectures website, Internet Archive capture of 23 September 2021, https://web.archive.org/web/20210923064824/https://tannerlectures.utah.edu/_resources/documents/a-to-z/b/baier92.pdf, which carries the volume's pagination. https://web.archive.org/web/20210923064824/https://tannerlectures.utah.edu/_resources/documents/a-to-z/b/baier92.pdf Accessed 2026-10-29.

Bainbridge, Lisanne. 1983. "Ironies of Automation." *Automatica* 19(6): 775–779. Section numbers cited in the text are those of the author's online version, now available as an Internet Archive capture (24 December 2019), <https://web.archive.org/web/20191224095616/http://www.bainbrdg.demon.co.uk/Papers/Ironies.html>. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)

Baldwin, Guy. 2023. "Rawls and Animal Moral Personality." *Animals* 13(7): 1238. <https://doi.org/10.3390/ani13071238>

Baldwin, Melinda. 2015. "Credibility, peer review, and Nature, 1945–1990." *Notes and Records of the Royal Society of London* 69(3): 337–352. Cited without page numbers from the publisher's version in PubMed Central, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4528400/>. <https://doi.org/10.1098/rsnr.2015.0029>

Bao, Honglin, Siyang Wu, Xiao Liu, Sida Li, Shiyun Cao, and James A. Evans. 2026. "Contemporary AI lacks the imagination to diverge or negate in science." *arXiv*. Preprint, version 3. <https://arxiv.org/abs/2606.08251v3>

Barandiaran, Xabier E. 2008. "A Naturalized Approach to the Autonomy of Cognitive Agents." University of the Basque Country. PhD dissertation, posted by the author; page numbers are those printed in the dissertation. https://xabier.barandiaran.net/wp-content/uploads/2010/06/barandiaran_-2008_phd_thesis.pdf

Barandiaran, Xabier E., Ezequiel Di Paolo, and Marieke Rohde. 2009. "Defining Agency: Individuality, Normativity, Asymmetry, and Spatio-temporality in Action." *Adaptive Behavior* 17(5): 367–386. Section numbers cited in the text refer to the authors' version 1.0 (July 2009), PhilArchive, <https://philarchive.org/rec/BARDAI-6>. <https://doi.org/10.1177/1059712309343819>

Barimah, George Kwasi. 2024. "Second-Order Assessment of Scientific Expert Claims and Sharing Epistemic Burdens in Science Communication." *Episteme* 21(2): 461–477. Published online 26 May 2022; the year and page numbers are those of the issue (21(2), 2024), which the publisher's PDF carries. <https://doi.org/10.1017/epi.2022.11>

Barman, Kristian Gonzalez, Sascha Caron, Tom Claassen, and Henk de Regt. 2024. "Towards a Benchmark for Scientific Understanding in Humans and Machines." *Minds and Machines* 34(1): 6. <https://doi.org/10.1007/s11023-024-09657-1>

Barnes, Eric Christian. 2022. "Prediction versus Accommodation." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/prediction-accommodation/> Accessed 2026-10-08.

Barrett, Paul H., Peter J. Gautrey, Sandra Herbert, David Kohn, and Sydney Smith (eds.). 1987. *Charles Darwin's Notebooks, 1836–1844: Geology, Transmutation of Species, Metaphysical Enquiries*. British Museum (Natural History) and Cambridge University Press. https://darwin-online.org.uk/converted/pdf/1987_Darwins_notebooks_F1817.pdf Accessed 2026-09-29.

Barros Sobrinho, Marcelo. 2015. "Evidência oral, evidência escrita e conceito de ciência — estudo de casos." Universidade de São Paulo. Corrected doctoral thesis in Social History, deposited in the University of São Paulo digital library of theses; page numbers are those printed in the thesis. https://www.teses.usp.br/teses/disponiveis/8/8138/tde-03122015-141856/publico/2015_MarceloBarrosSobrinho_VCorri.pdf

Bassi, Yasmin. 2013. "Interventionism and the Exclusion Problem." University of Warwick. PhD thesis, Department of Philosophy, University of Warwick, in the Warwick Research Archive Portal; page numbers are those printed in the thesis. https://wrap.warwick.ac.uk/id/eprint/58570/1/WRAP_THESIS_Bassi_2013.pdf

Bastounis, Alexander, Fabian Circelli, and Anders C. Hansen. 2026. "Navier-Stokes lost in translation: Why Lean verification of AI autoformalisation does not guarantee correct natural language proofs." *arXiv*. Preprint. <https://arxiv.org/abs/2610.08144>

- Baum, Kevin, Maximilian Kiener, Markus Langer, and Johann Laux. 2026. "Anticipatory Human Oversight of Agentic AI: A Philosophical Account." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2609.24242v1>
- Bean, Andrew M., Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, Maria Grandury, Sophia Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrom, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott Hale, Deborah Raji, Christopher Summerfield, Philip Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. 2025. "Measuring what Matters: Construct Validity in Large Language Model Benchmarks." In *Advances in Neural Information Processing Systems* 38, 19868–19949. Neural Information Processing Systems Foundation, Inc. (NeurIPS). <https://doi.org/10.52202/085713-0590>
- Beddor, Bob. 2024. "Inquiry beyond knowledge." *Philosophy and Phenomenological Research* 109(1): 330–356. Section numbers cited in the text refer to the author's version, PhilArchive, <https://philarchive.org/rec/BEDIBK>; they are the same as in the published article. <https://doi.org/10.1111/phpr.13042>
- Beel, Joeran, Min-Yen Kan, and Moritz Baumgart. 2025. "Evaluating Sakana's AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality Towards 'Artificial Research Intelligence' (ARI)?" *arXiv*. Preprint, version 3. <https://arxiv.org/abs/2502.14297v3>
- Beijing Municipal Committee of the Communist Party of China, Leading Group for Education, Science and Technology, and Talent Work (中共北京市委教育科技人才工作领导小组). 2026. *北京市加快推进人工智能赋能科学研究实施方案 (2026—2028年)*. 国家科技传播中心 (转载). Implementation plan of Beijing for accelerating AI-enabled scientific research, 2026–2028; text read on the reposting page cited. https://www.ncsti.gov.cn/zcfg/zcwj/202607/t20260702_250676.html Accessed 2026-10-28.
- Beisbart, Claus. 2025. "Digging deeper with deep learning? Explanatory understanding and deep neural networks." *European Journal for Philosophy of Science* 15(3): 40. <https://doi.org/10.1007/s13194-025-00668-y>
- Bell Burnell, S. Jocelyn. 1977. "Little Green Men, White Dwarfs or Pulsars?" *Annals of the New York Academy of Sciences* 302(1): 685–689. Cited without page numbers from the reprint "reproduced by permission" in *Cosmic Search* 1(1) (1979), on the magazine's website, <https://bigear.org/CSMO/HTML/CS01/cs01p16.htm>. <https://doi.org/10.1111/j.1749-6632.1977.tb37085.x>
- Bell, S. Jocelyn. 1968. "The Measurement of Radio Source Diameters Using a Diffraction Method." *Apollo - University of Cambridge Repository*. University of Cambridge. <https://doi.org/10.17863/cam.4926>
- Bender, Emily M. and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.463>
- Bengio, Yoshua, Patrice Simard, and Paolo Frasconi. 1994. "Learning Long-Term Dependencies with Gradient Descent is Difficult." *IEEE Transactions on Neural Networks* 5(2): 157–166. Section numbers cited in the text refer to the authors' preprint, <https://www.iro.umontreal.ca/~lisa/pointeurs/ieeetrnn94.pdf>. <https://doi.org/10.1109/72.279181>
- Bengio, Yoshua, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King. 2025. "Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?" *arXiv*. Position-paper preprint, version 2. <https://arxiv.org/abs/2502.15657>
- Bergamaschi Ganapini, Marianna. 2025. "Can AI make scientific discoveries?" *Philosophical Studies*. Published online ahead of issue; section numbers cited in the text refer to the author's accepted manuscript (penultimate draft), which can be read at PhilArchive, <https://philarchive.org/rec/BERCAM-8>; the work is cited as the published article. <https://doi.org/10.1007/s11098-025-02299-8>
- Bianchi, Federico, Owen Queen, Nitya Thakkar, Eric Sun, and James Zou. 2025. "Exploring the use of AI authors and reviewers at Agents4Science." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2511.15534>
- Bianchini, Stefano, Valentina Di Girolamo, Julien Ravet, and David Arranz. 2026. "AI in science: When and where it makes a difference." *Research Policy* 55(6): 105478. <https://doi.org/10.1016/j.respol.2026.105478>
- BigScience. 2021. "tr8-104B-wide chronicles." *bigscience-workshop/bigscience*, *GitHub*. <https://github.com/bigscience-workshop/bigscience/blob/9e64edec2ad1fb9263bf39ba616a84ff64b4d947/train/tr8-104B-wide/chronicles.md> Accessed 2026-09-29.
- BigScience. 2022. "tr11-176B-ml chronicles." *bigscience-workshop/bigscience*, *GitHub*. <https://github.com/bigscience-workshop/bigscience/blob/9e64edec2ad1fb9263bf39ba616a84ff64b4d947/train/tr11-176B-ml/chronicles.md> Accessed 2026-09-29.
- BigScience Workshop. 2022. "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model." *arXiv*. Preprint. <https://arxiv.org/abs/2211.05100>
- Binz, Marcel, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T. Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D. West, Qiong Zhang, Richard M. Shiffrin, Samuel J. Gershman, Vencislav Popov, Emily M. Bender, Marco Marelli, Matthew M. Botvinick, Zeynep Akata, and Eric Schulz. 2025. "How should the advancement of large language models affect the practice of science?" *Proceedings of the National Academy of*

- Sciences* 122(5): e2401227121. Four perspectives with responses; the parts cited in the text are the perspective of Botvinick and Gershman, the response of Bender, Bergstrom, and West, and the response of Schulz, Schad, Binz, and colleagues. <https://doi.org/10.1073/pnas.2401227121>
- Bird, Alexander. 2004. "Kuhn and Philosophy of Science in the Twentieth Century." *Annals of the Japan Association for Philosophy of Science* 12(2): 61–74. https://www.jstage.jst.go.jp/article/jafpos1956/12/2/12_2_61/_article
- Bird, Alexander. 2019. "The aim of belief and the aim of science." *THEORIA. An International Journal for Theory, History and Foundations of Science* 34(2): 171. <https://doi.org/10.1387/theoria.19351>
- Bird, Alexander. 2024. "Naturalized knowledge-first and the epistemology of groups." *Philosophy and Phenomenological Research* 109(3): 856–873. <https://doi.org/10.1111/phpr.13082>
- Bird, Alexander. 2025a. "Objective scientific standards and the function of science." *Asian Journal of Philosophy* 4(1): 54. Cited without page numbers from the author's text, Apollo – University of Cambridge Repository, <https://www.repository.cam.ac.uk/handle/1810/397239>. <https://doi.org/10.1007/s44204-025-00273-z>
- Bird, Alexander. 2025b. "Thomas Kuhn." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/thomas-kuhn/> Accessed 2026-09-29.
- Bisht, Harshit, Vinay Kumar, Kevin Maik Jablonka, Mausam, and N. M. Anoop Krishnan. 2026. "Agentic AI Scientists Are Not Built For Autonomous Scientific Discovery." *arXiv*. Preprint, version 2, which is the version consulted; the arXiv record states acceptance to the NeurIPS 2026 Position Paper Track. <https://arxiv.org/abs/2605.08956v2>
- Biswas, Joydeep, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. 2026. "AI-Assisted Peer Review at Scale: The AAAI-26 AI Review Pilot." *arXiv*. Preprint. <https://arxiv.org/abs/2604.13940>
- Blount, Zachary D., Christina Z. Borland, and Richard E. Lenski. 2008. "Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*." *Proceedings of the National Academy of Sciences* 105(23): 7899–7906. <https://doi.org/10.1073/pnas.0803151105>
- Boden, Margaret A. 1993. *Autonomy and Artificiality*. Cognitive Science Research Paper 307. School of Cognitive and Computing Sciences, University of Sussex. Plain-text version of the report (November 1993) in the University of Sussex Cognitive Science Research Papers archive; page numbers are those marked in the text. <https://www.sussex.ac.uk/informatics/cogslib/reports/csrp/csrp307.txt> Accessed 2026-10-29.
- Boge, Florian J. and Paul Grünke. 2019. "Computer Simulations, Machine Learning and the Laplacean Demon: Opacity in the Case of High Energy Physics." PhilSci-Archive. Preprint of a chapter announced for *The Science and Art of Simulation II*, edited by Michael M. Resch, Andreas Kaminski, and Petra Gehring (Springer); page numbers cited in the text refer to this preprint. <https://philsci-archive.pitt.edu/17637/> Accessed 2026-10-17.
- Boge, Florian J. and Henk W. de Regt. 2025. "Machine Learning Discoveries and Scientific Understanding in Particle Physics: Problems and Prospects." In *Philosophy of Science for Machine Learning*, edited by Juan M. Durán and Giorgia Pozzi, 403–434. Springer. https://doi.org/10.1007/978-3-032-03083-2_19
- Boge, Florian J., Annika Schuster, and Frauke Stoll. 2026. "Disentangling scientific and scientists' understanding." *Synthese* 207(6): 243. <https://doi.org/10.1007/s11229-026-05627-5>
- Boge, Florian J. 2026. "The Epistemic Alignment Problem of Machine Learning." PhilSci-Archive. Preprint of a paper accepted for presentation at PSA 2026; page numbers cited in the text refer to this preprint. <https://philsci-archive.pitt.edu/29796> Accessed 2026-09-07.
- Boon, Mieke. 2020. "How Scientists Are Brought Back into Science—The Error of Empiricism." In *A Critical Reflection on Automated Science: Will Science Remain Human?*, edited by Marta Bertolaso and Fabio Sterpetti, 43–65. Springer. Page and note numbers cited in the text refer to the author's preprint of 9 May 2019, University of Twente repository, <https://ris.utwente.nl/ws/portalfiles/portal/177935074>; the published pagination differs. https://doi.org/10.1007/978-3-030-25001-0_4
- Borgdorff, Henk. 2011. "The Production of Knowledge in Artistic Research." In *The Routledge Companion to Research in the Arts*, edited by Michael Biggs and Henrik Karlsson, 44–63. Routledge. Cited from the chapter posted by the author on the Research Catalogue, <https://www.researchcatalogue.net/profile/show-work?work=129647>, in the published layout and pagination. <https://www.taylorfrancis.com/chapters/edit/10.4324/9780203841327-4/production-knowledge-artistic-research-henk-borgdorff> Accessed 2026-09-07.
- Bostrom, Nick. 2012. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents." *Minds and Machines* 22(2): 71–85. Page numbers cited in the text refer to the author's copy at <https://nickbostrom.com/superintelligentwill.pdf>. <https://doi.org/10.1007/s11023-012-9281-3>
- Bostrom, Nick. 2014a. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press. Chapter and note numbers cited in the text follow an electronic edition of the first edition without print pagination. <https://global.oup.com/academic/product/superintelligence-9780199678112>

- Bostrom, Nick. 2014b. "Hail Mary, Value Porosity, and Utility Diversification." Working paper posted on the author's website, dated 19 December 2014 (first version 12 December 2014); page numbers are those of the PDF. <https://nickbostrom.com/papers/porosity.pdf> Accessed 2026-10-29.
- Bottou, Léon. 2017. "Foreword." In Marvin L. Minsky and Seymour A. Papert, *Perceptrons: An Introduction to Computational Geometry*, reissue of the 1988 expanded edition, vii–xiii. MIT Press. Author-hosted copy of the foreword. <https://leon.bottou.org/publications/pdf/perceptrons-2017.pdf> Accessed 2026-10-14.
- Bouchacourt, Diane and Marco Baroni. 2018. "How agents see things: On visual representations in an emergent language game." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 981–985. Association for Computational Linguistics. <https://doi.org/10.18653/v1/d18-1119>
- Boudry, Maarten and Michael Vlerick. 2014. "Natural Selection Does Care about Truth." *International Studies in the Philosophy of Science* 28(1): 65–77. Cited from the authors' deposit of the article in the journal's layout, Ghent University Academic Bibliography, <https://biblio.ugent.be/publication/4374618>, whose pages correspond to the journal pages 65–77. <https://doi.org/10.1080/02698595.2014.915651>
- Bowman, Samuel R., Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile 'Lukošiute,' Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jack Clark, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noémi Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. 2022. "Measuring Progress on Scalable Oversight for Large Language Models." *arXiv*. Preprint, version 2. <https://arxiv.org/abs/2211.03540v2>
- Boyd, Kenneth. 2017. "Testifying Understanding." *Episteme* 14(1): 103–127. Page numbers cited in the text refer to the author's manuscript, PhilArchive, <https://philarchive.org/rec/BOYTU>; the published pagination differs. <https://doi.org/10.1017/epi.2015.53>
- Bradie, Michael and William Harms. 2020. "Evolutionary Epistemology." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/epistemology-evolutionary/> Accessed 2026-09-29.
- Bradley, Adam and Bradford Saad. 2025. "AI Alignment vs. AI Ethical Treatment: 10 Challenges." *arXiv*. Preprint; forthcoming in *Analytic Philosophy*. <https://arxiv.org/abs/2510.12844>
- Bradley, Jean-Claude. 2008. "How should Open Notebook Science be used?" *Chemistry Wide Open, Science 2.0*. https://www.science20.com/chemistry_wide_open/how_should_open_notebook_science_be_used Accessed 2026-09-30.
- Bromberger, Sylvain. 1992. *On What We Know We Don't Know: Explanation, Theory, Linguistics, and How Questions Shape Them*. University of Chicago Press; Center for the Study of Language and Information. The complete book is posted by CSLI Publications, its co-publisher and copyright holder, at <https://web.stanford.edu/group/cslipublications/cslipublications/bromberger-corpus/On-What-We-Know-We-Dont-Know.pdf> (Online Collected Works of Sylvain Bromberger); the PDF keeps the printed pagination, which the text cites. <https://web.stanford.edu/group/cslipublications/cslipublications/bromberger-corpus/On-What-We-Know-We-Dont-Know.pdf> Accessed 2026-10-29.
- Brown, Jessica. 2023. "Lackey on group justified belief and evidence." *Asian Journal of Philosophy* 2(2): 36. Article 36; page numbers cited in the text are those of the article (pp. 1–7). <https://doi.org/10.1007/s44204-023-00091-1>
- Brown, Mark B. 2013. "Philip Kitcher, Science in a Democratic Society." *Minerva* 51(3): 389–397. Cited from the published review as posted on the author's faculty page at California State University, Sacramento, https://www.csus.edu/faculty/b/brownm/_internal/_documents/brown-review-of-science-in-a-democratic-society-by-philip-kitcher.pdf, which carries the journal pagination. <https://doi.org/10.1007/s11024-013-9233-y>
- Brown, Matthew J. 2010. "Genuine Problems and the Significance of Science." *Contemporary Pragmatism* 7(2): 131–153. Cited from the PDF of the published article on the author's website, which carries the journal pagination. <https://www.matthewjbrown.net/professional/papers/significance-problems.pdf> Accessed 2026-10-29.
- Brown, Matthew J. 2015. "A Critical Appreciation of Ronald N. Giere's 'Distributed Cognition without Distributed Knowing'." *Social Epistemology Review and Reply Collective* 4(6): 45–51. Cited without page numbers from the web version on the journal's website, which has none. <https://social-epistemology.com/2015/06/22/a-critical-appreciation-of-ronald-n-gieres-distributed-cognition-without-distributed-knowing-matthew-j-brown> Accessed 2026-10-29.
- Buckner, Cameron. 2020. "Adversarial Examples and the Deeper Riddle of Induction: The Need for a Theory of Artifacts in Deep Learning." *arXiv*. Preprint (arXiv:2003.11917); page numbers cited in the text refer to this preprint. A revised version was published in *Nature Machine Intelligence* 2 (2020): 731–736. <https://arxiv.org/abs/2003.11917>
- Burda, Yuri, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. 2018. "Large-Scale Study of Curiosity-Driven Learning." *arXiv*. Preprint, version 1; later published at the International Conference on Learning Representations (2019); section numbers cited in the text refer to arXiv:1808.04355v1. <https://arxiv.org/abs/1808.04355v1>

- Bush, Vannevar. 1945. *Science, the Endless Frontier: A Report to the President*. United States Government Printing Office. https://www.govinfo.gov/content/pkg/GOVPUB-PR32_400-e7966ee70a4f7b47f862431c9776f727/pdf/GOVPUB-PR32_400-e7966ee70a4f7b47f862431c9776f727.pdf
- Buss, Sarah and Andrea Westlund. 2018. "Personal Autonomy." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/personal-autonomy/> Accessed 2026-10-08.
- Butterfield, Jeremy. 2019. "Lost in Math? A review of 'Lost in Math: How Beauty Leads Physics Astray', by Sabine Hossenfelder." *arXiv*. arXiv version of a review published in *Physics in Perspective* 21(1): 63–86 (2019), <https://doi.org/10.1007/s00016-019-00233-0>; section numbers cited in the text refer to the arXiv version. <https://arxiv.org/abs/1902.03480>
- Buzzard, Kevin. 2022. "Beyond the Liquid Tensor Experiment." *Xena Project*. <https://xenaproject.wordpress.com/2022/09/12/beyond-the-liquid-tensor-experiment> Accessed 2026-09-30.
- Büter, Anke. 2012. *Das Wertfreiheitsideal in der Sozialen Erkenntnistheorie: Objektivität, Pluralismus und das Beispiel Frauengesundheitsforschung*. Epistemische Studien 26.ontos verlag. Edition hosted by De Gruyter; page numbers are those printed. Cited for its quotation, on p. 84, of Helen E. Longino, *The Fate of Knowledge* (Princeton University Press, 2002), pp. 135–136. <https://www.degruyterbrill.com/document/doi/10.1515/9783110322729/html>
- Campbell, Donald T. 1997. "From Evolutionary Epistemology via Selection Theory to a Sociology of Scientific Validity." *Evolution and Cognition* 3(1): 5–38. Published after the author's death, edited from his manuscript by C. Heyes and B. Frankel; the text reproduces, and marks, passages from his earlier papers, including "A Tribal Model of the Social System Vehicle Carrying Scientific Knowledge" (1979). Cited from the issue PDF on the website of the Konrad Lorenz Institute, <https://kli.ac.at/webroot/files/file/Evolution%20%26%20Cognition/1997%203-1.pdf>, which carries the journal pagination. <https://kli.ac.at/webroot/files/file/Evolution%20%26%20Cognition/1997%203-1.pdf> Accessed 2026-10-29.
- Campbell, Donald T. [1976] 2011. "Assessing the Impact of Planned Social Change." *Journal of MultiDisciplinary Evaluation* 7(15): 3–43. <https://doi.org/10.56645/jmde.v7i15.297>
- Cappelen, Herman and Josh Dever. 2024. "AI with Alien Content and Alien Metasemantics." In *The Oxford Handbook of Applied Philosophy of Language*, edited by Ernie Lepore and Luvell Anderson. Oxford University Press. Section numbers cited in the text refer to the authors' preprint. <https://arxiv.org/abs/2405.19808>
- Carlsmith, Joseph. 2022. "Is Power-Seeking AI an Existential Risk?" *arXiv*. Preprint, version 2; the report is dated April 2021. <https://arxiv.org/abs/2206.13353v2>
- Carlsmith, Joseph. 2025. "How Human-Like Do Safe AI Motivations Need to Be?" *AI Alignment Forum*, 12 November 2025. Forum post in the series "How Do We Solve the Alignment Problem?", published under the name Joe Carlsmith; cited by section number. <https://www.alignmentforum.org/posts/r8H4d3X7AKg9n8TQk/how-human-like-do-safe-ai-motivations-need-to-be> Accessed 2026-10-17.
- Carnap, Rudolf. [1950] 1984. "Empiricism, Semantics, and Ontology." 2nd ed. In *Philosophy of Mathematics: Selected Readings*, edited by Paul Benacerraf and Hilary Putnam, 241–257. Cambridge University Press. First published in *Revue Internationale de Philosophie* 4 (1950): 20–40; page numbers cited in the text refer to the reprint. <https://doi.org/10.1017/cbo9781139171519.013>
- Carroll, Micah, Davis Foote, Anand Siththaranjan, Stuart Russell, and Anca Dragan. 2024. "AI Alignment with Changing and Influenceable Reward Functions." In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235: 5706–5756. PMLR. <https://proceedings.mlr.press/v235/carroll24a.html> Accessed 2026-09-07.
- Castellacci, Fulvio, Tommaso Ciarli, Yuan Gao, Marianna Marino, Giacomo Marzi, Massimo Riccaboni, Maria Savona, and Simone Vannuccini. 2026. "Generative Artificial Intelligence in Scientific Research: Individual Benefits, Collective Risks, and a Framework for Responsible Research with AI." *arXiv*. Preprint. <https://arxiv.org/abs/2607.24879>
- Cat, Jordi. 2024. "Otto Neurath." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2026/entries/neurath/> Accessed 2026-10-29.
- Cautaerts, Niels. 2026. "eLabFTW Export of elabftw.nielscautaerts.xyz Instance." Zenodo. Static export of the author's eLabFTW instance, deposited as a dataset. <https://doi.org/10.5281/zenodo.20127631>
- CERN. 2011. "OPERA experiment reports anomaly in flight time of neutrinos from CERN to Gran Sasso." Press release of 23 September 2011, with updates of 18 November 2011 and 23 February 2012. <https://home.cern/opera-experiment-reports-anomaly-in-flight-time-of-neutrinos-from-cern-to-gran-sasso/> Accessed 2026-09-24.
- Chalmers, David J. 2010. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17(9–10): 7–65. Section numbers cited in the text refer to the author's version. <https://consc.net/papers/singularity.pdf>
- Chan, Alan, Christoph Winter, Andrew Barto, Jakub Pachocki, Geoffrey Hinton, Eric Horvitz, Yoshua Bengio, Dawn Song, Jack Clark, Hilary Greaves, Anton Korinek, Samuel Hammond, Thore Graepel, Ben Bariach, Philip H. S. Torr, Sheila A. McIlraith, Jeff Clune, Sam Manning, Girish Sastry, Tom Davidson, Daniel Eth, and Sören Mindermann. 2026. "What if automating AI R&D triggers an intelligence explosion?" *arXiv*. Preprint. <https://arxiv.org/abs/2609.36054>

- Chandra, A. Ramesh, Aaghaz Mahajan, Ainesh Sanyal, Mianqi Wang, Shubhadeep Chakraborty, Sudip Chakraborty, Shailesh Chandrasekharan, Rama Govindarajan, Akshit Goyal, Jonathan J. Heckman, Yasir Iqbal, Divya Jaganathan, Saumav Kapoor, Andreas Karch, Pushkar Khandare, Sandeep Krishna, Anupam Kundu, Alok Laddha, Tongyan Lin, Swarnendu Maity, Nilmani Mathur, Shani Nadir Meynet, Alessandro Mininno, Abhijeet Mishra, Vidur Niranjan, Abhishodh Prakash, Pratik Rath, Durbar Roy, Debargha Sarkar, Srikanth Sastry, Gary Shiu, Aninda Sinha, Thomas Van Riet, Praneeth Netrapalli, and Suvrat Raju. 2026. “TheoPhysEval: Assessing AI agents at autonomous theoretical physics problem solving.” *arXiv*. Preprint. <https://arxiv.org/abs/2610.16733>
- Chang, Hasok. 2007. “Beyond Case-Studies: History as Philosophy.” PhilSci-Archive. Text of a presentation at the first Integrated History and Philosophy of Science conference (&HPS1), University of Pittsburgh, 13 October 2007, PhilSci-Archive 3926; section numbers cited in the text refer to this text. <https://philsci-archive.pitt.edu/3926/> Accessed 2026-10-29.
- Chang, Hasok. 2010. “The Hidden History of Phlogiston: How Philosophical Failure Can Generate Historiographical Refinement.” *HYLE—International Journal for Philosophy of Chemistry* 16(2): 47–79. Cited from the journal's open HTML edition by section. <https://www.hyle.org/journal/issues/16-2/chang.htm>
- Chang, Hasok. 2017a. “Who cares about the history of science?” *Notes and Records: The Royal Society Journal of the History of Science* 71(1): 91–107. Cited by section heading from the article, also in PubMed Central, <https://pmc.ncbi.nlm.nih.gov/articles/PMC5311899/>. <https://doi.org/10.1098/rsnr.2016.0042>
- Chang, Hasok. 2017b. “Operational Coherence as the Source of Truth.” *Proceedings of the Aristotelian Society* 117(2): 103–122. Cited from the author's accepted manuscript in the University of Cambridge repository, <https://www.repository.cam.ac.uk/handle/1810/263994>; page numbers refer to the manuscript. The accepted manuscript was consulted. <https://www.repository.cam.ac.uk/handle/1810/263994>
- Chang, Hasok. 2021. “Presentist History for Pluralist Science.” *Journal for General Philosophy of Science* 52: 97–114. <https://doi.org/10.1007/s10838-020-09512-8>
- Chang, Hasok. 2023. “Hasok Chang Transcript - S2 Ep 11.” *The HPS Podcast*. HPS@UniMelb, University of Melbourne. Interview by Samara Greenwood; transcript published 13 December 2023. Cited by the transcript's time stamps. <https://www.hpsunimelb.org/post/hasok-chang-transcript-s2-ep-11> Accessed 2026-10-30.
- Chang, Ruth. 2002. “The Possibility of Parity.” *Ethics* 112(4): 659–688. Cited from the scan of the published article on PhilArchive, <https://philarchive.org/rec/CHATPO-5>, which carries the journal pagination. <https://doi.org/10.1086/339673>
- Chang, Ruth. 2013. “Grounding practical normativity: going hybrid.” *Philosophical Studies* 164(1): 163–187. <https://doi.org/10.1007/s11098-013-0092-z>
- Chen, Hui, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. 2025. “MLR-Bench: Evaluating AI Agents on Open-Ended Machine Learning Research.” In *Advances in Neural Information Processing Systems* 38, 131093–131141. Neural Information Processing Systems Foundation, Inc. (NeurIPS). <https://doi.org/10.52202/085713-3945>
- Chen, Zihan, Di Zhu, and Lei Nico Zheng. 2026. “Auditing the Product, Not the Recursion: The Evaluation Evidence Behind Recursive Self-Improvement Claims.” *arXiv*. Preprint. <https://arxiv.org/abs/2610.27610>
- Chen, Zirui, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. 2025. “ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery.” *International Conference on Learning Representations* 2025: 96934. https://proceedings.iclr.cc/paper_files/paper/2025/hash/f12b4df26344f3be803c06b555252efe-Abstract-Conference.html Accessed 2026-09-07.
- Cheng, Zhoujun, Shibo Hao, Tianyang Liu, Fan Zhou, Yutao Xie, Feng Yao, Yuexin Bian, Yonghao Zhuang, Nilabjo Dey, Yuheng Zha, Yi Gu, Kun Zhou, Yuqi Wang, Yuan Li, Richard Fan, Jianshu She, Chengqian Gao, Abulhair Saparov, Haonan Li, Taylor W. Killian, Mikhail Yurochkin, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2025. “Revisiting Reinforcement Learning for LLM Reasoning from a Cross-Domain Perspective.” *arXiv*. <https://arxiv.org/abs/2506.14965v1>
- Cherniak, Christopher. 1981. “Minimal Rationality.” *Mind* 90(358): 161–183. Cited from the scan of the published article posted on the author's University of Maryland page, https://terpconnect.umd.edu/~cherniak/Mind_81.pdf; it carries the journal pagination. <https://doi.org/10.1093/mind/XC.358.161>
- Cherniak, Christopher. 1983. “Rationality and the Structure of Human Memory.” *Synthese* 57(2): 163–186. Cited from the scan of the published article posted on the author's University of Maryland page, https://terpconnect.umd.edu/~cherniak/Synth_83.pdf; it carries the journal pagination. <https://doi.org/10.1007/BF01064000>
- Chiang, Ted. 2000. “Catching crumbs from the table.” *Nature* 405(6786): 517. Short fiction in the journal's “Futures” series. Cited from the article page that Nature formerly served free of charge, Internet Archive capture of 10 August 2020, <https://web.archive.org/web/20200810233948/https://www.nature.com/articles/35014679>. <https://doi.org/10.1038/35014679>

- Chinn, Clark A. and William F. Brewer. 1993. "The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction." *Review of Educational Research* 63(1): 1–49. Page numbers cited in the text refer to the technical-report version, Center for the Study of Reading, Technical Report No. 583 (University of Illinois at Urbana-Champaign, September 1993), IDEALS, <https://www.ideals.illinois.edu/items/17657>; the journal pagination differs. <https://doi.org/10.3102/00346543063001001>
- Chirumuuta, M. 2021. "Prediction versus understanding in computationally enhanced neuroscience." *Synthese* 199(1–2): 767–790. Page numbers cited in the text refer to the peer-reviewed author's version, Edinburgh Research Explorer, <https://www.research.ed.ac.uk/en/publications/cfaf1281-07ba-4fe0-a357-bade5e001ea4>; the published pagination differs. <https://doi.org/10.1007/s11229-020-02713-0>
- ChoGlueck, Christopher and Elisabeth A. Lloyd. 2023. "Values as heuristics: a contextual empiricist account of assessing values scientifically." *Synthese* 201(6): 220. Article 220; cited by section number. <https://doi.org/10.1007/s11229-023-04198-z>
- Christiano, Paul. 2019. "What Failure Looks Like." *LessWrong*, 17 March 2019. Forum post. <https://www.lesswrong.com/posts/HBxe6wdjxK239zajf/what-failure-looks-like> Accessed 2026-10-17.
- Christiano, Tom and Sameer Bajaj. 2024. "Democracy." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/democracy/> Accessed 2026-09-29.
- Christiano, Tom and Stefan Sciaraffa. 2025. "Authority." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/authority/> Accessed 2026-09-29.
- Christman, John. 2012. "Comments on Elizabeth Anderson, 'Epistemic Justice as a Virtue of Social Institutions'." *Social Epistemology Review and Reply Collective*. Published online 16 May 2012; cited from the web version, without page numbers. <https://social-epistemology.com/2012/05/16/john-christman-comments-on-elizabeth-anderson-epistemic-justice-as-a-virtue-of-social-institutions/> Accessed 2026-10-30.
- Christman, John. 2025. "Autonomy in Moral and Political Philosophy." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/autonomy-moral/> Accessed 2026-09-07.
- Chroscinski, Denise, Darryl Sampey, and Nimet Maherali. 2015. "Registered report: Tumour vascularization via endothelial differentiation of glioblastoma stem-like cells." *eLife* 4: e04363. <https://doi.org/10.7554/elife.04363>
- Chubachi, Shigeru. 1984. "Preliminary result of ozone observations at Syowa Station from February 1982 to January 1983." *Memoirs of National Institute of Polar Research. Special issue* 34: 13–19. <https://nipr.repo.nii.ac.jp/records/1666> Accessed 2026-09-07.
- Chubachi, Shigeru (忠鉢繁). 1993. "私がオゾンホールを発見した頃 [When I Discovered the Ozone Hole]." *日本物理学会誌 [Butsuri]* 48(10): 815–818. https://www.jstage.jst.go.jp/article/butsuri1946/48/10/48_KJ00001498367/_article/-char/ja
- Ciardelli, Ivano. 2022. "Why We Need a Question Semantics." In *Asking and Answering: Rivalling Approaches to Interrogative Methods*, edited by Moritz Cordes, 15–47. Narr Francke Attempto. <https://doi.org/10.24053/9783823394808>
- Clark, Andy and David Chalmers. 1998. "The Extended Mind." *Analysis* 58(1): 7–19. Section numbers cited in the text refer to the authors' version on David Chalmers's website, <https://consc.net/papers/extended.html>, which has no page numbers. <https://doi.org/10.1093/analys/58.1.7>
- Clark, Elinor and Donal Khosrowi. 2022. "Decentring the discoverer: how AI helps us rethink scientific discovery." *Synthese* 200(6): 463. <https://doi.org/10.1007/s11229-022-03902-9>
- Clay Mathematics Institute. 2026. "Navier-Stokes Announcement." *Clay Mathematics Institute*. <https://www.claymath.org/news/navier-stokes-announcement> Accessed 2026-10-28.
- Cleland, Carol E. 2002. "Methodological and Epistemic Differences between Historical Science and Experimental Science." *Philosophy of Science* 69(3): 474–496. Cited from the copy of the published article that the author posted on her University of Colorado page (Internet Archive capture of 10 November 2011, <https://web.archive.org/web/20111110141414/http://spot.colorado.edu/~cleland/articles/Cleland.PS.Pdf>), which carries the journal pagination. <https://doi.org/10.1086/342455>
- Coelho Mollo, Dimitri and Raphaël Milliére. 2026. "The Vector Grounding Problem." *Philosophy and the Mind Sciences* 7(1). <https://doi.org/10.33735/phimisci.2026.12307>
- Cohen, Michael K., Elliot Catt, and Marcus Hutter. 2021. "Curiosity Killed or Incapacitated the Cat and the Asymptotically Optimal Agent." *IEEE Journal on Selected Areas in Information Theory* 2(2): 665–677. Section numbers cited in the text refer to arXiv:2006.03357v2, <https://arxiv.org/abs/2006.03357v2>. <https://doi.org/10.1109/jsait.2021.3079722>
- Colas, Cédric, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. 2022. "Autotelic Agents with Intrinsically Motivated Goal-Conditioned Reinforcement Learning: A Short Survey." *Journal of Artificial Intelligence Research* 74: 1159–1199. <https://doi.org/10.1613/jair.1.13554>
- Cole, David. 2024. "The Chinese Room Argument." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/chinese-room/> Accessed 2026-10-11.

- Collingwood, R. G. 1940. *An Essay on Metaphysics*. Clarendon Press. Digitized book copy. <https://archive.org/details/in.ernet.dli.2015.187414> Accessed 2026-10-03.
- Collins, H. M. 2025. "Why artificial intelligence needs sociology of knowledge: parts I and II." *AI & Society* 40(3): 1249–1263. <https://doi.org/10.1007/s00146-024-01954-8>
- Collins, H. M. and Robert Evans. 2002. "The Third Wave of Science Studies: Studies of Expertise and Experience." *Social Studies of Science* 32(2): 235–296. Cited without page numbers from the working paper of March 2002, ORCA, Cardiff University, <https://orca.cardiff.ac.uk/78065/>, a lightly edited version of which was published. <https://doi.org/10.1177/0306312702032002003>
- Collins, H.M. 2003. "Lead into gold: the science of finding nothing." *Studies in History and Philosophy of Science Part A* 34(4): 661–691. Cited from the abstract on the author's website, <https://sites.cardiff.ac.uk/harrycollins/paper-10/>. <https://doi.org/10.1016/j.shpsa.2003.09.002>
- Commelin, Johan. 2022. "Breen-Deligne packages and the liquid tensor experiment." *MathOverflow*. Author answer. <https://mathoverflow.net/questions/415469/breen-deligne-packages-and-the-liquid-tensor-experiment/415540> Accessed 2026-09-07.
- Commelin, Johan, Mateja Jamnik, Rodrigo Ochigame, Lenny Taelman, and Akshay Venkatesh. 2026. "Shaping the Future of Mathematics in the Age of AI." *arXiv*. Preprint, version 2. <https://arxiv.org/abs/2603.24914v2>
- Copeland, Samantha. 2019. "On serendipity in science: discovery at the intersection of chance and wisdom." *Synthese* 196(6): 2385–2406. Cited without page numbers from the accepted author manuscript, TU Delft Repository, <http://resolver.tudelft.nl/uuid:1fc35dd6-2d88-4c89-a0c1-22e42ec74f47>. <https://doi.org/10.1007/s11229-017-1544-3>
- Crasnow, Sharon. 2025. "Feminist Perspectives on Science." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/feminist-science/> Accessed 2026-10-30.
- Creel, Kathleen A. 2020. "Transparency in Complex Computational Systems." *Philosophy of Science* 87(4): 568–589. <https://doi.org/10.1086/709729>
- Critch, Andrew. 2021. "What Multipolar Failure Looks Like, and Robust Agent-Agnostic Processes (RAAPs)." *LessWrong*, 31 March 2021. Forum post. <https://www.lesswrong.com/posts/LpM3EAakwYdS6aRKf/what-multipolar-failure-looks-like-and-robust-agent-agnostic> Accessed 2026-10-17.
- Croarken, Mary. 2002. "Providing Longitude for All: The Eighteenth-Century Computers of the Nautical Almanac." *Journal for Maritime Research* 4(1): 106–126. <https://doi.org/10.1080/21533369.2002.9668324>
- Cronbach, Lee J. and Paul E. Meehl. 1955. "Construct validity in psychological tests." *Psychological Bulletin* 52(4): 281–302. Page numbers cited in the text refer to the reprint "with minor alterations" in *Minnesota Studies in the Philosophy of Science* 1 (1956): 174–204, University of Minnesota Digital Conservancy, <https://hdl.handle.net/11299/184279>. <https://doi.org/10.1037/h0040957>
- Cross, Charles and Floris Roelofsen. 2026. "Questions." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/questions/> Accessed 2026-10-04.
- Crupi, Vincenzo. 2025. "Confirmation." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/confirmation/> Accessed 2026-10-11.
- Cudd, Ann and Seena Eftekhari. 2025. "Contractarianism." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/contractarianism/> Accessed 2026-10-08.
- Curtis-Trudel, André, Darrell P. Rowbottom, and Tjonnje G. F. Li. 2025. "Computational science and unconceived alternatives: lessons from, and for, gravitational-wave astronomy." *Synthese* 206(6): 297. <https://doi.org/10.1007/s11229-025-05362-3>
- Cutts, Zachary, John Coller, Rachel Chen, Hua Tang, Mathias Bellaïche, Ryder Easterlin, Dina Zverinski, Petar Sirkovic, Avinatan Hassidim, Yossi Matias, Pooja Rao, and Annalisa Pawlosky. 2026. *Aptamers as a sandbox for lab-in-the-loop agentic molecular discovery*. Google Research. https://www.gstatic.com/science-ai/agentic_aptamer_sandbox.pdf Accessed 2026-10-27.
- Danaher, John. 2016. "The Threat of Algocracy: Reality, Resistance and Accommodation." *Philosophy & Technology* 29(3): 245–268. Cited without page numbers from the author's version, marked as forthcoming in the journal, PhilArchive, <https://philarchive.org/rec/DANTTO-13>. <https://doi.org/10.1007/s13347-015-0211-1>
- Danaher, John and Sven Nyholm. 2021. "Automation, work and the achievement gap." *AI and Ethics* 1: 227–237. <https://doi.org/10.1007/s43681-020-00028-x>
- Danchev, Valentin, Andrey Rzhetsky, and James A. Evans. 2019. "Centralized scientific communities are less likely to generate replicable results." *eLife* 8: e43094. <https://doi.org/10.7554/eLife.43094>
- Darwin, Charles. 1958. *The Autobiography of Charles Darwin, 1809–1882: With the Original Omissions Restored*. Collins. Edited by Nora Barlow. Page image hosted by Darwin Online. <https://darwin-online.org.uk/content/frameset?itemID=F1497&viewtype=image&pageseq=124> Accessed 2026-09-30.
- Dasgupta, Partha and Paul A. David. 1994. "Toward a New Economics of Science." *Research Policy* 23(5): 487–521. [https://doi.org/10.1016/0048-7333\(94\)01002-1](https://doi.org/10.1016/0048-7333(94)01002-1)

Daston, Lorraine and Peter Galison. [2007] 2010. *Objectivity*. Zone Books. Page numbers are those of the 2007 hardcover (ISBN 9781890951788). The citation of p. 17 in Chapter 11, Section 3 links the publisher's free excerpt of the German translation, *Objektivität*, translated by Christa Krüger (Suhrkamp, suhrkamp taschenbuch wissenschaft 2458, 2024; German first edition 2007), https://res.cloudinary.com/suhrkamp/images/q_auto/v1784939850/150935/objektivitaet_9783518300589_leseprobe.pdf, where the passage is on p. 17. <https://search.worldcat.org/isbn/9781890951795> Accessed 2026-09-07.

Davidovic, Jovana. 2026. "The End of Human Judgment in the Kill Chain? Relocating Initiative and Interpretation with Agentic AI." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2604.06300v1>

Dawid, Richard, Stephan Hartmann, and Jan Sprenger. 2012. "The No Alternatives Argument." PhilSci-Archive preprint dated February 14, 2012; section numbers cited in the text refer to this preprint. Published in *The British Journal for the Philosophy of Science* 66(1): 213–234 (2015). <https://philsci-archive.pitt.edu/9038/>

de Blanc, Peter. 2011. "Ontological Crises in Artificial Agents' Value Systems." *arXiv*. Preprint. <https://arxiv.org/abs/1105.3821>

De Filippi, Primavera, Morshed Mannan, and Wessel Reijers. 2020. "Blockchain as a confidence machine: The problem of trust & challenges of governance." *Technology in Society* 62: 101284. <https://doi.org/10.1016/j.techsoc.2020.101284>

De Filippi, Primavera, Morshed Mannan, and Wessel Reijers. 2022. "The alegality of blockchain technology." *Policy and Society* 41(3): 358–372. Page numbers cited in the text refer to the authors' manuscript, HAL hal-03513113, <https://hal.science/hal-03513113>. <https://doi.org/10.1093/polsoc/puac006>

de Regt, Henk W. and Dennis Dieks. 2005. "A Contextual Approach to Scientific Understanding." *Synthese* 144(1): 137–170. Section numbers cited in the text refer to the authors' revised manuscript (Word file "Synthese-revision"), PhilSci-Archive, <https://philsci-archive.pitt.edu/1354/>. <https://doi.org/10.1007/s11229-005-5000-4>

de Regt, Henk W. 2017. *Understanding Scientific Understanding*. Oxford University Press. <https://doi.org/10.1093/oso/9780190652913.001.0001>

de Regt, Henk W. 2018. "Review of Catherine Z. Elgin, True Enough." *Notre Dame Philosophical Reviews*: 2018.04.13. <https://ndpr.nd.edu/reviews/true-enough/> Accessed 2026-10-29.

de Regt, Henk W. 2019. "From Explanation to Understanding: Normativity Lost?" *Journal for General Philosophy of Science* 50: 327–343. <https://doi.org/10.1007/s10838-019-09477-3>

de Regt, Henk W. 2020. "Understanding, Values, and the Aims of Science." *Philosophy of Science* 87(5): 921–932. Cited from the publisher's version deposited in the Radboud Repository, <https://hdl.handle.net/2066/227612>, which carries the journal pagination. <https://doi.org/10.1086/710520>

DeDeo, Simon. 2024. "AlephZero and Mathematical Experience." *Bulletin of the American Mathematical Society* 61(3): 375–386. Section numbers cited in the text refer to arXiv:2307.05599v2. <https://arxiv.org/abs/2307.05599v2>

Deguchi, Yasuo. 2025a. "The WE-turn of Action: Principles." *Open Philosophy* 8(1): 20250092. A letter to the editors stating the principles of the WE-turn for a special issue on the author's work. Page numbers cited in the text are those of the article PDF (1–14). <https://doi.org/10.1515/opphil-2025-0092>

Deguchi, Yasuo. 2025b. "The WE-turn of Values: Principles." *Tetsugaku Kenkyu* (Kyoto Philosophical Society) 614: 1–43. https://doi.org/10.14989/JPS_614_1

Deguchi, Yasuo. 2026a. "Kyodo saishu ketteisha toshite no AI [AI as Co-Final-Decision-Maker]." *Kagaku* (Iwanami Shoten) 96(1), January 2026: 28–33. Page numbers cited in the text were read from the page images of the Kindle edition. <https://www.iwanami.co.jp/kagaku/> Accessed 2026-10-17.

Deguchi, Yasuo. 2026b. "AI jidai ni ningen wa fuyo ka? [Are Humans Unnecessary in the Age of AI?]." National Institute of Informatics. Lecture slides (in Japanese) for a public event, 24 April 2026, posted by the organizer; cited by slide number. <https://www.nii.ac.jp/event/upload/20260424-deguchi.pdf> Accessed 2026-10-29.

Dellsén, Finnur. 2016. "Scientific progress: Knowledge versus understanding." *Studies in History and Philosophy of Science Part A* 56: 72–83. Section numbers cited in the text refer to the author's penultimate draft, UCD Research Repository, <https://hdl.handle.net/10197/8173>; they are the same as in the published article. <https://doi.org/10.1016/j.shpsa.2016.01.003>

Dellsén, Finnur. 2017. "Understanding without Justification or Belief." *Ratio* 30(3): 239–254. Page numbers cited in the text refer to the author's penultimate draft, PhilArchive, <https://philarchive.org/rec/DELUWJ>; the published pagination differs. <https://doi.org/10.1111/rati.12134>

Dellsén, Finnur. 2018. "Scientific Progress: Four Accounts." *Philosophy Compass* 13(11): e12525. Section numbers cited in the text refer to the accepted manuscript, PhilSci-Archive 15529, <https://philsci-archive.pitt.edu/15529/>. <https://doi.org/10.1111/phc3.12525>

Dellsén, Finnur. 2020. "Beyond Explanation: Understanding as Dependency Modelling." *The British Journal for the Philosophy of Science* 71(4): 1261–1286. Section numbers cited in the text refer to the accepted manuscript, PhilSci-Archive 15528, <https://philsci-archive.pitt.edu/15528/>; they are the same as in the published article. <https://doi.org/10.1093/bjps/axy058>

- Dellsén, Finnur. 2021. "Understanding scientific progress: the noetic account." *Synthese* 199(3-4): 11249–11278. <https://doi.org/10.1007/s11229-021-03289-z>
- Dellsén, Finnur. 2024. "Expanding the Empirical Realm: Constructive Empiricism and Augmented Observation." In *Scientific Theories and Philosophical Stances: Themes from van Fraassen*, edited by Claus Beisbart and Michael Frauchiger, 127–146. De Gruyter. Page numbers cited in the text refer to the author's preprint (marked "To appear in M. Frauchiger (ed.), Themes from van Fraassen"), PhilSci-Archive 18686, <https://philsci-archive.pitt.edu/18686/>, not to the published chapter. <https://doi.org/10.1515/978311019802-009>
- Dellsén, Finnur. 2025. "Promoting scientific progress." *Philosophical Studies* 182(8): 2123–2143. <https://doi.org/10.1007/s11098-025-02335-7>
- Dennett, D. C. 1971. "Intentional Systems." *Journal of Philosophy* 68(4): 87–106. Cited from the scan of the published article in the Daniel C. Dennett Papers, Tufts Digital Library, <https://dl.tufts.edu/concern/pdfs/t435gq74n>, which carries the journal pagination. <https://doi.org/10.2307/2025382>
- DePaul, Michael R. and Stephen R. Grimm. 2007. "Review Essay on Jonathan Kvanvig's *The Value of Knowledge and the Pursuit of Understanding*." *Philosophy and Phenomenological Research* 74(2): 498–514. Cited from the published essay as posted on Stephen R. Grimm's faculty page at Fordham University, https://faculty.fordham.edu/sgrimm/Site/Papers_and_Presentations_files/Grimm-2007-PPR-Kvanvig%20Essay.pdf, which carries the journal pagination. <https://doi.org/10.1111/j.1933-1592.2007.00034.x>
- Dewey, John. 1910. *How We Think*. D. C. Heath. Page numbers cited in the text refer to the 1910 D. C. Heath edition, as marked in the Project Gutenberg transcription (eBook 37423), <https://www.gutenberg.org/ebooks/37423>. <https://www.gutenberg.org/ebooks/37423> Accessed 2026-10-29.
- Dewey, John. 1916. *Essays in Experimental Logic*. Chicago: University of Chicago Press. Cited from the Project Gutenberg edition (eBook 40794) of the third impression (1920), which marks the original page numbers. <https://www.gutenberg.org/ebooks/40794> Accessed 2026-10-29.
- Dewey, John. 1927. *The Public and Its Problems*. New York: Henry Holt. Page references are to the Alan Swallow (Denver) printing digitized at the archive.org item below. <https://archive.org/details/publicitsproblem00dewe> Accessed 2026-10-18.
- Di Liscia, Daniel A. 2025. "Johannes Kepler." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/kepler/> Accessed 2026-10-04.
- Di Paolo, Ezequiel A. 2005. "Autopoiesis, Adaptivity, Teleology, Agency." *Phenomenology and the Cognitive Sciences* 4(4): 429–452. Page numbers cited in the text refer to the author's manuscript (COGS, University of Sussex), posted on the author's Sussex home page (Internet Archive capture of 12 January 2011, https://web.archive.org/web/20110112055529/http://www.informatics.sussex.ac.uk:80/users/ezequiel/autopoiesis_teleology_2005.pdf); the published pagination differs. <https://doi.org/10.1007/s11097-005-9002-y>
- Di Paolo, Ezequiel A. and Hiroyuki Iizuka. 2008. "How (not) to model autonomous behaviour." *Biosystems* 91(2): 409–423. Page numbers cited in the text refer to the authors' preprint, Cognitive Science Research Paper CSRP 588 (University of Sussex, April 2007), linked from Ezequiel Di Paolo's Sussex home page (Internet Archive capture of 12 January 2011, <https://web.archive.org/web/20110112055550/http://www.informatics.sussex.ac.uk:80/users/ezequiel/csrp588.pdf>); the published pagination differs. <https://doi.org/10.1016/j.biosystems.2007.05.016>
- Dietrich, Franz and Kai Spiekermann. 2021. "Jury Theorems." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/jury-theorems/> Accessed 2026-10-08.
- Dimitrakos, Thodoris. 2023. "Do Kuhnians have to be anti-realists? Towards a realist reconception of Kuhn's historiography." *Synthese* 202: 21. <https://link.springer.com/article/10.1007/s11229-023-04225-z>
- Dixon, Dominic. 2025. "Social Epistemology as an Epistemological Foundation for Library and Information Science." *Social Epistemology*. Published online 12 November 2025; no volume or issue assigned. Section numbers cited in the text refer to the peer-reviewed accepted manuscript in Cambridge's Apollo repository. <https://www.repository.cam.ac.uk/items/724d4ac4-380c-4164-9785-aac82eb97c82>
- Doboszewski, Juliusz and Jamee Elder. 2024. "Robustness and the Event Horizon Telescope: the case of the first image of M87*." *Philosophy of Physics* 2(1): 3. <https://doi.org/10.31389/pop.74>
- Douglas, Heather. 2000. "Inductive Risk and Values in Science." *Philosophy of Science* 67(4): 559–579. Section numbers cited in the text refer to the author's preprint (received February 2000, revised July 2000), PhilSci-Archive 27688, <https://philsci-archive.pitt.edu/27688/>. <https://doi.org/10.1086/392855>
- Douglas, Heather. 2003. "The Moral Responsibilities of Scientists: Tensions between Autonomy and Responsibility." *American Philosophical Quarterly* 40: 59–68. Cited without page numbers from the author's preprint as submitted to the journal, PhilSci-Archive 27686, <https://philsci-archive.pitt.edu/27686/>. <https://philsci-archive.pitt.edu/27686/> Accessed 2026-10-29.
- Douglas, Heather. 2006. "Norms for Values in Scientific Belief Acceptance." PhilSci-Archive. Paper for the PSA 2006 contributed papers, deposited on PhilSci-Archive (3024) on 3 November 2006; page numbers are those of the posted text. <https://philsci-archive.pitt.edu/3024/> Accessed 2026-10-29.

- Douglas, Heather. 2007. "Rejecting the Ideal of Value-Free Science." PhilSci-Archive. Author's preprint, marked as forthcoming in *Value-Free Science: Ideal or Illusion*, edited by John Dupré, Harold Kincaid, and Alison Wylie (published by Oxford University Press, 2007), PhilSci-Archive 27681. Page numbers cited in the text are the preprint's own. <https://philsci-archive.pitt.edu/27681/> Accessed 2026-10-29.
- Douglas, Heather. 2008. "The Role of Values in Expert Reasoning." PhilSci-Archive. Author's preprint, marked as forthcoming in *Public Affairs Quarterly*, PhilSci-Archive 27679. Page numbers cited in the text are the preprint's own. <https://philsci-archive.pitt.edu/27679/> Accessed 2026-10-29.
- Douglas, Heather. 2016. "Values in Science." PhilSci-Archive. Author's preprint, marked as forthcoming in the *Oxford Handbook of Philosophy of Science* (edited by Paul Humphreys, Oxford University Press, 2016), PhilSci-Archive 27649. Page numbers cited in the text are the preprint's own. <https://philsci-archive.pitt.edu/27649/> Accessed 2026-10-29.
- Douglas, Heather. 2017. "Why Inductive Risk Requires Values in Science." PhilSci-Archive. Author's preprint of the essay published in *Current Controversies in Values and Science*, edited by Kevin C. Elliott and Daniel Steel (Routledge, 2017), PhilSci-Archive 27646. Page numbers cited in the text are the preprint's own. <https://philsci-archive.pitt.edu/27646/> Accessed 2026-10-29.
- Duede, Eamon. 2022. "Instruments, Agents, and Artificial Intelligence: Novel Epistemic Categories of Reliability." *Synthese* 200(6): 491. Section numbers cited in the text refer to the preprint, PhilSci-Archive 21352 (dated 6 November 2022), <https://philsci-archive.pitt.edu/21352/>. <https://doi.org/10.1007/s11229-022-03975-6>
- Duede, Eamon. 2023. "Deep Learning Opacity in Scientific Discovery." *Philosophy of Science* 90(5): 1089–1099. <https://doi.org/10.1017/psa.2023.8>
- Duede, Eamon and Kevin Davey. 2025. "A Priori Knowledge in an Era of Computational Opacity: The Role of Artificial Intelligence in Mathematical Discovery." *Philosophy of Science* 92(5): 1394–1404. <https://doi.org/10.1017/psa.2025.10160>
- Duede, Eamon and Daniel Friedman. 2025. "Epistemic Gaps and the Attribution of (AI) Discovery." PhilSci-Archive. Preprint. <https://philsci-archive.pitt.edu/27719> Accessed 2026-09-07.
- Duerr, Patrick M. and Enno Fischer. 2025. "Rationally warranted promise: the virtue-economic account of pursuit-worthiness." *Synthese* 206(2): 68. <https://doi.org/10.1007/s11229-025-05077-5>
- Dufner, Annette. 2025. "Personal Identity and Ethics." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/identity-ethics/> Accessed 2026-10-08.
- Duhem, Pierre. 1906. *La théorie physique: son objet et sa structure*. Chevalier & Rivière. Cited by part, chapter and section from the scan of the first edition digitized by the Internet Archive from a University of Toronto copy. The English translation of the second edition (1914), *The Aim and Structure of Physical Theory*, translated by Philip P. Wiener (Princeton University Press, 1954), has the same section numbering in Part II, chapter VI. <https://archive.org/details/lathoriephysiq00duhe> Accessed 2026-10-29.
- Dunbar, Kevin. 2000. "How Scientists Think in the Real World: Implications for Science Education." *Journal of Applied Developmental Psychology* 21(1): 49–58. [https://doi.org/10.1016/s0193-3973\(99\)00050-7](https://doi.org/10.1016/s0193-3973(99)00050-7)
- Dung, Leonard. 2024. "Is Superintelligence Necessarily Moral?" *Analysis* 84(4): 730–738. Cited from the author's preprint, PhilArchive, <https://philarchive.org/rec/DUNISN>; page numbers in the text refer to the preprint. <https://doi.org/10.1093/analys/anae033>
- Dung, Leonard. 2025. "Understanding Artificial Agency." *The Philosophical Quarterly* 75(2): 450–472. Section numbers cited in the text refer to the author's preprint, PhilArchive, <https://philarchive.org/rec/DUNUAA>. <https://doi.org/10.1093/pq/pqae010>
- Dunn, Jeff. 2021. "Review of Jennifer Lackey, *The Epistemology of Groups*." *Notre Dame Philosophical Reviews*: 2021.11.03. <https://ndpr.nd.edu/reviews/the-epistemology-of-groups/> Accessed 2026-10-29.
- Durán, Juan M. and Nico Formanek. 2018. "Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism." *Minds and Machines* 28(4): 645–666. Section and note numbers cited in the text refer to the preprint, arXiv:1904.01052v1. <https://doi.org/10.1007/s11023-018-9481-6>
- Durán, Juan M. 2025a. "Beyond Transparency: Computational Reliabilism as an Externalist Epistemology of Algorithms." In *Philosophy of Science for Machine Learning*, edited by Juan M. Durán and Giorgia Pozzi, 55–79. Springer. Section numbers cited in the text refer to the preprint, arXiv:2502.20402. https://doi.org/10.1007/978-3-032-03083-2_4
- Durán, Juan M. 2025b. "In Defense of Reliabilist Epistemology of Algorithms." *European Journal for Philosophy of Science* 15: 37. <https://link.springer.com/article/10.1007/s13194-025-00664-2>
- Dutilh Novaes, Catarina and Çağlar Dede. 2025. "Critical Contextual Empiricism for Busy People: Scientific Argumentation as Epistemic Exchange." *Topoi* 44: 733–747. <https://doi.org/10.1007/s11245-025-10198-0>
- Eigi, Jaana. 2012. "Two Millian Arguments: Using Helen Longino's Approach to Solve the Problems Philip Kitcher Targeted with His Argument on Freedom of Inquiry." *Studia Philosophica Estonica* 5(1): 44–63. <https://doi.org/10.12697/spe.2012.5.1.03>
- Eigi, Jaana. 2016. "The Social Organisation of Science as a Question for Philosophy of Science." University of Tartu. PhD thesis (version deposited for the defence), DSpace, University of Tartu. Read from the Internet Archive capture of 4 May 2026 of the repository's own file,

- <https://web.archive.org/web/20260504023055/https://dspace.ut.ee/server/api/core/bitstreams/f902e5f6-f02c-48aa-a1ab-529f51067d48/content>. Elgi cites Science, Truth, and Democracy from its 2003 printing. Also cited for its quotation, on p. 136, of Helen E. Longino, *The Fate of Knowledge* (Princeton University Press, 2002), pp. 133–135. <http://hdl.handle.net/10062/52179> Accessed 2026-10-30.
- El Mawas, Omar. 2021. “Probing ‘operational coherence’ in Hasok Chang’s pragmatic realism.” *European Journal for Philosophy of Science* 11(4): 108. <https://doi.org/10.1007/s13194-021-00425-x>
- El Shazly, Adham. 2025. “Inquiring to understand.” *The Philosophical Quarterly*: pqaf098. <https://doi.org/10.1093/pq/pqaf098>
- Elabbar, Ahmad. 2026. “Varying Evidential Standards as a Matter of Justice.” *The British Journal for the Philosophy of Science* 77(3): 643–669. Section numbers cited in the text refer to the accepted manuscript, Apollo – University of Cambridge Repository, <https://www.repository.cam.ac.uk/handle/1810/366372>. <https://doi.org/10.1086/727429>
- Elgin, Catherine Z. 2002. “Art in the Advancement of Understanding.” *American Philosophical Quarterly* 39(1): 1–12. Page numbers cited in the text refer to the author’s version on her website, <https://catherineelgin.com/undg/artcogapq.pdf>; the published pagination (1–12) differs. <https://catherineelgin.com/undg/artcogapq.pdf>
- Elgin, Catherine Z. 2004. “True Enough.” *Philosophical Issues* 14(1): 113–131. Page numbers cited in the text refer to the author’s version on her website, <https://catherineelgin.com/undg/trueno.pdf>; the published pagination (113–131) differs. <https://doi.org/10.1111/j.1533-6077.2004.00023.x>
- Elgin, Catherine Z. 2006. “From Knowledge to Understanding.” In *Epistemology Futures*, edited by Stephen Hetherington, 199–215. Clarendon Press. Page numbers cited in the text refer to the author’s version on her website, <https://catherineelgin.com/undg/kntoun.pdf>; the published pagination (199–215) differs. <https://doi.org/10.1093/oso/9780199273317.003.0012>
- Elgin, Catherine Z. 2007. “Understanding and the Facts.” *Philosophical Studies* 132(1): 33–42. Page numbers cited in the text refer to the author’s version on her website, <https://catherineelgin.com/undg/nonfactive.pdf>; the published pagination (33–42) differs. <https://doi.org/10.1007/s11098-006-9054-z>
- Elgin, Catherine Z. 2018. “Epistemic Virtues in Understanding.” In *The Routledge Handbook of Virtue Epistemology*, edited by Heather Battaly, 330–339. Routledge. Page numbers cited in the text refer to the author’s version on her website, https://catherineelgin.com/undg/epistemic_virtues_in_understanding.pdf; the published pagination (330–339) differs. https://catherineelgin.com/undg/epistemic_virtues_in_understanding.pdf Accessed 2026-10-29.
- Elgin, Catherine Z. 2022. “Models as Felicitous Falsehoods.” *Principia* 26(1): 7–23. <https://doi.org/10.5007/1808-1711.2022.e84576>
- Elgin, Catherine Z. 2025. *Epistemic Ecology*. MIT Press. <https://doi.org/10.7551/mitpress/15707.001.0001>
- Elgin, Catherine Z. n.d. “Ignorance, Error, and the Advancement of Understanding.” catherineelgin.com. Author’s English text, posted on her website, of a paper published in Spanish translation (by Remei Capdevila) as “Ignorancia, error y el avance de la comprensión,” in *Error y conocimiento*, edited by Victòria Campos, Anna Estany, and Mercè Izquierdo (Comares). The file shows no date; page numbers cited in the text are those of the PDF. <https://catherineelgin.com/undg/error.pdf> Accessed 2026-10-29.
- Elish, Madeleine Clare. 2019. “Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction.” *Engaging Science, Technology, and Society* 5: 40–60. <https://doi.org/10.17351/ests2019.260>
- Elliott, Kevin C. 2012. “Epistemic and methodological iteration in scientific research.” *Studies in History and Philosophy of Science Part A* 43(2): 376–382. Section numbers cited in the text refer to the author’s preprint, PhilSci-Archive 9041, <https://philsci-archive.pitt.edu/9041/>, which has no page numbers. <https://doi.org/10.1016/j.shpsa.2011.12.034>
- Elliott, Kevin C. 2013. “Douglas on Values: From Indirect Roles to Multiple Goals.” *Studies in History and Philosophy of Science Part A* 44. Section numbers cited in the text refer to the author’s preprint (unpaginated), headed as forthcoming in *Studies in History and Philosophy of Science*, PhilSci-Archive 9856, <https://philsci-archive.pitt.edu/9856/>. <https://philsci-archive.pitt.edu/9856/> Accessed 2026-10-29.
- Elliott, Kevin C. 2025. “Arguments against the Value-Free Ideal.” In *The Routledge Handbook of Values and Science*, edited by Kevin C. Elliott and Ted Richards. Routledge. Preprint of a forthcoming chapter, PhilSci-Archive 27172; page numbers cited in the text refer to this preprint. <https://philsci-archive.pitt.edu/27172/> Accessed 2026-10-29.
- Eltis, David. [2008] 2024. “Methodology.” *SlaveVoyages*. Essay first published 2008; reposted 14 March 2024. <https://legacy.slavevoyages.org/blog/methodology-trans-atlantic> Accessed 2026-09-26.
- Errington, Timothy M, Alexandria Denis, Nicole Perfito, Elizabeth Iorns, and Brian A Nosek. 2021a. “Challenges for assessing replicability in preclinical cancer biology.” *eLife* 10: e67995. <https://doi.org/10.7554/elife.67995>
- Errington, Timothy M, Alexandria Denis, Anne B Allison, Renee Araiza, Pedro Aza-Blanc, Lynette R Bower, Jessica Campos, Heidi Chu, Sarah Denson, Cristine Donham, Kaitlyn Harr, Babette Haven, Elizabeth Iorns, Jennie Kwok, Elysia McDonald, Steven Pelech, Nicole Perfito, Amanda Pike, Darryl Sampey, Michael Settles, David A Scott, Vidhu Sharma, Todd Tolentino, Angela Trinh, Rachel Tsui, Brandon Willis, Joshua Wood,

- and Lisa Young. 2021b. “Experiments from unfinished Registered Reports in the Reproducibility Project: Cancer Biology.” *eLife* 10: e73430. <https://doi.org/10.7554/elife.73430>
- European Southern Observatory. 2020. “Newly processed Venus ALMA data made available to the community.” www.eso.org. <https://www.eso.org/public/announcements/ann20030> Accessed 2026-09-29.
- Evans, Owain, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. “Truthful AI: Developing and Governing AI That Does Not Lie.” *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2110.06674v1>
- Event Horizon Telescope Collaboration. 2019. “First M87 Event Horizon Telescope Results. IV. Imaging the Central Supermassive Black Hole.” *The Astrophysical Journal Letters* 875(1): L4. <https://doi.org/10.3847/2041-8213/ab0e85>
- Event Horizon Telescope Collaboration. 2022. “Imaging Reanalyses of EHT Data.” *Event Horizon Telescope (blog)*. Blog statement, 14 June 2022; archived copy: <https://web.archive.org/web/20220618000549/https://eventhorizontelescope.org/blog/imaging-reanalyses-eh-t-data> <https://eventhorizontelescope.org/blog/imaging-reanalyses-eh-t-data> Accessed 2026-09-30.
- Fagan, Melinda. 2012. “Collective Scientific Knowledge.” *Philosophy Compass* 7(12): 821–831. The citation by “abstract” is to the article's abstract. <https://doi.org/10.1111/j.1747-9991.2012.00528.x>
- Fairhurst, Jordi, José Antonio Pérez-Escobar, and Deniz Sarikaya. 2026. “Against a global conception of mathematical hinges.” *The Philosophical Quarterly* 76(3): 900–922. <https://doi.org/10.1093/pq/pqae090>
- Falbo, Arianna. 2024. “Inquiry for the mistaken and confused.” *Philosophy and Phenomenological Research* 109(3): 962–985. <https://doi.org/10.1111/phpr.13089>
- Fazelpour, Sina, Joseph O'Brien, and Hannah Rubin. 2026. “Navigating Epistemic Monocultures in AI-Driven Science: A Simulation Study.” *arXiv*. Preprint; accepted at *Philosophy of Science*. <https://arxiv.org/abs/2608.19390>
- Fazi, M. Beatrice. 2021. “Beyond Human: Deep Learning, Explainability and Representation.” *Theory, Culture & Society* 38(7–8): 55–77. <https://doi.org/10.1177/0263276420966386>
- Feldman, Richard and Earl Conee. 1985. “Evidentialism.” *Philosophical Studies* 48(1): 15–34. <https://doi.org/10.1007/bf00372404>
- Feldman, Richard. 2005. “Review of William P. Alston, Beyond ‘Justification’: Dimensions of Epistemic Evaluation.” *Notre Dame Philosophical Reviews*: 2005.08.07. <https://ndpr.nd.edu/reviews/beyond-justification-dimensions-of-epistemic-evaluation/> Accessed 2026-10-29.
- Fellows, Richard and Anita Liu. 2015. *Research Methods for Construction*. 4th ed. Wiley-Blackwell. Cited from the chapter 1 excerpt the publisher serves free; page numbers are those printed in the book. The book quotes Kuhn, *The Structure of Scientific Revolutions*, p. 37. <https://catalogimages.wiley.com/images/db/pdf/9781118915745.excerpt.pdf>
- Feng, Tony, Trieu Trinh, Garrett Bingham, Jiwon Kang, Shengtong Zhang, Sang-hyun Kim, Kevin Barreto, Carl Schildkraut, Junehyuk Jung, Jaehyeon Seo, Carlo Pagano, Yuri Chervonyi, Dawsen Hwang, Kaiying Hou, Sergei Gukov, Cheng-Chiang Tsai, Hyunwoo Choi, Youngbeom Jin, Wei-Yuan Li, Hao-An Wu, Ruey-An Shiu, Yu-Sheng Shih, Quoc V. Le, and Thang Luong. 2026. “Semi-Autonomous Mathematics Discovery with Gemini: A Case Study on the Erdős Problems.” *arXiv*. Preprint, version 3; page numbers cited in the text refer to this version. Version 3 includes the post-publication addendum on Erdős-935; the counts of autonomous solutions differ between versions. <https://arxiv.org/abs/2601.22401v3>
- Ferrario, Andrea, Michele Loi, and Eleonora Viganò. 2020. “In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions.” *Philosophy & Technology* 33(3): 523–539. <https://doi.org/10.1007/s13347-019-00378-3>
- Festenstein, Matthew. 2023. “Dewey's Political Philosophy.” *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/dewey-political/> Accessed 2026-10-08.
- Feyerabend, Paul K. 1962. “Explanation, Reduction, and Empiricism.” In *Scientific Explanation, Space, and Time*, edited by Herbert Feigl and Grover Maxwell, 28–97. University of Minnesota Press. <https://hdl.handle.net/11299/184633>
- Firsching, Moritz, Paul Lezeau, Salvatore Mercuri, Miklós Z. Horváth, Yaël Dillies, Calle Sönne, Eric Wieser, Fred Zhang, Thomas Hubert, Blaise Agüera y Arcas, and Pushmeet Kohli. 2026. “Formal Conjectures: An Open and Evolving Benchmark for Verified Discovery in Mathematics.” *arXiv*. Preprint. <https://arxiv.org/abs/2605.13171>
- Fischer, Enno. 2023. “Naturalness and the Forward-Looking Justification of Scientific Principles.” *Philosophy of Science* 90(5): 1050–1059. <https://doi.org/10.1017/psa.2023.5>
- Fitts, Paul M. (ed.). 1951. *Human Engineering for an Effective Air-Navigation and Traffic-Control System*. Washington, DC: National Research Council. <https://apps.dtic.mil/sti/tr/pdf/ADB815893.pdf>
- Fleisher, Will. 2022. “Understanding, Idealization, and Explainable AI.” *Episteme* 19(4): 534–560. <https://doi.org/10.1017/epi.2022.39>
- Flexner, Abraham. 1939. “The Usefulness of Useless Knowledge.” *Harper's Magazine* 179: 544–552. <https://www.ias.edu/sites/default/files/library/UsefulnessHarpers.pdf>

- Flores, Carolina and Elise Woodard. 2023. "Epistemic norms on evidence-gathering." *Philosophical Studies* 180(9): 2547–2571. <https://doi.org/10.1007/s11098-023-01978-8>
- Frances, Bryan and Jonathan Matheson. 2024. "Disagreement." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/disagreement/> Accessed 2026-10-29.
- Franklin, Allan and Slobodan Perović. 2023. "Experiment in Physics." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/physics-experiment/> Accessed 2026-10-04.
- Fraser, Nancy. 2007. "Transnationalizing the Public Sphere: On the Legitimacy and Efficacy of Public Opinion in a Post-Westphalian World." *Theory, Culture & Society* 24(4): 7–30. Cited from the version published in *transversal texts* (eipcp), issue "publicum", March 2007, which has no page numbers. <https://transversal.at/transversal/0605/fraser/en> Accessed 2026-10-18.
- Freeman, Samuel. 2023. "Original Position." Summer 2026 archived edition. *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/sum2026/entries/original-position/> Accessed 2026-11-01.
- Fricker, Elizabeth. 2026. "On the metaphysical and epistemic contrasts between human and AI testimony." *Inquiry* 69(6): 2859–2884. <https://doi.org/10.1080/0020174x.2025.2553297>
- Fricker, Miranda. 2008a. "Précis." *Theoria: An International Journal for Theory, History and Foundations of Science* 23(1): 69–71. In the forum on Miranda Fricker's *Epistemic Injustice: Power and the Ethics of Knowing*. <https://doi.org/10.1387/theoria.7>
- Fricker, Miranda. 2008b. "Replies to Critics." *Theoria: An International Journal for Theory, History and Foundations of Science* 23(1): 81–86. <https://doi.org/10.1387/theoria.10>
- Friederich, Simon. 2024. "Symbiosis, not alignment, as the goal for liberal democracies in the transition to artificial general intelligence." *AI and Ethics* 4(2): 315–324. <https://doi.org/10.1007/s43681-023-00268-7>
- Friedman, Daniel C. 2024. "Intentions and Inquiry." *Mind*. Published online ahead of issue; cited without page numbers from the author's version, PhilArchive, <https://philarchive.org/rec/FRILAI-8>. The issue version is *Mind* 134(533): 85–106 (2025). <https://doi.org/10.1093/mind/fzae056>
- Friedman, Jane. 2013. "Question-Directed Attitudes." *Philosophical Perspectives* 27(1): 145–174. The author's version dated January 2014, <https://jfriedmanphilosophy.github.io/QDA.pdf>, was consulted; no page numbers are cited. <https://doi.org/10.1111/phpe.12026>
- Friedman, Jane. 2019. "Inquiry and Belief." *Noûs* 53(2): 296–315. Cited without page numbers from the author's version dated August 2017, on the author's website, <https://jfriedmanphilosophy.github.io/IB.pdf>. <https://doi.org/10.1111/nous.12222>
- Friedman, Jane. 2020. "The Epistemic and the Zetetic." *The Philosophical Review* 129(4): 501–536. Page numbers cited in the text refer to the author's version dated July 2020, on the author's website, <https://jfriedmanphilosophy.github.io/E&Z.pdf>; the published pagination differs. <https://doi.org/10.1215/00318108-8540918>
- Friedman, Jane. 2024. "The aim of inquiry?" *Philosophy and Phenomenological Research* 108(2): 506–523. Cited without page numbers from the author's version dated 12 May 2023, on the author's website, <https://jfriedmanphilosophy.github.io/AIMS.pdf>. <https://doi.org/10.1111/phpr.12982>
- Frigg, Roman and Stephan Hartmann. 2025. "Models in Science." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/models-science/> Accessed 2026-10-29.
- Frise, Matthew. 2023. "Epistemological Problems of Memory." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/memory-episprob/> Accessed 2026-10-08.
- Froese, Tom, Nathaniel Virgo, and Eduardo Izquierdo. 2007. "Autonomy: A Review and a Reappraisal." In *Advances in Artificial Life: 9th European Conference, ECAL 2007*, 455–464. Springer. Section numbers cited in the text refer to the manuscript version, Cognitive Science Research Paper 591, University of Sussex, <https://www.sussex.ac.uk/informatics/cogslib/reports/csrp/csrp591.pdf>. https://doi.org/10.1007/978-3-540-74913-4_46
- Funtowicz, Silvio O. and Jerome R. Ravetz. 1993. "Science for the Post-Normal Age." *Futures* 25(7): 739–755. Read in the authors' republication with a new foreword, Commonplace, 14 May 2020 (<https://doi.org/10.21428/6ffd8432.8a99dd09>); cited by section. <https://commonplace.knowledgefutures.org/pub/6qqfgms5/release/1>
- Gabriel, Iason. 2020. "Artificial Intelligence, Values, and Alignment." *Minds and Machines* 30(3): 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
- Gabriel, Iason and Geoff Keeling. 2025. "A Matter of Principle? AI Alignment as the Fair Treatment of Claims." *Philosophical Studies* 182(7): 1951–1973. <https://doi.org/10.1007/s11098-025-02300-4>
- Gabrielsen, Ane Møller. 2020. "Openness and trust in data-intensive science: the case of biocuration." *Medicine, Health Care and Philosophy* 23(3): 497–504. <https://doi.org/10.1007/s11019-020-09960-5>

- Gaillard, Maxence. 2013. "The Governance of "Well-Ordered Science", from Ideal Conversation to Public Debate." *Theoria: An International Journal for Theory, History and Foundations of Science* 28(2): 245–256. <https://doi.org/10.1387/theoria.6498>
- Galison, Peter. 1999. "Objectivity Is Romantic." In *The Humanities and the Sciences*. American Council of Learned Societies. Text of a lecture given at the ACLS Annual Meeting on 1 May 1999 and published in the ACLS Occasional Paper; cited by section heading from the text formerly on the ACLS website, Internet Archive capture of 1 January 2003, <https://web.archive.org/web/20030101232935/http://www.acls.org:80/op47-3.htm>. <https://web.archive.org/web/20030101232935/http://www.acls.org:80/op47-3.htm> Accessed 2026-10-29.
- Gaver, William. 2012. "What should we expect from research through design?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 937–946. ACM. Cited without page numbers from the author's accepted manuscript, Goldsmiths Research Online, <https://research.gold.ac.uk/id/eprint/14818/>. <https://doi.org/10.1145/2207676.2208538>
- Gelman, Andrew and Cosma Rohilla Shalizi. 2013. "Philosophy and the practice of Bayesian statistics." *British Journal of Mathematical and Statistical Psychology* 66(1): 8–38. Page and note numbers cited in the text refer to the authors' final preprint, arXiv:1006.3868v4 (28 June 2011), <https://arxiv.org/abs/1006.3868v4>; the published pagination differs. <https://doi.org/10.1111/j.2044-8317.2011.02037.x>
- Genin, Konstantin and Franz Huber. 2020. "Formal Representations of Belief." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/formal-belief/> Accessed 2026-10-04.
- Genus, Audley and Andy Stirling. 2018. "Collingridge and the dilemma of control: Towards responsible and accountable innovation." *Research Policy* 47(1): 61–69. <https://doi.org/10.1016/j.respol.2017.09.012>
- Genus, Audley and Marfuga Iskandarova. 2018. "Responsible innovation: its institutionalisation and a critique." *Technological Forecasting and Social Change* 128: 1–9. Page numbers cited in the text refer to the authors' accepted manuscript, Kingston University Research Repository, <https://eprints.kingston.ac.uk/id/eprint/39359/>; the published pagination differs. <https://doi.org/10.1016/j.techfore.2017.09.029>
- Ghins, Michel. 2003. "Thomas Kuhn on the existence of the world." *International Studies in the Philosophy of Science* 17(3): 265–279. Cited from the copy archived on PhilArchive; page numbers follow the published pagination. <https://philarchive.org/archive/GHITKO>
- Giere, Ronald N. 2007. "Distributed Cognition without Distributed Knowing." *Social Epistemology* 21(3): 313–320. Cited from the page proof posted on the author's University of Minnesota page (Internet Archive capture of 3 July 2012, <https://web.archive.org/web/20120703034005/http://www.tc.umn.edu:80/~giere/DCwoDK.pdf>), which carries the journal pagination. <https://doi.org/10.1080/02691720701674197>
- Gieryn, Thomas F. 1983. "Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists." *American Sociological Review* 48(6): 781–795. The citation by "abstract" is to the article's abstract. <https://www.jstor.org/stable/2095325> Accessed 2026-09-07.
- Gijsbers, Victor. 2007. "Why Unification Is Neither Necessary Nor Sufficient for Explanation." *Philosophy of Science* 74(4): 481–500. Page numbers cited in the text refer to chapter 2 of the author's doctoral thesis, *Explanation and Determination* (Leiden University, 2011), which the thesis describes as a slightly revised version of the article, Leiden University Scholarly Publications, <https://scholarlypublications.universiteitleiden.nl/download/1887/17879/7>; the published pagination differs. <https://doi.org/10.1086/524420>
- Gijsbers, Victor. 2011. "Explanation and Determination." Leiden University. PhD thesis (defended 28 September 2011), Leiden University Scholarly Publications. <https://hdl.handle.net/1887/17879> Accessed 2026-10-29.
- Gillies, Donald. 2014. "Selecting applications for funding: why random choice is better than peer review." *RT. A Journal on Research Policy and Evaluation* Vol 2: No 1 (2014). <https://doi.org/10.13130/2282-5398/3834>
- Gilmore, Sarah A., Michael W. Schelle, Cynthia M. Holsclaw, Clifton D. Leigh, Madhulika Jain, Jeffery S. Cox, Julie A. Leary, and Carolyn R. Bertozzi. 2012. "Sulfolipid-1 Biosynthesis Restricts Mycobacterium tuberculosis Growth in Human Macrophages." *ACS Chemical Biology* 7(5): 863–870. <https://doi.org/10.1021/cb200311s>
- Godin, Benoît and Désirée Schauz. 2016. "The changing identity of research: A cultural and conceptual history." *History of Science* 54(3): 276–306. Cited without page numbers from the accepted version, EspaceINRS, <https://espace.inrs.ca/id/eprint/6070/>. <https://doi.org/10.1177/0073275316656007>
- Goldberg, Sanford C. 2020. "Epistemically engineered environments." *Synthese* 197(7): 2783–2802. The citations by "abstract" and "n. 9" are to the article's abstract and notes. <https://doi.org/10.1007/s11229-017-1413-0>
- Goldman, Alvin. 2007. "Social Epistemology." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2007/entries/epistemology-social/> Accessed 2026-10-31.
- Goldman, Alvin and Thomas Blanchard. 2015. "Social Epistemology." *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/sum2015/entries/epistemology-social/>

- Goldman, Alvin and Bob Beddor. 2021. "Reliabilist Epistemology." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/reliabilism/> Accessed 2026-09-29.
- Good, I. J. 1965. "Speculations Concerning the First Ultrainelligent Machine." In *Advances in Computers*, vol. 6, edited by Franz L. Alt and Morris Rubinoff, 31–88. Academic Press. The page cited in the text (33) is in the scan of pp. 31–33 in VTechWorks, Virginia Tech, <https://vtechworks.lib.vt.edu/handle/10919/89424>. [https://doi.org/10.1016/S0065-2458\(08\)60418-0](https://doi.org/10.1016/S0065-2458(08)60418-0)
- Google DeepMind. 2025. "AlphaEvolve: A Gemini-Powered Coding Agent for Designing Advanced Algorithms." *Google DeepMind (blog)*. <https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms> Accessed 2026-10-28.
- Gopnik, Alison. 2020. "Childhood as a Solution to Explore–Exploit Tensions." *Philosophical Transactions of the Royal Society B: Biological Sciences* 375(1803): 20190502. <https://doi.org/10.1098/rstb.2019.0502>
- Gordon, Emma C. n.d. "Understanding in Epistemology." *Internet Encyclopedia of Philosophy*. The entry shows no date of publication. <https://iep.utm.edu/understa> Accessed 2026-10-29.
- Gottweis, Juraj, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2025. "Towards an AI co-scientist." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2502.18864v1>
- Gottweis, Juraj, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2026. "Accelerating scientific discovery with Co-Scientist." *Nature* 655(8122): 487–496. <https://doi.org/10.1038/s41586-026-10644-y>
- Graßhoff, Gerd and Michael May. 1995. "From Historical Case Studies to Systematic Methods of Discovery." In *Systematic Methods of Scientific Discovery: Papers from the 1995 AAAI Spring Symposium*, 46–57. AAAI Press. AAAI Technical Report SS-95-03. <https://aaai.org/papers/0010-ss95-03-010-from-historical-case-studies-to-systematic-methods-of-discovery/>
- Grasswick, Heidi. 2018. "Feminist Social Epistemology." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/feminist-social-epistemology/> Accessed 2026-10-29.
- Greaves, Jane S., Anita M. S. Richards, William Bains, Paul B. Rimmer, Hideo Sagawa, David L. Clements, Sara Seager, Janusz J. Petkowski, Clara Sousa-Silva, Sukrit Ranjan, Emily Drabek-Maunder, Helen J. Fraser, Annabel Cartwright, Ingo Mueller-Wodarg, Zhuchang Zhan, Per Friberg, Iain Coulson, E'lisa Lee, and Jim Hoge. 2020. "Phosphine gas in the cloud decks of Venus." *Nature Astronomy* 5(7): 655–664. Cited without page numbers from the submitted version, arXiv:2009.06593, <https://arxiv.org/abs/2009.06593>. <https://doi.org/10.1038/s41550-020-1174-4>
- Grimm, Stephen. 2021. "Understanding." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/understanding/> Accessed 2026-09-07.
- Grimm, Stephen R. 2008. "Epistemic Goals and Epistemic Values." *Philosophy and Phenomenological Research* 77(3): 725–744. <https://doi.org/10.1111/j.1933-1592.2008.00217.x>
- Grimm, Stephen R. 2012. "The Value of Understanding." *Philosophy Compass* 7(2): 103–117. The author's version was consulted. https://faculty.fordham.edu/sgrimm/site/Papers_and_Presentations_files/value%20of%20understanding-philosophy%20compass-8-14-11.pdf
- Grimm, Stephen R. 2017. "Understanding and Transparency." In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, edited by Stephen R. Grimm, Christoph Baumberger, and Sabine Ammon, 212–229. Routledge. Cited without page numbers; the claims cited are from the author's draft of 13 May 2015. <https://philpapers.org/rec/STEQAT-2>
- Grimm, Stephen R. 2019. "Transmitting Understanding and Know-How." In *What the Ancients Offer to Contemporary Epistemology*, edited by Stephen Hetherington and Nicholas D. Smith, 124–139. Routledge. Section numbers cited in the text refer to the author's draft of 28 March 2019, PhilArchive, <https://philarchive.org/rec/GRITUA>. <https://doi.org/10.4324/9780429344138-9>
- Grote, Thomas, Konstantin Genin, and Emily Sullivan. 2024. "Reliability in Machine Learning." *Philosophy Compass* 19(5): e12974. <https://doi.org/10.1111/phc3.12974>
- Gruber, Matthias J., Bernard D. Gelman, and Charan Ranganath. 2014. "States of Curiosity Modulate Hippocampus-Dependent Learning via the Dopaminergic Circuit." *Neuron* 84(2): 486–496. The accepted author manuscript (PubMed Central, PMC4252494) was consulted; it has no publisher page numbers. <https://doi.org/10.1016/j.neuron.2014.08.060>

- Guerrero, Alexander A. 2017. "Living with Ignorance in a World of Experts." In *Perspectives on Ignorance from Moral and Social Philosophy*, edited by Rik Peels, 156–185. Routledge. Page numbers cited in the text refer to the author's proofs, linked from his publication page; they follow the printed pagination. https://static1.squarespace.com/static/67a77b56f412ee35d35dd875/t/67a77bd1f3695e0f982b2adf/1585769117063/Proofs_Living_with_Ignorance.pdf
- Habgood-Coote, Joshua. 2022. "Group Inquiry." *Erkenntnis* 87(3): 1099–1123. <https://doi.org/10.1007/s10670-020-00232-5>
- Habgood-Coote, Joshua. 2026. "Asking a Question is a Proposal in Shared Inquiry." *Australasian Journal of Philosophy*: 1–19. Published online ahead of issue; page numbers cited in the text refer to the online version. <https://doi.org/10.1080/00048402.2026.2680296>
- Hacking, Ian. 1983. *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511814563>
- Hadfield-Menell, Dylan and Gillian K. Hadfield. 2019. "Incomplete Contracting and AI Alignment." In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 417–422. ACM. <https://doi.org/10.1145/3306618.3314250>
- Haely, Karen Cordrick. 2003. "Objectivity in the Feminist Philosophy of Science." The Ohio State University. PhD dissertation, OhioLINK ETD Center; page numbers are those printed. https://etd.ohiolink.edu/acprod/odb_etd/ws/send_file/send?accession=osu1064415629&disposition=inline Accessed 2026-10-31.
- Hagger, M. S., N. L. D. Chatzisarantis, H. Alberts, C. O. Anggono, C. Batailler, A. R. Birt, R. Brand, M. J. Brandt, G. Brewer, S. Bruyneel, D. P. Calvillo, W. K. Campbell, P. R. Cannon, M. Carlucci, N. P. Carruth, T. Cheung, A. Crowell, D. T. D. De Ridder, S. Dewitte, M. Elson, J. R. Evans, B. A. Fay, B. M. Fennis, A. Finley, Z. Francis, E. Heise, H. Hoemann, M. Inzlicht, S. L. Koole, L. Koppel, F. Kroese, F. Lange, K. Lau, B. P. Lynch, C. Martijn, H. Merckelbach, N. V. Mills, A. Michirev, A. Miyake, A. E. Mosser, M. Muise, D. Muller, M. Muzi, D. Nalis, R. Nurwanti, H. Otgaar, M. C. Philipp, P. Primoceri, K. Rentzsch, L. Ringos, C. Schlinkert, B. J. Schmeichel, S. F. Schoch, M. Schrama, A. Schütz, A. Stamos, G. Tinghög, J. Ullrich, M. vanDellen, S. Wimbarti, W. Wolff, C. Yusainy, O. Zerhouni, and M. Zwienenberg. 2016. "A Multilab Preregistered Replication of the Ego-Depletion Effect." *Perspectives on Psychological Science* 11(4): 546–573. <https://doi.org/10.1177/1745691616652873>
- Hakli, Raul and Pekka Mäkelä. 2016. "Robots, Autonomy, and Responsibility." In *What Social Robots Can and Should Do*, 145–154. IOS Press. Cited from the authors' copy of the published paper on PhilPapers, <https://philpapers.org/rec/HAKRAA>, which carries the proceedings pagination. <https://doi.org/10.3233/978-1-61499-708-5-145>
- Hamburg, Sarah. 2022. "A Guide to DeSci, the Latest Web3 Movement." *a16z crypto*, 9 February 2022. Web article. <https://a16zcrypto.com/posts/article/what-is-decentralized-science-aka-desci/> Accessed 2026-10-18.
- Hao, Qianyu, Fengli Xu, Yong Li, and James Evans. 2026. "Artificial intelligence tools expand scientists' impact but contract science's focus." *Nature* 649(8099): 1237–1243. Page numbers cited in the text refer to arXiv:2412.07727v4, headed "Just accepted by Nature," <https://arxiv.org/abs/2412.07727v4>. <https://doi.org/10.1038/s41586-025-09922-y>
- Haque, Albert, Michelle Guo, Alexandre Alahi, Serena Yeung, Zelun Luo, Alisha Rege, Jeffrey Jopling, Lance Downing, William Beninati, Amit Singh, Terry Platchek, Arnold Milstein, and Li Fei-Fei. 2017. "Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance." *arXiv*. Preprint. <https://arxiv.org/abs/1708.00163>
- Harding, Sandra. 2007. "Gender, Democracy, and Philosophy of Science." *RECIIS – Electronic Journal of Communication, Information and Innovation in Health* 1(1): 161–166. https://www.reciis.icict.fiocruz.br/index.php/reciis/article/download/891/pdf_20 Accessed 2026-10-29.
- Harding, Sandra. 2010. "Standpoint Methodologies and Epistemologies: A Logic of Scientific Inquiry for People." In *World Social Science Report 2010: Knowledge Divides*, 173–175. UNESCO and International Social Science Council. Cited from the copy of the report in UNESCO's web archive, <https://webarchive.unesco.org/web/20151214152723/http://unesdoc.unesco.org/images/0018/001883/188333e.pdf>, which carries the report's pagination. <https://webarchive.unesco.org/web/20151214152723/http://unesdoc.unesco.org/images/0018/001883/188333e.pdf> Accessed 2026-10-29.
- Harding, Sandra. 2015. "After Mr. Nowhere: What Kind of Proper Self for a Scientist?" *Feminist Philosophy Quarterly* 1(1): article 2. <https://doi.org/10.5206/fpq/2015.1.2>
- Hardwig, John. 1985. "Epistemic Dependence." *The Journal of Philosophy* 82(7): 335–349. Cited from the copy linked from the author's publications page at the University of Tennessee, a scan of the published article with a JSTOR cover page (Internet Archive capture of 11 February 2005, <https://web.archive.org/web/20050211011904/http://web.utk.edu:80/~jhardwig/EpDep.pdf>); it carries the journal pagination. <https://doi.org/10.2307/2026523>
- Hardwig, John. 1991. "The Role of Trust in Knowledge." *The Journal of Philosophy* 88(12): 693–708. Cited from the copy linked from the author's publications page at the University of Tennessee, a scan of the published article with a JSTOR cover page (Internet Archive capture of 15 June 2004, <https://web.archive.org/web/20040615221132/http://web.utk.edu:80/~jhardwig/RoleTrust.pdf>); it carries the journal pagination. <https://doi.org/10.2307/2027007>
- Harkness, Deborah E. 2009. "Finding the Story." *Perspectives on History*. American Historical Association. Researcher self-account. <https://www.historians.org/perspectives-article/finding-the-story-january-2009> Accessed 2026-09-07.

- Harnad, Stevan. 1990. "The symbol grounding problem." *Physica D: Nonlinear Phenomena* 42(1-3): 335–346. Cited from the scan of the published article in ePrints Soton, University of Southampton, <https://eprints.soton.ac.uk/250382/>, which carries the journal pagination. [https://doi.org/10.1016/0167-2789\(90\)90087-6](https://doi.org/10.1016/0167-2789(90)90087-6)
- Harris, Keith Raymond. 2024. "Scientific Progress and Collective Attitudes." *Episteme* 21(1): 127–146. <https://doi.org/10.1017/epi.2021.6>
- Hartmann, Stephan. 1999. "Models and Stories in Hadron Physics." In *Models as Mediators*, edited by Mary S. Morgan and Margaret Morrison, 326–346. Cambridge University Press. Page numbers cited in the text refer to the author's preprint, PhilSci-Archive 2433, <https://philsci-archive.pitt.edu/2433/>; a slightly revised version appeared in the volume, whose pagination differs. <https://doi.org/10.1017/CBO9780511660108.012>
- Hassabis, Demis and James M. Manyika. 2026. "AI as the Ultimate Tool for Science: A Conversation with Demis Hassabis." *Dædalus*. <https://www.amacad.org/publication/daedalus/ai-ultimate-tool-science-conversation-demis-hassabis> Accessed 2026-10-27.
- Hauswald, Rico. 2025. "Artificial Epistemic Authorities." *Social Epistemology* 39(6): 716–725. <https://doi.org/10.1080/02691728.2025.2449602>
- Havstad, Joyce C. 2020. "Forty Years after Laboratory Life." *Philosophy, Theory, and Practice in Biology* 12(2020112). <https://doi.org/10.3998/ptpbio.16039257.0012.003>
- Hayek, F. A. [1968] 2002. "Competition as a Discovery Procedure." Translated by Marcellus S. Snow. *Quarterly Journal of Austrian Economics* 5(3): 9–23. Translation of "Der Wettbewerb als Entdeckungsverfahren," a 1968 lecture published as Kieler Vorträge 56. https://cdn.mises.org/qjae5_3_3.pdf Accessed 2026-10-18.
- Hayek, F. A. 1945. "The Use of Knowledge in Society." *American Economic Review* 35(4): 519–530. Cited by section from the reprint on Econlib (Liberty Fund), <https://www.econlib.org/library/Essays/hykKw.html>, which states that it is reprinted with permission of the *American Economic Review*; the reprint has no page numbers. <https://www.jstor.org/stable/1809376> Accessed 2026-09-29.
- Haziza, Eliran. 2023. "Norms of Inquiry." *Philosophy Compass* 18(12): e12952. <https://doi.org/10.1111/phc3.12952>
- Henderson, Leah, Noah D. Goodman, Joshua B. Tenenbaum, and James F. Woodward. 2010. "The Structure and Dynamics of Scientific Theories: A Hierarchical Bayesian Perspective." *Philosophy of Science* 77(2): 172–200. Cited from the copy of the published article on the website of co-author Noah Goodman's laboratory, <https://cocolab.stanford.edu/papers/HendersonEtAl2010-PhilosSci.pdf>, which carries the journal pagination. <https://doi.org/10.1086/651319>
- Henderson, Leah. 2022. "The Problem of Induction." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/induction-problem/> Accessed 2026-09-07.
- Henne, Céline. 2022. "Framed and Framing Inquiry: Development and Defence of John Dewey's Theory of Knowledge." University of Cambridge. PhD thesis, University of Cambridge, deposited in the Apollo repository; page numbers are those printed in the thesis. <https://doi.org/10.17863/CAM.87164>
- Henne, Céline. 2023. "Framed and framing inquiry: a pragmatist proposal." *Synthese* 201(2): 60. Section numbers cited in the text refer to the author's manuscript, PhilPapers, <https://philpapers.org/rec/HENFAF-3>; they are the same as in the published article. <https://doi.org/10.1007/s11229-023-04059-9>
- Herndon, Thomas, Michael Ash, and Robert Pollin. 2013. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." PERI Working Paper 322, Political Economy Research Institute, University of Massachusetts Amherst. <https://peri.umass.edu/wp-content/uploads/joomla/images/WP322.pdf> Accessed 2026-09-24.
- Herzog, Lisa. 2020. "The epistemic division of labour in markets: knowledge, global trade and the preconditions of morally responsible agency." *Economics and Philosophy* 36(2): 266–286. Cited from the publisher's version in the University of Groningen research portal, <https://hdl.handle.net/11370/a44e0296-74b9-4a19-8df5-adc369f6424e>, which carries the journal pagination. <https://doi.org/10.1017/s0266267119000130>
- Hessels, Laurens K., Thomas Franssen, Wout Scholten, and Sarah de Rijcke. 2019. "Variation in Valuation: How Research Groups Accumulate Credibility in Four Epistemic Cultures." *Minerva* 57(2): 127–149. Cited from the publisher's PDF deposited in the Leiden University Scholarly Publications repository, <https://hdl.handle.net/1887/77958>, which carries the journal pagination. <https://doi.org/10.1007/s11024-018-09366-x>
- Hilbert, David. 1902. "Mathematical Problems." *Bulletin of the American Mathematical Society* 8(10): 437–479. Translated by Mary Winston Newson. <https://doi.org/10.1090/s0002-9904-1902-00923-3>
- Hildebrand, David. 2024. "John Dewey." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/dewey/> Accessed 2026-10-29.
- Hill, Ryan R. and Carolyn Stein. 2026. "How Artificial Intelligence Shapes Science: Evidence from AlphaFold." NBER Working Paper 35143. NBER. Figures and pages cited in the text are from the revision dated 17 July 2026 (first version 11 March 2026). <https://www.nber.org/papers/w35143>

- Hills, Alison. 2020. "Moral Testimony: Transmission Versus Propagation." *Philosophy and Phenomenological Research* 101(2): 399–414. Page numbers cited in the text refer to the author manuscript, Oxford University Research Archive, <https://ora.ox.ac.uk/objects/uuid:63e396fc-c058-4822-b51f-a0d78f2ef167>; the published pagination differs. <https://doi.org/10.1111/phpr.12595>
- Hills, Alison. 2022. "Aesthetic Testimony, Understanding and Virtue." *Noûs* 56: 21–39. <https://doi.org/10.1111/nous.12344>
- Hinton, Geoffrey. 2024. "AICS CS History Interview 005: Geoffrey Hinton." Interview by Vassilis Galanos, 14 June 2024. University of Edinburgh, AI and Computer Science history project. Edited transcript; cited by page. https://groups.inf.ed.ac.uk/aics_history/AICSCSHistory_Interview005_GeoffreyHinton_publish.pdf Accessed 2026-09-29.
- Hirosige, Tetu (広重徹). 1964. "10a-F-6 特殊相対論の起原 [The Origin of Special Relativity]." In *日本物理学会年会講演予稿集 [Proceedings of the Annual Meeting of the Physical Society of Japan]* 19(1), 107. The Physical Society of Japan. Conference abstract. https://doi.org/10.11316/jpsgaiyoi.19.1.0_107
- Hochreiter, Sepp and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 9(8): 1735–1780. Cited without page numbers from the authors' version on the website of the Institute for Machine Learning, Johannes Kepler University Linz, <https://www.bioinf.jku.at/publications/older/2604.pdf>. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Horvitz, Eric. 1999. "Principles of Mixed-Initiative User Interfaces." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*, 159–166. ACM. Cited without page numbers from the author's copy on his website, <http://erichorvitz.com/chi99horvitz.pdf>. <https://doi.org/10.1145/302979.303030>
- Hossenfelder, Sabine. 2018a. *Lost in Math: How Beauty Leads Physics Astray*. Basic Books. Page numbers cited in the text refer to the first-edition hardcover (ISBN 978-0-465-09425-7). <https://www.hachettebookgroup.com/titles/sabine-hossenfelder/lost-in-math/9780465094264> Accessed 2026-09-30.
- Hossenfelder, Sabine. 2018b. "Screams for Explanation: Finetuning and Naturalness in the Foundations of Physics." *arXiv*. Section numbers cited in the text refer to arXiv:1801.02176v2 (10 January 2018); a version was later published in *Synthese*. <https://arxiv.org/abs/1801.02176>
- Hossenfelder, Sabine. 2018c. "A Theory with No Strings Attached: Can Beautiful Physics Be Wrong? [Excerpt]." *Scientific American*. Adapted excerpt from Hossenfelder 2018a, published online 11 June 2018; cited without page numbers. <https://www.scientificamerican.com/article/a-theory-with-no-strings-attached-can-beautiful-physics-be-wrong-excerpt/> Accessed 2026-10-29.
- Hossenfelder, Sabine. 2018d. "The Alluring Beauty of Supersymmetry." *Science Friday*. Excerpt from Hossenfelder 2018a, published online 2 August 2018; cited without page numbers. <https://www.sciencefriday.com/articles/the-alluring-beauty-of-supersymmetry/> Accessed 2026-10-29.
- Howells-Whitaker, Ned and Seth Lazar. 2026. "Artificial Persons." *arXiv*. The preprint was consulted. <https://arxiv.org/abs/2607.08695> Accessed 2026-11-01.
- Hoyningen-Huene, Paul. 1989. "The Concept of Nature, the Epistemic Ideal, and Experiment: Why Is Modern Science Technologically Exploitable?" Translated by Alexander Levine. *Zeitschrift für Wissenschaftsforschung* 5: 43–55. The English translation by Alexander Levine, deposited by the author on PhilSci-Archive, was consulted; page numbers are those of the translation. <https://philsci-archive.pitt.edu/21766/> Accessed 2026-11-01.
- Hoyningen-Huene, Paul. 2018. "Replies." PhilSci-Archive. Manuscript version of 17 February 2018 (revised submission SYNT-D-17-01020R1) of the author's replies in the *Synthese* special issue "Systematicity: The Nature of Science," PhilSci-Archive 14398; page numbers cited in the text are the manuscript's own, and <https://philsci-archive.pitt.edu/14398/> Accessed 2026-10-29.
- Hoyningen-Huene, Paul. 2020. "The Heart of Science: Systematicity." In *Competing Knowledges – Wissen im Widerstreit*, edited by A. M. Horatschek, 85–102. De Gruyter. <https://doi.org/10.1515/9783110659658-005>
- Hu, Botao Amber, Yuhan Liu, and Helena Rong. 2025. "Trustless Autonomy: Understanding Motivations, Benefits and Governance Dilemmas in Self-Sovereign Decentralized AI Agents." *arXiv*. Preprint, version 2. <https://arxiv.org/abs/2505.09757v2>
- Hu, Xingming. 2021. "The Epistemic Value of Understanding-why." *Episteme*. Cited from the publisher's version posted by the author on PhilArchive, <https://philarchive.org/rec/HUTEV>; page numbers are those of the online version. <https://doi.org/10.1017/epi.2021.4>
- Huang, Chih-Tung. 2012. "Shaping Environmental Justice: Applying Science, Technology and Society Boundary Work." *Electronic Green Journal* 1(34). The article has no journal pagination; page numbers cited in the text refer to the PDF on eScholarship. <https://doi.org/10.5070/G313410934>
- Huebner, Bryce, Rebecca Kukla, and Eric Winsberg. 2017. "Making an Author in Radically Collaborative Research." In *Scientific Collaboration and Collective Knowledge*, edited by Thomas Boyer-Kassem, Conor Mayo-Wilson, and Michael Weisberg. Oxford University Press. Section numbers cited in the text refer to the authors' revised final manuscript, linked from Eric Winsberg's website (Internet Archive capture of 30 November 2020, https://web.archive.org/web/20201130151851/https://www.winsberg.net/uploads/7/4/4/3/74439291/hkw_-_authorship_revised_final-1.pdf). <https://doi.org/10.1093/oso/9780190680534.003.0005>

- Hugging Face. 2024. "Fixing Gradient Accumulation." *Hugging Face Blog*, 16 October 2024. https://huggingface.co/blog/gradient_accumulation Accessed 2026-10-14.
- Hugging Face Transformers. 2021. "Gradient accumulation causing different training curves (issue #14638)." *huggingface/transformers*, *GitHub*. <https://github.com/huggingface/transformers/issues/14638> Accessed 2026-10-14.
- Hugging Face TRL. 2024. "Gradient accumulation yields worse results than the equivalent batch size (issue #2175)." *huggingface/trl*, *GitHub*. <https://github.com/huggingface/trl/issues/2175> Accessed 2026-10-14.
- Hughes, Edward, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktäschel. 2024. "Open-Endedness is Essential for Artificial Superhuman Intelligence." *arXiv*. Preprint, version 1; page numbers cited in the text refer to this version. <https://arxiv.org/abs/2406.04268v1>
- Huh, Minyoung, Brian Cheung, Tongzhou Wang, and Phillip Isola. 2024. "The Platonic Representation Hypothesis." In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235. Page numbers cited in the text refer to arXiv:2405.07987v5. <https://arxiv.org/abs/2405.07987v5>
- Human Rights Watch and International Human Rights Clinic, Harvard Law School. 2012. *Losing Humanity: The Case against Killer Robots*. Human Rights Watch. <https://www.hrw.org/report/2012/11/19/losing-humanity/case-against-killer-robots>
- Hume, David. 1751. *An Enquiry concerning the Principles of Morals*. Hume Texts Online. <https://davidhume.org/texts/m/3> Accessed 2026-10-07.
- Humphreys, Paul. 1995. "Computational Science and Scientific Method." *Minds and Machines* 5: 499–512. Only the abstract is cited. <https://link.springer.com/article/10.1007/BF00974980>
- Humphreys, Paul. 2002. "Computational Models." *Philosophy of Science* 69(S3): S1–S11. Only the abstract is cited. <https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/computational-models/3937A1B08890C8163E268C750E69D00B>
- Humphreys, Paul. 2006. "Twenty-First Century Epistemology." University of Virginia (author's page). Author's English text (© 2006 Paul W. Humphreys), posted on his University of Virginia page and captured by the Wayback Machine on 17 January 2015; the file's header gives it as to appear in Spanish in *Anthropos* (2007), which has not been verified. The file has no page numbers; a paragraph is identified by its opening words. <https://web.archive.org/web/20150117101744/http://people.virginia.edu:80/~pwh2a/anthropos%20final.doc> Accessed 2026-10-29.
- Humphreys, Paul. 2009a. "The Philosophical Novelty of Computer Simulation Methods." *Synthese* 169: 615–626. Cited by section from the author-posted PDF on his University of Virginia page. https://uva.theopenscholar.com/files/paul-humphreys/files/the_philosophical_novelty_8.pdf
- Humphreys, Paul. 2009b. "Computational Economics." In *The Oxford Handbook of Philosophy of Economics*, edited by Harold Kincaid and Don Ross, 371–385. Oxford University Press. Section numbers cited in the text refer to the author's final version (Word file), posted on his University of Virginia page and now available as an Internet Archive capture (17 January 2015), <https://web.archive.org/web/20150117101759/http://people.virginia.edu:80/~pwh2a/handbook%20of%20economics%20article%20final%20version.doc> it has no page numbers. <https://doi.org/10.1093/oxfordhb/9780195189254.003.0013>
- Humphreys, Paul. 2010. "Conceptual Sea Changes." *Spontaneous Generations: A Journal for the History and Philosophy of Science* 4(1): 111–115. <https://doi.org/10.4245/sponge.v4i1.14185>
- Humphreys, Paul. 2019. "Knowledge Transfer across Scientific Disciplines." *Studies in History and Philosophy of Science Part A* 77: 112–119. Page numbers cited in the text refer to the author-posted advance version on his University of Virginia page. https://uva.theopenscholar.com/files/paul-humphreys/files/knowledge_transfer_across_8.pdf
- Humphreys, Paul. 2020. "Why Automated Science Should Be Cautiously Welcomed." In *A Critical Reflection on Automated Science: Will Science Remain Human?*, edited by Marta Bertolaso and Fabio Sterpetti, 11–26. Springer. https://doi.org/10.1007/978-3-030-25001-0_2
- Humphreys, Paul. n.d. "What Are Data About?" University of Virginia (author's page). Final revised version, as named in the file's address, posted on the author's University of Virginia page and captured by the Wayback Machine on 28 November 2016. The file shows no date or publication details; page numbers are those of the PDF. <https://web.archive.org/web/20161128140336/http://people.virginia.edu/~pwh2a/stuttgart%20volume%20paper%20humphreys%20final%20revised.pdf> Accessed 2026-10-29.
- Huq, Aziz Z. 2020. "A Right to a Human Decision." *Virginia Law Review* 106(3): 611–688. <https://virginialawreview.org/articles/right-human-decision/>
- Hussain, Waheed and Margaret Kohn. 2024. "The Common Good." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/common-good/> Accessed 2026-10-08.
- Hutchins, Edwin. 1995. *Cognition in the Wild*. The MIT Press. <https://doi.org/10.7551/mitpress/1881.001.0001>
- Hüttemann, Andreas. 2025. "Arguing for the aim of science." *Asian Journal of Philosophy* 4(1): 65. <https://doi.org/10.1007/s44204-025-00290-y>
- Ichikawa, Jonathan and Matthias Steup. 2026. "The Analysis of Knowledge." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/knowledge-analysis/> Accessed 2026-10-08.

2026-09-07.

Irvine, Elizabeth. 2016. "Model-Based Theorising in Cognitive Neuroscience." *British Journal for the Philosophy of Science* 67: 143–168. Page numbers cited in the text refer to the author's submitted manuscript in the Cardiff University repository (ORCA). https://orca.cardiff.ac.uk/id/eprint/81122/3/MBT_3_submission_BJPS_singlespaced.pdf

Irving, Geoffrey and Amanda Askill. 2019. "AI Safety Needs Social Scientists." *Distill* 4(2). Section titles cited in the text refer to the online article. <https://doi.org/10.23915/distill.00014>

Irzik, Gürol and Faik Kurtulmus. 2019. "What Is Epistemic Public Trust in Science?" *The British Journal for the Philosophy of Science* 70(4): 1145–1166. Section numbers cited in the text refer to the authors' preprint, PhilSci-Archive 15370, <https://philsci-archive.pitt.edu/15370/>; they are the same as in the published article. <https://doi.org/10.1093/bjps/axy007>

Ivanova, Milena. 2010. "Pierre Duhem's good sense as a guide to theory choice." *Studies in History and Philosophy of Science Part A* 41(1): 58–64. Section numbers cited in the text refer to the author's preprint, PhilSci-Archive 8679, <https://philsci-archive.pitt.edu/8679/>. <https://doi.org/10.1016/j.shpsa.2009.12.009>

Jaffe, Arthur and Frank Quinn. 1993. "Theoretical Mathematics': Toward a Cultural Synthesis of Mathematics and Theoretical Physics." *Bulletin of the American Mathematical Society* 29(1): 1–13. <https://doi.org/10.1090/s0273-0979-1993-00413-0>

Jaffe, Oliver and Dane Sherburn. 2026. "TasteVal: Measuring the Experimental Research Taste of AI Systems Against Human Experts." *arXiv*. Preprint, version 2; page numbers cited in the text refer to this version. <https://arxiv.org/abs/2610.06824v2>

Janssen, Michel and Matthew Mecklenburg. 2004. *Electromagnetic Models of the Electron and the Transition from Classical to Relativistic Mechanics*. Preprint 277. Max Planck Institute for the History of Science. Page numbers cited in the text refer to this preprint. A revised version was published as "From Classical to Relativistic Mechanics: Electromagnetic Models of the Electron," in *Interactions: Mathematics, Physics and Philosophy, 1860–1930*, Boston Studies in the Philosophy of Science 251, Springer, 2007, 65–134, https://doi.org/10.1007/978-1-4020-5195-1_3. <https://mpiwg-berlin.mpg.de/sites/default/files/migrated/p277.pdf> Accessed 2026-10-29.

Jarvie, I. C. 2001. "Science in a Democratic Republic." *Philosophy of Science* 68(4): 545–564. Page numbers cited in the text refer to the published article; the citation by "abstract" is to the article's abstract. <https://doi.org/10.1086/392942>

Jaworska, Agnieszka and Julie Tannenbaum. 2021. "The Grounds of Moral Status." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/grounds-moral-status/> Accessed 2026-09-07.

Jeannot, Emmanuel. 2026. "The Industrialization of Research: On AI-Driven Science and Its Consequences." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2607.15164v1>

Jin, Yiqiao, Yiyang Wang, Lucheng Fu, Bing He, Siheng Xiong, Yijia Xiao, B. Aditya Prakash, Josiah Hester, Srikanth Kumar, James Evans, and Jindong Wang. 2026. "Science Utopia? Closed-Loop LLM Simulation of Academic Research Ecosystems." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2610.01257v1>

John, Stephen. 2015a. "Inductive risk and the contexts of communication." *Synthese* 192(1): 79–96. Section numbers cited in the text refer to the accepted manuscript. <https://doi.org/10.1007/s11229-014-0554-7>

John, Stephen. 2015b. "The example of the IPCC does not vindicate the Value Free Ideal: a reply to Gregor Betz." *European Journal for Philosophy of Science* 5(1): 1–13. Section numbers cited in the text refer to the accepted manuscript, Apollo – University of Cambridge Repository, <https://www.repository.cam.ac.uk/handle/1810/246084>. <https://doi.org/10.1007/s13194-014-0095-4>

John, Stephen. 2025. "Against alignment: the value of non-democratic science." *European Journal for Philosophy of Science* 15(4): 48. <https://doi.org/10.1007/s13194-025-00680-2>

Johnson, Michael R. 1994. "The Galileo high gain antenna deployment anomaly." *NASA. Lewis Research Center; The 28th Aerospace Mechanisms Symposium*. <https://ntrs.nasa.gov/citations/19940028813> Accessed 2026-09-29.

Ju, Haocheng, Guoxiong Gao, Jiedong Jiang, Bin Wu, Zeming Sun, Shurui Liu, Leheng Chen, Yutong Wang, Yuefeng Wang, Zichen Wang, Wanyi He, Peihao Wu, Liang Xiao, Ruochuan Liu, Bryan Dai, and Bin Dong. 2026. "Automated Conjecture Resolution with Formal Verification." *arXiv*. Preprint. <https://arxiv.org/abs/2604.03789>

Juelsholt, Mikkel. 2026. "Continued challenges in high-throughput materials predictions: MatterGen predicts compounds from the training dataset." *Materials Horizons* 13(11): 5672–5679. Consulted the copy deposited in the Humboldt-Universität zu Berlin repository (edoc, <https://doi.org/10.18452/37763>), which carries the journal pagination. <https://doi.org/10.1039/d6mh00268d>

Kallestrup, Jesper. 2022. "Comments on Hiller and Randall's 'Epistemic Structure in Non-Summative Social Knowledge'." *Social Epistemology Review and Reply Collective* 11(11): 38–41. Cited without page numbers from the web version on the journal's website, which has none. <https://social-epistemology.com/2022/11/18/comments-on-hiller-and-randalls-epistemic-structure-in-non-summative-social-knowledge-jesper-kallestrup/> Accessed 2026-10-30.

- Kallestrup, Jesper. 2024. "Nonreductive Group Knowledge Revisited." *Episteme* 21(2): 565–588. <https://doi.org/10.1017/epi.2022.19>
- Karaca, Koray. 2017. "A case study in experimental exploration: exploratory data selection at the Large Hadron Collider." *Synthese* 194: 333–354. <https://doi.org/10.1007/s11229-016-1206-x>
- Karnofsky, Holden. 2021. "Forecasting Transformative AI, Part 1: What Kind of AI?" *Cold Takes*, 10 August 2021. Blog post; cited by section heading and note number. <https://www.cold-takes.com/transformative-ai-timelines-part-1-of-4-what-kind-of-ai/> Accessed 2026-10-17.
- Kaul, Inge and Ronald U. Mendoza. 2003. "Advancing the Concept of Public Goods." In *Providing Global Public Goods: Managing Globalization*, edited by Inge Kaul, Pedro Conceição, Katell Le Goulven, and Ronald U. Mendoza, 78–111. Oxford University Press for the United Nations Development Programme. Internet Archive capture of the UNDP copy. <https://web.archive.org/web/20030628130148/http://www.undp.org:80/globalpublicgoods/globalization/pdfs/KaulMendoza.pdf> Accessed 2026-10-18.
- Kauppinen, Antti. 2026. "Reasons for questions and the norms of inquiry." *Philosophical Studies*. Published online ahead of issue; section numbers cited in the text refer to the author's version. <https://doi.org/10.1007/s11098-026-02570-6>
- Kawall, Jason. 2024. "Epistemic Autonomy and the Shaping of Our Epistemic Lives." *Social Epistemology* 38(3): 374–391. <https://doi.org/10.1080/02691728.2024.2326840>
- Kearl, Timothy R. 2023. "Epistemic control without voluntarism." *Philosophical Issues* 33(1): 95–109. <https://doi.org/10.1111/phils.12245>
- Keeling, Charles D. 1998. "Rewards and Penalties of Monitoring the Earth." *Annual Review of Energy and the Environment* 23(1): 25–82. Cited from the Annual Reviews reprint ("Reprinted, with permission") formerly posted by the Scripps CO2 Program, now available as an Internet Archive capture (25 April 2026), https://web.archive.org/web/20260425085457/https://scrippsco2.ucsd.edu/assets/publications/keeling_autobiography.pdf; it carries the journal pagination. <https://doi.org/10.1146/annurev.energy.23.1.25>
- Keeling, Sophie. 2023. "Accounting for doxastic agency: Mental action and self-awareness." *Synthese* 201(6): 185. <https://doi.org/10.1007/s11229-023-04069-7>
- Kelly, Thomas. 2014. "Evidence." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/evidence/> Accessed 2026-10-05.
- Kelp, Christoph. 2014. "Knowledge, Understanding and Virtue." In *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*, edited by Abrol Fairweather, 347–360. Springer. The author's version was consulted. https://doi.org/10.1007/978-3-319-04672-3_20
- Kelp, Christoph. 2015. "Understanding Phenomena." *Synthese* 192: 3799–3816. Page numbers cited in the text refer to the author's submitted version, Lirias (KU Leuven), <https://lirias.kuleuven.be/retrieve/3f99f143-0d17-4e32-a41f-672cb68f78e5>; the published pagination differs. <https://lirias.kuleuven.be/retrieve/3f99f143-0d17-4e32-a41f-672cb68f78e5> Accessed 2026-10-29.
- Kelp, Christoph. 2017. "Towards a Knowledge-Based Account of Understanding." In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, edited by Stephen R. Grimm, Christoph Baumberger, and Sabine Ammon, 251–271. Routledge. Page numbers cited in the text refer to the accepted manuscript, Enlighten (University of Glasgow), <https://eprints.gla.ac.uk/140976/>; the published pagination differs. <https://eprints.gla.ac.uk/140976/> Accessed 2026-10-29.
- Kelp, Christoph. 2021b. "Inquiry, knowledge and understanding." *Synthese* 198(Suppl 7): S1583–S1593. <https://doi.org/10.1007/s11229-018-1803-y>
- Kelp, Christoph. 2021c. "Theory of inquiry." *Philosophy and Phenomenological Research* 103(2): 359–384. Page numbers cited in the text refer to the accepted manuscript, Enlighten (University of Glasgow), <http://eprints.gla.ac.uk/219477/>; the published pagination differs. <https://doi.org/10.1111/phpr.12719>
- Kepler, Johannes. [1609] 1937. "Astronomia nova." In *Johannes Kepler Gesammelte Werke, vol. 3*, edited by Max Caspar. C. H. Beck; digital edition Bayerische Akademie der Wissenschaften. <https://publikationen.badw.de/de/002334739> Accessed 2026-09-29.
- Keren, Arnon. 2013. "Kitcher on Well-Ordered Science: Should Science Be Measured against the Outcomes of Ideal Democratic Deliberation?" *Theoria: An International Journal for Theory, History and Foundations of Science* 28(2): 233–244. <https://doi.org/10.1387/theoria.6590>
- Khalifa, Kareem. 2019. "Is Verstehen Scientific Understanding?" *Philosophy of the Social Sciences* 49(4): 282–306. Page numbers cited in the text refer to the author's version ("Forthcoming in Philosophy of the Social Sciences"), PhilSci-Archive, <https://philsci-archive.pitt.edu/16248/>; the published pagination differs. <https://doi.org/10.1177/0048393119847104>
- Khalifa, Kareem. 2022. "Should Friends and Frenemies of Understanding Be Friends? Discussing de Regt." In *Scientific Understanding and Representation: Modeling in the Physical Sciences*, edited by Kareem Khalifa, Insa Lawler, and Elay Shech, 33–50. Routledge. Page numbers cited in the text refer to the author's archive version (pp. 1–30), <https://philpapers.org/archive/KHASFA.pdf>. <https://doi.org/10.4324/9781003202905-4>

- Khan, Akbir, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. "Debating with More Persuasive LLMs Leads to More Truthful Answers." In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235. Page numbers cited in the text refer to arXiv version 4. <https://arxiv.org/abs/2402.06782v4>
- Khosrowi, Donal. 2026. "Automating Pursuitworthiness: Four Concerns About 'AI Scientists' and the Proper Roles for Machine Learning Systems in Scientific Discovery." Preprint, March 2026. PhilSci-Archive. <https://philsci-archive.pitt.edu/28854/>
- Kieval, Phillip Hintikka. 2024. "Artificial achievements." *Analysis* 84(1): 32–41. <https://doi.org/10.1093/analysis/anad052>
- Kim, Elliot Myunghoon, Avi Garg, Kenny Peng, and Nikhil Garg. 2025. "Correlated Errors in Large Language Models." In *Proceedings of the 42nd International Conference on Machine Learning*, 30038–30066. PMLR. <https://proceedings.mlr.press/v267/kim25e.html> Accessed 2026-09-29.
- Kim, Seungone, Dongkeun Yoon, Kiril Gashteovski, Juyoung Suk, Jinheon Baek, Pranjal Aggarwal, Ian Wu, Viktor Zaverkin, Spase Petkoski, Daniel R. Schrider, Ilija Dukovski, Francesco Santini, Biljana Mitreska, Yong Jeong, Kyeongha Kwon, Young Min Sim, Dragana Manasova, Arthur Porto, Biljana Mojsoska, Makoto Takamoto, Marko Shuntov, Ruqi Liu, Hyunjoon Jenny Lee, Niyazi Ulas Dinç, Yehyun Jo, Sunkyu Han, Chungwoo Lee, Huishan Li, Esther H. R. Tsai, Ergun Simsek, Khushboo Shafi, Yeonseung Chung, Jihye Park, Aleksandar Shulevski, Henrik Christiansen, Yoosang Son, Elly Knight, Amanda Montoya, Jeongyoun Ahn, Christian Langkammer, Heera Moon, Changwon Yoon, Nikola Stikov, Mooseok Jang, Edward Choi, Junhan Kim, Yeon Sik Jung, Woo Youn Kim, Jae Kyoung Kim, Ishraq Md Anjum, Hyun Uk Kim, Drew Bridges, Carolin Lawrence, Xiang Yue, Alice Oh, Akari Asai, Sean Welleck, and Graham Neubig. 2026. "On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2605.20668v1>
- King, Loren, Brandon Morgan-Olsen, and James Wong. 2016. "Identifying Difference, Engaging Dissent: What Is at Stake in Democratizing Knowledge?" *Foundations of Science* 21(1): 69–88. Read in the authors' manuscript on PhilPapers; page numbers are those of the manuscript. <https://philpapers.org/archive/KINIDE.pdf>
- King, Ross D. 2026. "The Past and Future of AI Scientists." *arXiv*. Preprint (arXiv:2608.14407). <https://arxiv.org/abs/2608.14407>
- King, Ross D., Kenneth E. Whelan, Ffion M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. 2004. "Functional genomic hypothesis generation and experimentation by a robot scientist." *Nature* 427(6971): 247–252. <https://doi.org/10.1038/nature02236>
- Kitano, Hiroaki. 2021. "Nobel Turing Challenge: creating the engine for scientific discovery." *npj Systems Biology and Applications* 7(1): 29. <https://doi.org/10.1038/s41540-021-00189-3>
- Kitcher, Philip. 2013. "Toward a Pragmatist Philosophy of Science." *Theoria: An International Journal for Theory, History and Foundations of Science* 28(2): 185–231. <https://doi.org/10.1387/theoria.7014>
- Kitcher, Philip. 2024. "The Scientist, Qua Scientist, Is an Ethical Agent." *Filozofia* 79(3): 231–243. <https://doi.org/10.31577/filozofia.2024.79.3.1>
- Klimovsky, Gregorio. [1971] 1975. "Ciencia e ideología." In *Ciencia e ideología: aportes polémicos*, 11–40. Ediciones Ciencia Nueva. Interview first published in *Ciencia Nueva* (1971). <https://repositorio.esocite.la/388> Accessed 2026-10-08.
- Klug, Dana M., Edwin G. Tse, Daniel G. Silva, Yafeng Cao, Susan A. Charman, Jyoti Chauhan, Elly Crighton, Maria Dichiaro, Chris Drake, David Drewry, Flavio da Silva Emery, Lori Ferrins, Lee Graves, Emily Hopkins, Thomas A. C. Kresina, Álvaro Lorente-Macias, Benjamin Perry, Richard Phipps, Bruno Quiroga, Antonio Quotadamo, Giada N. Sabatino, Anthony Sama, Andreas Schätzlein, Quillon J. Simpson, Jonathan Steele, Julia Shanu-Wilson, Peter Sjö, Paul Stapleton, Christopher J. Swain, Alexandra Vaideanu, Huanxu Xie, William Zuercher, and Matthew H. Todd. 2023. "Open Source Antibiotics: Simple Diarylimidazoles Are Potent against Methicillin-Resistant *Staphylococcus aureus*." *ACS Infectious Diseases* 9(12): 2423–2435. Supporting information cited as SI. <https://doi.org/10.1021/acsinfecdis.3c00286>
- Knorr Cetina, Karin. 1993. "Liminal and Referent Epistemologies in Contemporary Science: An Ethnography of the Empirical in Two Sciences." *Teoria sociologica* 2: 258–282. Text of a paper presented at the Institute for Advanced Study, Princeton, on 3 December 1992. Cited from the scan in KOPS, the University of Konstanz repository, <https://kops.uni-konstanz.de/handle/123456789/11523>, which carries the journal pagination. <https://kops.uni-konstanz.de/handle/123456789/11523> Accessed 2026-10-29.
- Knorr Cetina, Karin. 1999. *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press. Page numbers cited in the text refer to the German edition, *Wissenskulturen: Ein Vergleich naturwissenschaftlicher Wissensformen* (Suhrkamp, 2002), whose full text is in KOPS, the University of Konstanz repository, <https://kops.uni-konstanz.de/handle/123456789/11729>. Quotations are this book's own translations from the German. <https://www.jstor.org/stable/j.ctvxw3q7f> Accessed 2026-09-07.
- Knuuttila, Tarja and Andrea Loettgers. 2020. "Magnetized Memories: Analogies and Templates in Model Transfer." In *Philosophical Perspectives on the Engineering Approach in Biology: Living Machines?*, edited by Sune Holm and Maria Serban, 123–140. Routledge. Page numbers cited in the text refer to the authors' manuscript on PhilArchive. <https://philpapers.org/archive/KNUMMA-2.pdf>
- Knuuttila, Tarja and Andrea Loettgers. 2023. "Model Templates: Transdisciplinary Application and Entanglement." *Synthese* 201: 200. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10238306/>

- Koons, Robert. 2025. "Defeasible Reasoning." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/reasoning-defeasible/> Accessed 2026-10-11.
- Korcz, Keith Allen. 2026. "The Epistemic Basing Relation." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/basing-epistemic/> Accessed 2026-09-07.
- Koskinen, Inkeri. 2024a. "We Have No Satisfactory Social Epistemology of AI-Based Science." *Social Epistemology* 38(4): 458–475. <https://doi.org/10.1080/02691728.2023.2286253>
- Koskinen, Inkeri. 2024b. "We Still Have No Satisfactory Social Epistemology of AI-Based Science: A Response to Peters." *Social Epistemology Review and Reply Collective* 13(5): 9–14. Read as the web version, which has no page numbers. <https://wp.me/p1Bfg0-8LL>
- Kovanis, Michail, Raphaël Porcher, Philippe Ravaud, and Ludovic Trinquart. 2016. "The Global Burden of Journal Peer Review in the Biomedical Literature: Strong Imbalance in the Collective Enterprise." *PLOS ONE* 11(11): e0166387. Page numbers are those of the article PDF. <https://doi.org/10.1371/journal.pone.0166387>
- Kraft, Angelie. 2025. "On Knowledge in AI: Epistemic and Ethical Limitations of Language Models and Knowledge Graphs." Universität Hamburg. PhD dissertation (submitted 20 November 2025; defended 4 February 2026), Universität Hamburg repository; page numbers are those printed. https://ediss.sub.uni-hamburg.de/bitstream/ediss/12231/1/dissertation_angelie_kraft.pdf Accessed 2026-10-31.
- Krenn, Mario, Jakob S. Kottmann, Nora Tischler, and Alán Aspuru-Guzik. 2021. "Conceptual Understanding through Efficient Automated Design of Quantum Optical Experiments." *Physical Review X* 11(3): 031044. <https://doi.org/10.1103/physrevx.11.031044>
- Krenn, Mario, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, Zhenpeng Yao, and Alán Aspuru-Guzik. 2022. "On scientific understanding with artificial intelligence." *Nature Reviews Physics* 4(12): 761–769. <https://doi.org/10.1038/s42254-022-00518-3>
- Krenn, Mario and Heather Champion. 2026. "Philosophy of Autonomous Science: Ten Questions for the Coming Age of Artificial Scientists." *Dædalus* 155(1–2): 335–349. <https://doi.org/10.1162/daed.a.991>
- Krämer, Fabian. 2018. "Shifting Demarcations: An Introduction." *History of Humanities* 3(1): 5–14. <https://doi.org/10.1086/696298>
- Kuhn, Thomas S. 1990. "Dubbing and Redubbing: The Vulnerability of Rigid Designation." In *Scientific Theories*, edited by C. Wade Savage, 298–318. University of Minnesota Press. <https://hdl.handle.net/11299/185720> Accessed 2026-10-29.
- Kuhn, Thomas S. [1990] 2012. "Cesta od Struktury." Translated by Martin Vondrášek. *Teorie vědy / Theory of Science* 34(2): 125–139. Czech translation of "The Road since Structure," in PSA 1990: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol. 2 (1990), 3–13. The passages cited are paraphrased from the Czech. <https://doi.org/10.46938/tv.2012.163>
- Kuhn, Thomas S. [1984] 2017. *Desarrollo científico y cambio de léxico: Conferencias Thalheimer*. FIC-UdelaR, ANII and SADAF. Spanish translation by Leandro Giri of the Thalheimer Lectures that Kuhn delivered at Johns Hopkins University, 12–19 November 1984, edited by Pablo Melogno and Hernán Miguel; the English text is unpublished. Open copy in MIT Libraries, Distinctive Collections (DOME), <https://hdl.handle.net/1721.3/189540>. The passage cited is paraphrased from the Spanish. <https://hdl.handle.net/1721.3/189540> Accessed 2026-10-29.
- Kuhn, Thomas S. [1951] 2021. *The Quest for Physical Theory: Problems in the Methodology of Scientific Research*. Thomas S. Kuhn's Lowell Lectures, delivered in March 1951, edited with an introduction by George A. Reisch. Open copy in MIT Libraries, Distinctive Collections (DOME); revised version of the edition, file of 25 June 2022; page numbers are those of the edition. <https://hdl.handle.net/1721.3/189338> Accessed 2026-10-29.
- Kulkarni, Deepak and Herbert A. Simon. 1988. "The Processes of Scientific Discovery: The Strategy of Experimentation." *Cognitive Science* 12(2): 139–175. Page numbers cited in the text refer to the manuscript version in the Herbert A. Simon Collection, Carnegie Mellon University Libraries, linked here. https://iiif.library.cmu.edu/file/Simon_box00082_fld06584_bd0001_doc0001/Simon_box00082_fld06584_bd0001_doc0001.pdf
- Kulveit, Jan, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. 2025. "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development." *arXiv*. Preprint, version 2 (29 January 2025). <https://arxiv.org/abs/2501.16946>
- Kumar, Pawan, Michael W. Schelle, Madhulika Jain, Fiona L. Lin, Christopher J. Petzold, Michael D. Leavell, Julie A. Leary, Jeffery S. Cox, and Carolyn R. Bertozzi. 2007. "PapA1 and PapA2 are acyltransferases essential for the biosynthesis of the Mycobacterium tuberculosis virulence factor Sulfolipid-1." *Proceedings of the National Academy of Sciences* 104(27): 11221–11226. <https://doi.org/10.1073/pnas.0611649104>
- Kummerfeld, Erich and Kevin J. S. Zollman. 2016. "Conservatism and the Scientific State of Nature." *The British Journal for the Philosophy of Science* 67(4): 1057–1076. Section numbers cited in the text refer to the authors' version dated 30 July 2014, posted on Kevin Zollman's website, <https://www.kevinzollman.com/uploads/5/0/3/6/50361245/conservativismv16.pdf>; they are the same as in the published article. <https://doi.org/10.1093/bjps/axv013>
- Kuorikoski, Jaakko and Petri Ylikoski. 2015. "External representations and scientific understanding." *Synthese* 192(12): 3817–3837. Section numbers cited in the text refer to the SocArXiv preprint, <https://osf.io/download/qewtn/>. <https://doi.org/10.1007/s11229-014-0591-2>

- Kureha, Makoto and Minao Kukita. 2019. "AI jidai no kagaku tetsugaku [Philosophy of Science in the Age of AI]." Japan Association for Philosophy of Science. Conference abstract. https://phsc.jp/dat/rsm/20190515_16AMA2.pdf
- Kureha, Makoto. 2023. "Implications of Automating Science: The Possibility of Artificial Creativity and the Future of Science." *Journal of Philosophy of Life* 13: 44–63. <https://doi.org/10.20556/0002001002>
- Kureha, Makoto. 2024. "AI kudō kagaku no tetsugaku no kore made to kore kara [The Philosophy of AI-Driven Science: So Far and from Here]." Philosophy of Science Society, Japan. Slides (in Japanese) for an invited talk in the symposium "AI wa kagaku o dō kaeru no ka?" [How Will AI Change Science?] at the 2024 annual meeting of the Philosophy of Science Society, Japan, Waseda University, 15 June 2024, posted on the author's researchmap page; cited by slide number. https://researchmap.jp/whiteelephant0901/presentations/46786110/attachment_file.pdf Accessed 2026-10-29.
- Kusumegi, Keigo, Xinyu Yang, Paul Ginsparg, Mathijs de Vaan, Toby Stuart, and Yian Yin. 2025. "Scientific production in the era of large language models." *Science*, published 18 December 2025. Page numbers cited in the text refer to the authors' accepted version, <https://arxiv.org/abs/2601.13187>. <https://doi.org/10.1126/science.adw3000>
- Kvanvig, Jonathan L. 2003. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511498909>
- Kwolek, Stephanie L. 1998. "Oral history interview with Stephanie L. Kwolek." *Science History Institute Digital Collections*. Science History Institute. <https://digital.sciencehistory.org/works/vh53ww75r> Accessed 2026-09-29.
- Lakatos, Imre. 1973. "Science and Pseudoscience." *London School of Economics and Political Science*. Transcript of a 1973 lecture; Internet Archive copy of the page. <https://web.archive.org/web/20110218195726/http://www2.lse.ac.uk:80/philosophy/about/lakatos/scienceAndPseudoscienceTranscript.aspx> Accessed 2026-09-29.
- Lammenranta, Markus. n.d. "Epistemic Circularity." *Internet Encyclopedia of Philosophy*. The entry shows no date of publication. <https://iep.utm.edu/ep-circ/> Accessed 2026-10-30.
- Lange, Robert Tjarko, Luca Grillotti, Yutaro Yamada, Rujikorn Charakorn, and David Ha. 2026. "The Agentic Scientist: Orchestrating Coding Agents for Automated Scientific (Co-)Discovery." *arXiv*. Preprint, version 1; page numbers cited in the text refer to this version. <https://arxiv.org/abs/2610.18933v1>
- Latour, Bruno and Steve Woolgar. 1986. *Laboratory Life: The Construction of Scientific Facts*. 2nd ed. Princeton University Press. <https://doi.org/10.1515/9781400820412>
- Latour, Bruno. 1986. "Visualisation and Cognition: Drawing Things Together." In *Knowledge and Society: Studies in the Sociology of Culture Past and Present*, edited by Henrika Kuklick, 1–40. JAI Press. Page numbers cited in the text refer to the author's English version on his website, <https://www.bruno-latour.fr/sites/default/files/21-DRAWING-THINGS-TOGETHER-GB.pdf>; the published pagination differs. <https://www.bruno-latour.fr/sites/default/files/21-DRAWING-THINGS-TOGETHER-GB.pdf> Accessed 2026-10-29.
- Latour, Bruno. 1992. "Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts." In *Shaping Technology/Building Society: Studies in Sociotechnical Change*, edited by Wiebe E. Bijker and John Law, 225–258. MIT Press. Page numbers cited in the text refer to the reprint in *Technology and Society: Building Our Sociotechnical Future*, edited by Deborah G. Johnson and Jameson M. Wetmore, 151–180 (MIT Press, 2008), linked here from the author's site. <https://www.bruno-latour.fr/sites/default/files/50-MISSING-MASSSES-GB.pdf>
- Latour, Bruno. 1994. "On Technical Mediation: Philosophy, Sociology, Genealogy." *Common Knowledge* 3(2): 29–64. <https://www.bruno-latour.fr/sites/default/files/54-TECHNIQUES-GB.pdf>
- Latour, Bruno. 1996. "On Actor-Network Theory: A Few Clarifications." *Soziale Welt* 47(4): 369–381. Page numbers cited in the text refer to the author's English version, "On Actor-Network Theory: A Few Clarifications, Plus More Than a Few Complications," linked here. <https://www.bruno-latour.fr/sites/default/files/P-67%20ACTOR-NETWORK.pdf>
- Lattimore, Tor and Csaba Szepesvári. 2020. *Bandit Algorithms*. Cambridge University Press. Section numbers cited in the text refer to the authors' free online edition, <https://tor-lattimore.com/downloads/book/book.pdf>, whose content is that of the print edition with minor corrections; its page numbers differ. <https://doi.org/10.1017/9781108571401>
- Laudan, Larry, Arthur Donovan, Rachel Laudan, Peter Barker, Harold Brown, Jarrett Leplin, Paul Thagard, and Steve Wykstra. 1993. "Mudança científica: modelos filosóficos e pesquisa histórica." *Estudos Avançados* 7(19): 7–89. Translated by Caetano Ernesto Plastino from "Scientific Change: Philosophical Models and Historical Research," *Synthese* 69 (1986): 141–223. Page numbers cited in the text are those of the translation; the theses it lists carry the authors' page references to Laudan's *Progress and Its Problems* (1977) and *Science and Values* (1984). <https://doi.org/10.1590/s0103-40141993000300002>
- Law, John. 1992. "Notes on the Theory of the Actor-Network: Ordering, Strategy, and Heterogeneity." *Systems Practice* 5(4): 379–393. Page numbers cited in the text refer to the author's online version, linked here. <http://www.heterogeneities.net/publications/Law1992NotesOnTheTheoryOfTheActorNetwork.pdf>

- Lazar, Seth. 2024. "Legitimacy, Authority, and Democratic Duties of Explanation." *Oxford Studies in Political Philosophy* 10. Section numbers cited in the text are those of the arXiv version 4. <https://arxiv.org/abs/2208.08628v4>
- Le Bihan, Soazig. 2017. "Enlightening Falsehoods: A Modal View of Scientific Understanding." In *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, edited by Stephen R. Grimm, Christoph Baumberger, and Sabine Ammon, 111–135. Routledge. Cited without page numbers; the claims cited are from the author's final draft. <https://philpapers.org/rec/LEBEFA-2>
- LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521(7553): 436–444. Cited from the PDF of the published article posted on Geoffrey Hinton's University of Toronto page, <https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf>, which carries the journal pagination. <https://doi.org/10.1038/nature14539>
- Lee, Carole J. 2026. "Critically Engaged Pragmatism: Scientific Norm and Social, Pragmatist Epistemology for AI Science Evaluation Tools." *Social Epistemology* 40(5): 584–598. Page numbers cited in the text refer to the version published online on 9 June 2026 (pp. 1–15). <https://doi.org/10.1080/02691728.2026.2669578>
- Lee, Wooram. 2026. "Belief, doubt, and inquiry." *Synthese* 208(1): 50. <https://doi.org/10.1007/s11229-026-05707-6>
- Leeman, Josh, Yuhua Liu, Joseph Stiles, Scott B. Lee, Prajna Bhatt, Leslie M. Schoop, and Robert G. Palgrave. 2024. "Challenges in High-Throughput Inorganic Materials Prediction and Autonomous Synthesis." *PRX Energy* 3(1): 011002. <https://doi.org/10.1103/prxenergy.3.011002>
- Lehman, Joel and Kenneth O. Stanley. 2011. "Abandoning Objectives: Evolution through the Search for Novelty Alone." *Evolutionary Computation* 19(2): 189–223. Section numbers cited in the text refer to the authors' manuscript, formerly on the authors' laboratory website and now available as an Internet Archive capture, https://web.archive.org/web/20220707/http://eplex.cs.ucf.edu/papers/lehman_ecj11.pdf. https://doi.org/10.1162/EVCO_a_00025
- Leibo, Joel Z., Alexander Sasha Vezhnevets, William A. Cunningham, Sébastien Krier, Manfred Diaz, and Simon Osindero. 2025. "Societal and technological progress as sewing an ever-growing, ever-changing, patchy, and polychrome quilt." *arXiv*. Position-paper preprint, version 1. <https://arxiv.org/abs/2505.05197>
- Leike, Jan and Marcus Hutter. 2015. "Bad Universal Priors and Notions of Optimality." In *Proceedings of the 28th Conference on Learning Theory*, 1244–1259. PMLR. Page numbers cited in the text refer to the PDF's own pagination (1–16). <https://proceedings.mlr.press/v40/Leike15.html> Accessed 2026-10-02.
- Leitgeb, Hannes and André Carus. 2020. "Rudolf Carnap." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/carnap/> Accessed 2026-09-07.
- Leonard, Nick. 2021. "Epistemological Problems of Testimony." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/testimony-episprob/> Accessed 2026-09-29.
- Leonelli, Sabina. 2015. "What Counts as Scientific Data? A Relational Framework." *Philosophy of Science* 82(5): 810–821. Cited from the author manuscript in PubMed Central, <https://pmc.ncbi.nlm.nih.gov/articles/PMC4747116/>, which has no page numbers. <https://doi.org/10.1086/684083>
- Leonelli, Sabina. 2023. "Philosophy of Open Science." *Elements in the Philosophy of Science*. <https://doi.org/10.1017/9781009416368>
- Levy, Steven. 2024. "8 Google Employees Invented Modern AI. Here's the Inside Story." *WIRED*, 20 March 2024. Retrospective interviews with the authors of "Attention Is All You Need." <https://www.wired.com/story/eight-google-employees-invented-modern-ai-transformers-paper/> Accessed 2026-10-14.
- Lewens, Tim. 2006. "Distinguishing treatment from research: a functional approach." *Journal of Medical Ethics* 32(7): 424–429. <https://doi.org/10.1136/jme.2005.013078>
- Li, Fei-Fei. 2023. *The Worlds I See: Curiosity, Exploration, and Discovery at the Dawn of AI*. Flatiron Books. <https://books.google.com/books?id=CBKhEAAQBAJ> Accessed 2026-09-30.
- Li, Runhong (黎润红) and Zhang Daqing (张大庆). 2019. "青蒿素：从中国传统药方到全球抗疟良药 [Artemisinin: From Traditional Chinese Prescription to Global Antimalarial Drug]." *中国科学院院刊 [Bulletin of Chinese Academy of Sciences]* 34(9): 1046–1057. <https://doi.org/10.16418/j.issn.1000-3045.2019.09.009>
- Li, Xiang, Shanghong Xie, Peter McColgan, Sarah J. Tabrizi, Rachael I. Scahill, Donglin Zeng, and Yuanjia Wang. 2018. "Learning Subject-Specific Directed Acyclic Graphs With Mixed Effects Structural Equation Models From Observational Data." *Frontiers in Genetics* 9: 430. <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2018.00430/full>
- Liang, Weixin, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. "Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews." In *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235. Page numbers cited in the text refer to arXiv:2403.07183v4. <https://arxiv.org/abs/2403.07183>
- Licklider, J. C. R. 1960. "Man-Computer Symbiosis." *IRE Transactions on Human Factors in Electronics* HFE-1(1): 4–11. Section numbers cited in the text are those of the reprint, "reprinted, with permission," in *In Memoriam: J. C. R. Licklider 1915–1990*, SRC Research Report 61 (Palo Alto:

- Digital Equipment Corporation Systems Research Center, 1990); the text links the Internet Archive capture of 20 August 2022, <https://web.archive.org/web/20220820021938/https://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-61.pdf>. <https://doi.org/10.1109/THFE2.1960.4503259>
- Lin, Hanti. 2022. "Bayesian Epistemology." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/epistemology-bayesian/> Accessed 2026-10-13.
- Lipton, Peter. 1993. "Making a Difference." *Philosophica* 51: 39–54. <https://doi.org/10.21825/philosophica.82386>
- Lipton, Peter. 1998. "Binding the Mind." In *Consciousness and Human Identity*, edited by John Cornwell, 212–224. Oxford University Press. Page numbers cited in the text refer to the manuscript posted on the website of the Department of History and Philosophy of Science, University of Cambridge, <https://www.hps.cam.ac.uk/files/lipton-binding-mind.pdf>; the published pagination differs. <https://www.hps.cam.ac.uk/files/lipton-binding-mind.pdf> Accessed 2026-10-29.
- Lipton, Peter. 2001. "What Good Is an Explanation?" In *Explanation: Theoretical Approaches and Applications*, edited by Giora Hon and Sam S. Rakover, 43–59. Springer. Page numbers cited in the text refer to the manuscript posted on the website of the Department of History and Philosophy of Science, University of Cambridge, <https://www.hps.cam.ac.uk/files/lipton-what-good.pdf>; the published pagination differs. https://doi.org/10.1007/978-94-015-9731-9_2
- List, Christian and Philip Pettit. 2002. "Aggregating Sets of Judgments: An Impossibility Result." *Economics and Philosophy* 18(1): 89–110. Cited from the authors' electronic version of the published article, LSE Research Online, <https://researchonline.lse.ac.uk/id/eprint/704/>, which carries the journal pagination. <https://doi.org/10.1017/S0266267102001098>
- List, Christian. 2021. "Group Agency and Artificial Intelligence." *Philosophy & Technology* 34(4): 1213–1242. <https://doi.org/10.1007/s13347-021-00454-7>
- Long, Robert, Jeff Sebo, and Toni Sims. 2025. "Is there a tension between AI safety and AI welfare?" *Philosophical Studies* 182(7): 2005–2033. <https://doi.org/10.1007/s11098-025-02302-2>
- Longino, Helen. 2021. "Naturalism? What Naturalism?" *Proceedings and Addresses of the American Philosophical Association* 95: 140–157. 2021 Romanell Lecture, presented at the Pacific Division meeting of the American Philosophical Association, 7 April 2021. Page numbers cited in the text refer to the lecture text posted on PhilSci-Archive (19917); the published pagination differs. <https://philsci-archive.pitt.edu/19917/> Accessed 2026-10-29.
- Longino, Helen. 2025. "The Social Dimensions of Scientific Knowledge." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/scientific-knowledge-social/> Accessed 2026-10-29.
- Longino, Helen E. [2008] 2017. "Valores, heurística e política do conhecimento." Translated by Débora Aymoré. *Scientiae Studia* 15(1): 39–57. Authorized Portuguese translation of "Values, Heuristics, and the Politics of Knowledge" (2008); page numbers are those printed. <https://revistas.usp.br/ss/en/article/download/133642/129612/0>
- Longino, Helen E. 2025b. "Interaction, pluralism, and community in conflictual contexts." *Studies in History and Philosophy of Science* 112: 12–20. <https://doi.org/10.1016/j.shpsa.2025.05.004>
- Lovett, Frank. 2022. "Republicanism." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/republicanism/> Accessed 2026-09-29.
- Lowe, Ryan, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. "Towards an Automatic Turing Test: Learning to Evaluate Dialogue Responses." In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1116–1126. <https://aclanthology.org/P17-1103>
- Lowe, Ryan. 2019. "A Retrospective for 'Towards an Automatic Turing Test – Learning to Evaluate Dialogue Responses'." *ML Retrospectives*. https://ml-retrospectives.github.io/published_retrospectives/2019/adem Accessed 2026-09-29.
- Lu, Chris, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. 2026. "Towards end-to-end automation of AI research." *Nature* 651(8107): 914–919. <https://doi.org/10.1038/s41586-026-10265-5>
- Lu, Sirui, Ruixuan Deng, Yanqiao Zhu, and Zhengfeng Ji. 2026. "Long-horizon autoformalization of a core theorem underlying MIP* = RE." *arXiv*. Preprint. <https://arxiv.org/abs/2609.19814>
- Ludwig, Roman. 2023. "Modelling Lymphatic Metastatic Progression in Head and Neck Cancer." University of Zurich. Page numbers cited in the text refer to the version released on GitHub as v1.5 (2023-01-09). <https://doi.org/10.5167/uzh-231470>
- Luo, Ziming, Atoosa Kasirzadeh, and Nihar B. Shah. 2025. "The More You Automate, the Less You See: Hidden Pitfalls of AI Scientist Systems." *arXiv*. Preprint, version 2 (20 December 2025). <https://arxiv.org/abs/2509.08713v2>
- Lusk, Greg and Kevin C. Elliott. 2024. "Against global aims for science: values, epistemic priority, and a local aims approach." *Synthese* 204(2): 47. <https://doi.org/10.1007/s11229-024-04633-9>

- LVD and OPERA Collaborations. 2012. “Determination of a time-shift in the OPERA set-up using high energy horizontal muons in the LVD and OPERA detectors.” *arXiv*. Preprint. <https://arxiv.org/abs/1206.2488>
- Lyons, Timothy D. 2006. “Scientific Realism and the Stratagema de Divide et Impera.” *The British Journal for the Philosophy of Science* 57(3): 537–560. Cited from the copy of the published article on the author’s PhilArchive record, <https://philarchive.org/rec/LYOSRA>, which carries the journal pagination. <https://doi.org/10.1093/bjps/axl021>
- Machado, Diogo. 2026. “Generative AI bias against scientific novelty: a cautionary tale from a small-sample evaluation of research proposals.” *Scientometrics* 131(10): 7479–7497. <https://doi.org/10.1007/s11192-026-05792-x>
- Malfatti, Federica I. 2020. “Can Testimony Transmit Understanding?” *Theoria* 86(1): 54–72. <https://doi.org/10.1111/theo.12220>
- Malpas, Jeff. 2023. “Donald Davidson.” *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/davidson/> Accessed 2026-10-29.
- Manheim, David and Scott Garrabrant. 2019. “Categorizing Variants of Goodhart’s Law.” *arXiv*. Preprint, version 4. <https://arxiv.org/abs/1803.04585v4>
- Mansbridge, Jane, James Bohman, Simone Chambers, Thomas Christiano, Archon Fung, John Parkinson, Dennis F. Thompson, and Mark E. Warren. 2012. “A systemic approach to deliberative democracy.” In *Deliberative Systems: Deliberative Democracy at the Large Scale*, edited by John Parkinson and Jane Mansbridge, 1–26. Cambridge University Press. Cited from the chapter in the published layout deposited in WRAP, University of Warwick, <https://doi.org/10.82444/warw.33655837>, which carries the book’s pagination. <https://doi.org/10.1017/cbo9781139178914.002>
- Marin Community. 2024. “Run Extraction Method Ablation on DCLM-baseline (issue #660).” *marin-community/marin*, *GitHub*. <https://github.com/marin-community/marin/issues/660> Accessed 2026-09-30.
- Marin Community. 2025. “Experiment: Anneal ema beta to 1 during deep cooldown (issue #1283).” *marin-community/marin*, *GitHub*. <https://github.com/marin-community/marin/issues/1283> Accessed 2026-09-30.
- Marin Community. 2026a. “Experiment: MoE Isoflop on new lr recipe (issue #4447).” *marin-community/marin*, *GitHub*. <https://github.com/marin-community/marin/issues/4447> Accessed 2026-09-29.
- Marin Community. 2026b. “Experiment: 1e23 MoE Run (issue #4697).” *marin-community/marin*, *GitHub*. <https://github.com/marin-community/marin/issues/4697> Accessed 2026-09-29.
- Marin Community. 2026c. “Experiment: Measure MoE sensitivity to tokenizer choice (issue #5821).” *marin-community/marin*, *GitHub*. <https://github.com/marin-community/marin/issues/5821> Accessed 2026-09-30.
- Marsili, Neri. 2023. “Group Assertions and Group Lies.” *Topoi* 42(2): 369–384. <https://doi.org/10.1007/s11245-022-09875-1>
- Martela, Frank. 2025. “Artificial intelligence and free will: generative agents utilizing large language models have functional free will.” *AI and Ethics* 5(4): 4389–4400. <https://doi.org/10.1007/s43681-025-00740-6>
- Mattarollo, Livio. 2022. “¿Tomando seriamente a Dewey? Una revisión crítica del ideal de ciencia bien ordenada de Philip Kitcher.” *Ideas y Valores* 71(180): 9–33. <https://doi.org/10.15446/ideasyvalores.v72n181.84190>
- Matthewson, John and Michael Weisberg. 2009. “The Structure of Tradeoffs in Model Building.” *Synthese* 170(1): 169–190. Cited without page numbers from the authors’ draft of 12 September 2007, Penn ScholarlyCommons, https://repository.upenn.edu/philosophy_papers/16; the published version was revised after this draft. <https://doi.org/10.1007/s11229-008-9366-y>
- Mayo, Deborah G. 1996. *Error and the Growth of Experimental Knowledge*. University of Chicago Press. Cited from chapter 4, “Duhem, Kuhn, and Bayes” (pp. 102–127), as posted by the author on her blog Error Statistics Philosophy; the file carries the book’s pagination. https://errorstatistics.com/wp-content/uploads/2021/03/egek-pdf-red_chapter-4.pdf Accessed 2026-10-29.
- Mayo, Deborah G. and Aris Spanos. 2006. “Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction.” *British Journal for the Philosophy of Science* 57(2): 323–357. Cited from the typeset article posted by the first author on her blog Error Statistics Philosophy, https://errorstatistics.com/wp-content/uploads/2013/12/2006mayo_spanos_severe_testing.pdf, which carries the journal pagination. <https://doi.org/10.1093/bjps/axl003>
- Mayo-Wilson, Conor, Kevin J. S. Zollman, and David Danks. 2011. “The Independence Thesis: When Individual and Social Epistemology Diverge.” *Philosophy of Science* 78(4): 653–677. <https://doi.org/10.1086/661777>
- McCammon, Richard C. 2025. “Domination.” *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/domination/> Accessed 2026-10-08.
- McClintock, Barbara. 1944. “Maize Genetics.” *Carnegie Institution of Washington Year Book* 43: 127–135. <https://digirepo.nlm.nih.gov/ext/document/101584613X28/PDF/101584613X28.pdf> Accessed 2026-09-29.
- McClintock, Barbara. 1949a. “Memorandum to Marcus Rhoades, First Section.” National Library of Medicine, Profiles in Science (Barbara McClintock Papers). <https://digirepo.nlm.nih.gov/ext/document/101584613X139/PDF/101584613X139.pdf> Accessed 2026-09-29.

- McClintock, Barbara. 1949b. "The Ds Locus. Part III. Transposition of the Ds Locus." National Library of Medicine, Profiles in Science (Barbara McClintock Papers). <https://digirepo.nlm.nih.gov/ext/document/101584613X149/PDF/101584613X149.pdf> Accessed 2026-09-29.
- McClintock, Barbara. 1950. "The Origin and Behavior of Mutable Loci in Maize." *Proceedings of the National Academy of Sciences* 36(6): 344–355. <https://doi.org/10.1073/pnas.36.6.344>
- McDowell, John. 2000. "Experiencing the World." In *John McDowell: Reason and Nature. Lecture and Colloquium in Münster 1999*, edited by Marcus Willaschek. Lit. The out-of-print volume is offered for download by the Philosophisches Seminar of the University of Münster, which organized the lecture: <https://www.uni-muenster.de/imperia/md/content/philosophischesseminar/veranstaltungen/mvph/mcdowell/mcdowellkolloq.pdf>. McDowell presents the lecture as an introduction to the conception of experience of *Mind and World*. <https://www.uni-muenster.de/imperia/md/content/philosophischesseminar/veranstaltungen/mvph/mcdowell/mcdowellkolloq.pdf> Accessed 2026-10-29.
- McHugh, Conor. 2013. "Epistemic Responsibility and Doxastic Agency." *Philosophical Issues* 23(1): 132–157. Section numbers cited in the text refer to the author's manuscript, linked from Conor McHugh's homepage, <https://sites.google.com/site/conormchughhomepage>, <https://docs.google.com/file/d/0B9AO9EjCCp3EZUhrbzVPdlpZVW8/edit?usp=sharing>; the manuscript has no journal page numbers. <https://doi.org/10.1111/phils.12007>
- McLeod, Carolyn. 2020. "Trust." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/trust/> Accessed 2026-10-29.
- McPherson, Tristram and David Plunkett. 2024. "Topic Continuity in Conceptual Engineering and Beyond." *Inquiry* 67(9): 2847–2873. <https://doi.org/10.1080/0020174x.2021.1980095>
- Messeri, Lisa and M. J. Crockett. 2024. "Artificial intelligence and illusions of understanding in scientific research." *Nature* 627(8002): 49–58. Cited from the PDF of the published article posted on M. J. Crockett's laboratory website, https://www.crockettlab.org/s/MesseriCrockett_2024_Nature.pdf, which carries the journal pagination. <https://doi.org/10.1038/s41586-024-07146-0>
- Meta AI. 2022. "OPT-175B Logbook." *facebookresearch/metaseq, GitHub*. Logbook of the training run, October 2021 to January 2022, released with redactions; cited by PDF page. https://github.com/facebookresearch/metaseq/blob/f7ffa5fd61cf90f498a36d365c13dd7f1a912ff7/projects/OPT/chronicles/OPT175B_Logbook.pdf Accessed 2026-09-29.
- Michael, Julian, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. 2023. "Debate Helps Supervise Unreliable Experts." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2311.08702v1>
- Michaelian, Kourken and John Sutton. 2017. "Memory." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/memory/> Accessed 2026-10-25.
- Miller, David. 2026. "Justice." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/justice/> Accessed 2026-10-08.
- Miller, Louis H. and Xinzhuan Su. 2011. "Artemisinin: Discovery from the Chinese Herbal Garden." *Cell* 146(6): 855–858. <https://doi.org/10.1016/j.cell.2011.08.024>
- Millière, Raphaël. 2025. "Normative Conflicts and Shallow AI Alignment." *Philosophical Studies* 182(7): 2035–2078. Page numbers cited in the text refer to arXiv:2506.04679v1. <https://doi.org/10.1007/s11098-025-02347-3>
- Mirsky, Reuth. 2025. "Artificial Intelligent Disobedience: Rethinking the Agency of Our Artificial Teammates." *AI Magazine* 46(2): e70011. Page numbers cited in the text are those of the arXiv version 1 (arXiv:2506.22276v1). <https://doi.org/10.1002/aaai.70011>
- MIT (Massachusetts Institute of Technology). 2025. "Assuring an Accurate Research Record." *MIT Stone Center on Inequality and Shaping the Future of Work*. Massachusetts Institute of Technology. <https://shapingwork.mit.edu/news/assuring-an-accurate-research-record> Accessed 2026-09-07.
- Mitchell, Sandra D. 2020. "Instrumental Perspectivism: Is AI Machine Learning Technology Like NMR Spectroscopy?" In *A Critical Reflection on Automated Science: Will Science Remain Human?*, edited by Marta Bertolaso and Fabio Sterpetti, 27–42. Springer. Page numbers cited in the text refer to the author's preprint of 12 February 2019, PhilSci-Archive 15738, <https://philsci-archive.pitt.edu/15738/>; the published pagination differs. https://doi.org/10.1007/978-3-030-25001-0_3
- Mittag, Daniel M. n.d. "Evidentialism." *Internet Encyclopedia of Philosophy*. The entry shows no date of publication. <https://iep.utm.edu/evidentialism> Accessed 2026-10-29.
- Miyoshi, Makoto, Yoshiaki Kato, and Junichiro Makino. 2022a. "The Jet and Resolved Features of the Central Supermassive Black Hole of M87 Observed with the Event Horizon Telescope (EHT)." *The Astrophysical Journal* 933(1): 36. <https://doi.org/10.3847/1538-4357/ac6ddb>

- Miyoshi, Makoto, Yoshiaki Kato, and Junichiro Makino. 2022b. "Comments on 'Imaging Reanalyses of EHT Data'." *arXiv*. Preprint. <https://arxiv.org/abs/2207.13279>
- Mladenović, Bojana. 2022. "Editor's Introduction." In *Thomas S. Kuhn, The Last Writings of Thomas S. Kuhn: Incommensurability in Science*, edited by Bojana Mladenović, xi–xlviii. University of Chicago Press. Free PDF of the editor's introduction, served by De Gruyter Brill; the identifier in its address is not a registered DOI, so the address is given in full. <https://www.degruyterbrill.com/document/doi/10.7208/chicago/9780226516301-002/pdf>
- Mohler, Chad and Bradley Monton. 2025. "Constructive Empiricism." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/constructive-empiricism/> Accessed 2026-09-29.
- Moody, Merijn and Henk W. de Regt. 2026. "Beyond Stochastic Parrots: Examining the Scientific Creativity of Large Language Models." *PhilSci-Archive*. Preprint. <https://philsci-archive.pitt.edu/30509> Accessed 2026-10-04.
- Moreno, Eric A., Samuel Bright-Thonney, Andrzej Novak, Dolores Garcia, and Philip Harris. 2026. "AI Agents Can Already Autonomously Perform Experimental High Energy Physics." *arXiv*. Preprint. <https://arxiv.org/abs/2603.20179>
- Morgan, Mary S. and Margaret Morrison (eds.). 1999. *Models as Mediators*. Cambridge University Press. <https://doi.org/10.1017/cbo9780511660108>
- Moro Simpson, Thomas. 1975. "Irracionalidad, ideología y objetividad." In *Ciencia e ideología: aportes polémicos*, 79–115. Ediciones Ciencia Nueva. <https://repositorio.esocite.la/388> Accessed 2026-10-08.
- Morrison, Margaret and Mary S. Morgan. 1999. "Models as Mediating Instruments." In *Models as Mediators*, edited by Mary S. Morgan and Margaret Morrison, 10–37. Cambridge University Press. <https://doi.org/10.1017/cbo9780511660108.003>
- Moxham, Noah and Aileen Fyfe. 2018. "The Royal Society and the prehistory of peer review, 1665–1965." *The Historical Journal* 61(4): 863–889. Page numbers cited in the text refer to the author's accepted manuscript, Kent Academic Repository, <https://kar.kent.ac.uk/65042/>; the published pagination differs. <https://doi.org/10.1017/S0018246X17000334>
- Musgrave, Alan and Charles Pigden. 2026. "Imre Lakatos." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/lakatos/> Accessed 2026-09-29.
- Myers-Schulz, Blake and Eric Schwitzgebel. 2013. "Knowing That P without Believing That P." *Noûs* 47(2): 371–384. Cited without page numbers from the authors' draft of 19 July 2011 on Eric Schwitzgebel's website, <https://faculty.ucr.edu/~eschwitz/SchwitzPapers/KB-110715.pdf>; the published version was revised after this draft. <https://doi.org/10.1111/nous.12022>
- Nersessian, Nancy J. 2009. "How Do Engineering Scientists Think? Model-Based Simulation in Biomedical Engineering Research Laboratories." *Topics in Cognitive Science* 1(4): 730–757. Cited from the publisher's PDF on the author's website, <https://sites.cc.gatech.edu/aimosaic/faculty/nersessian/papers/how-do-engineering-scientists-think.pdf>, which carries the journal pagination. <https://doi.org/10.1111/j.1756-8765.2009.01032.x>
- Nersessian, Nancy J. 2024. "In Vitro Analogies: Simulation Modeling in Biomedical Engineering Sciences." In *The Routledge Handbook of Philosophy of Scientific Modeling*, edited by Tarja Knuuttila, Natalia Carrillo, and Rami Koskinen, 469–481. Routledge. <https://doi.org/10.4324/9781003205647-40>
- Neurath, Otto. 1919. *Durch die Kriegswirtschaft zur Naturalwirtschaft*. Georg D. W. Callwey. Cited from the retro-digitized scan of the first edition in ZBW/EconBiz; the passages cited are in the essay "Die Utopie als gesellschaftstechnische Konstruktion" (pp. 228–231). Quotations are this book's own translations. An English translation of the essay, "Utopia as a Social Engineer's Construction," is in Otto Neurath, *Empiricism and Sociology*, edited by Marie Neurath and Robert S. Cohen (D. Reidel, 1973), 150–155. <https://digi.econbiz.de/viewer/image/891221816/242/> Accessed 2026-10-29.
- Neylon, Cameron. 2012. "Laboratory Notebook for doi:10.1186/1756-0500-4-138." figshare. Static export of the LabTrove notebook "Investigations into neutral drift," University of Southampton (snapshot of 7 January 2010), deposited as a dataset; cited by post number. <https://doi.org/10.6084/m9.figshare.92444.v1>
- Nickles, Thomas. 1981. "What is a problem that we may solve it?" *Synthese* 47(1): 85–118. <https://doi.org/10.1007/bf01064267>
- Nickles, Thomas. 1985. "Beyond Divorce: Current Status of the Discovery Debate." *Philosophy of Science* 52(2): 177–206. <https://doi.org/10.1086/289239>
- Nickles, Thomas. 1990. "Discovery Logics." *Philosophica* 45: 7–32. <https://doi.org/10.21825/philosophica.82435>
- Nickles, Thomas. 2017. "Historicist Theories of Scientific Rationality." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/rationality-historicist/> Accessed 2026-10-05.

- Nickles, Thomas. 2019. "The Crowbar Model of Method and Its Implications." *Theoria: An International Journal for Theory, History and Foundations of Science* 34(3): 357–372. <https://doi.org/10.1387/theoria.19070>
- Nielsen, Michael and Kanjun Qiu. 2022. "A Vision of Metascience: An Engine of Improvement for the Social Processes of Science." Science++, <https://scienceplusplus.org/metascience> Accessed 2026-10-07.
- Niiniluoto, Ilkka. 2024. "Scientific Progress." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/scientific-progress/> Accessed 2026-10-29.
- Niwa, Hitoshi. 2016. "Investigation of the cellular reprogramming phenomenon referred to as stimulus-triggered acquisition of pluripotency (STAP)." *Scientific Reports* 6(1): 28003. <https://doi.org/10.1038/srep28003>
- Norton, John D. 2026. "The Terminal Obsession." Chapter 11 of an author-posted book manuscript on the author's University of Pittsburgh page, revised through 31 May 2026; page numbers are those of the PDF. https://sites.pitt.edu/~jdnorton/papers/small-e_empiricism/11_Terminal_obsession.pdf
- NOVA. 1997. "The Proof." WGBH/PBS. Transcript of the television programme. <https://www.pbs.org/wgbh/nova/transcripts/2414proof.html> Accessed 2026-09-30.
- Novak, Joseph D. and Alberto J. Cañas. 2008. *The Theory Underlying Concept Maps and How to Construct and Use Them*. Florida Institute for Human and Machine Cognition. <https://cmap.ihmc.us/docs/pdf/TheoryUnderlyingConceptMaps.pdf> Accessed 2026-11-01.
- Nyholm, Sven. 2018. "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci." *Science and Engineering Ethics* 24(4): 1201–1219. <https://doi.org/10.1007/s11948-017-9943-x>
- Nyrup, Rune. 2015. "How Explanatory Reasoning Justifies Pursuit: A Peircean View of IBE." *Philosophy of Science* 82(5): 749–760. Section numbers cited in the text refer to the author's preprint (PSA 2014 paper), PhilSci-Archive 11126, <https://philsci-archive.pitt.edu/11126/>; the published version was revised. <https://doi.org/10.1086/683262>
- O'Connor, J. J. and E. F. Robertson. 1996. "The Four Colour Theorem." *MacTutor History of Mathematics Archive*. https://mathshistory.st-andrews.ac.uk/HistTopics/The_four_colour_theorem Accessed 2026-09-30.
- O'Neill, Onora. 2002. "Trust and Transparency." *A Question of Trust*, BBC Reith Lectures 2002, Lecture 4. Transcript. <http://www.bbc.co.uk/radio4/reith2002/lecture4.shtml> Accessed 2026-10-19.
- Oberheim, Eric and Paul Hoyningen-Huene. 2024. "The Incommensurability of Scientific Theories." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/incommensurability/> Accessed 2026-10-29.
- Ochiai, Koji, Yuya Tahara-Arai, Akari Kato, Kazunari Kaizu, Hirokazu Kariyazaki, Makoto Umeno, Koichi Takahashi, Genki N. Kanda, and Haruka Ozaki. 2025. "Automating care by self-maintainability for full laboratory automation." *Digital Discovery* 4(9): 2285–2297. <https://doi.org/10.1039/D5DD000151J>
- Ochigame, Rodrigo. 2024. "Automated Mathematics and the Reconfiguration of Proof and Labor." *Bulletin of the American Mathematical Society* 61(3): 423–437. Section numbers cited in the text refer to arXiv:2309.11457v1. <https://arxiv.org/abs/2309.11457v1>
- Oddie, Graham and Gustavo Cevolani. 2022. "Truthlikeness." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/truthlikeness/> Accessed 2026-10-08.
- OECD. 2015. *Frascati Manual 2015*. OECD Publishing. <https://doi.org/10.1787/9789264239012-en>
- Olazaran, Mikel. 1991. "Historical Sociology of Neural Network Research." PhD thesis, University of Edinburgh. Catalogued under the name Jose Miguel Olazaran Rodriguez; cited by the thesis's printed page numbers. <https://hdl.handle.net/1842/20075> Accessed 2026-10-14.
- Olson, Eric T. 2023. "Personal Identity." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/identity-personal/> Accessed 2026-10-08.
- Open Source Mycetoma. 2022. "Biological evaluation 2022 (issue #71)." *OpenSourceMycetoma/Series-1-Fenarimols*, *GitHub*. <https://github.com/OpenSourceMycetoma/Series-1-Fenarimols/issues/71> Accessed 2026-09-30.
- OpenAI. 2023. *GPT-4 Technical Report*. <https://arxiv.org/abs/2303.08774v6>
- OpenAI. 2025. "Multi-Modal Gesture Recognition Data Leakage (issue #77)." *openai/mle-bench*, *GitHub*. Maintainer reply, 8 October 2025. <https://github.com/openai/mle-bench/issues/77> Accessed 2026-09-29.
- OpenAI. 2026a. *Ten Advances in Mathematics and Theoretical Computer Science*. OpenAI. <https://cdn.openai.com/pdf/ten-proofs-oai.pdf> Accessed 2026-10-28.
- OpenAI. 2026b. "ten-proofs." *openai/ten-proofs*, *GitHub*. <https://github.com/openai/ten-proofs> Accessed 2026-10-28.
- OpenAI. 2026c. *Finite Time Blowup for Navier-Stokes*. OpenAI. <https://cdn.openai.com/pdf/32d9f210-8b73-45e0-91bc-82a30aef8a9a/navier-stokes.pdf> Accessed 2026-10-28.

- OpenAI. 2026d. “math.” *openai/math*, *GitHub*. <https://github.com/openai/math> Accessed 2026-10-28.
- OpenAI. 2026e. “Update README.md (pull request #143).” *openai/mle-bench*, *GitHub*. <https://github.com/openai/mle-bench/pull/143> Accessed 2026-09-29.
- OPERA Collaboration. 2011. “Measurement of the neutrino velocity with the OPERA detector in the CNGS beam.” *arXiv*. Preprint, version 1; cited by PDF page. <https://arxiv.org/abs/1109.4897v1> Accessed 2026-09-24.
- Oreskes, Naomi, Kristin Shrader-Frechette, and Kenneth Belitz. 1994. “Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences.” *Science* 263(5147): 641–646. Cited without page numbers from a text-only copy of the article (a database export) on the website of co-author Kristin Shrader-Frechette, https://sites.nd.edu/kristin-shrader-frechette/files/2025/01/Verification_validation_and_co.pdf. <https://doi.org/10.1126/science.263.5147.641>
- Ortmann, Jakob. 2025. “Of opaque oracles: epistemic dependence on AI in science poses no novel problems for social epistemology.” *Synthese* 205(2): 80. <https://doi.org/10.1007/s11229-025-04930-x>
- Ostrom, Elinor. 2009. “Beyond Markets and States: Polycentric Governance of Complex Economic Systems.” Prize Lecture, 8 December 2009. In *Les Prix Nobel 2009*, 408–444. The Nobel Foundation. https://www.nobelprize.org/uploads/2018/06/ostrom_lecture.pdf Accessed 2026-10-18.
- Otsuka, Jun. 2018. “Inga suiron o sasaeru sonzairon to ninshikiron [Ontology and Epistemology Underlying Causal Inference].” In *Proceedings of the Annual Meeting of the Behaviormetric Society of Japan* 46, 188–189. Behaviormetric Society of Japan. In Japanese; session SB17-1. The text paraphrases it in this book's translation. https://doi.org/10.20742/pbsj.46.0_188
- Otsuka, Jun. 2021. “Tokeigaku wa naze tetsugaku no mondai ni naru no ka [Why Does Statistics Matter to Philosophy?].” *Tetsugaku Kenkyu* 606: 1–24. In Japanese. The author's English version (“a minimally edited machine translation”) is on PhilSci-Archive, <https://philsci-archive.pitt.edu/22945/>. Page numbers cited in the text are those of the Japanese article; the text paraphrases it in this book's translation. https://doi.org/10.14989/JPS_606_1
- Otsuka, Jun. 2023. *Thinking About Statistics: The Philosophical Foundations*. Routledge. <https://doi.org/10.4324/9781003319061>
- Otsuka, Jun. 2026. “What Does It Mean to Understand AI?” *Harvard Data Science Review* 8(2). <https://doi.org/10.1162/99608f92.cdf80349>
- Oudeyer, Pierre-Yves and Frédéric Kaplan. 2007. “What is intrinsic motivation? A typology of computational approaches.” *Frontiers in Neurorobotics* 1. <https://doi.org/10.3389/neuro.12.006.2007>
- O’Loughlin, Ryan. 2023. “Diagnosing errors in climate model intercomparisons.” *European Journal for Philosophy of Science* 13(2): 20. Section numbers cited in the text refer to the accepted manuscript, PhilSci-Archive 21947, <https://philsci-archive.pitt.edu/21947/>. <https://doi.org/10.1007/s13194-023-00522-z>
- Pagin, Peter and Neri Marsili. 2025. “Assertion.” *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/assertion/> Accessed 2026-10-08.
- Pamuk, Zeynep. 2019. “Justifying Public Funding for Science.” *British Journal of Political Science* 49(1): 1–16. <https://doi.org/10.1017/S0007123418000431>
- Pan, Wenbo, Zhichao Liu, Shujie Liu, Jingying Zeng, Chin-Yew Lin, Xianfeng Tang, Yan Lu, Qi He, and Xiaohua Jia. 2026. “The Tasteful Agent: Measuring and Improving Taste in Long-Horizon Tasks.” *arXiv*. Preprint, version 2; page numbers cited in the text refer to this version. <https://arxiv.org/abs/2609.25804v2>
- Parasuraman, Raja, Thomas B. Sheridan, and Christopher D. Wickens. 2000. “A Model for Types and Levels of Human Interaction with Automation.” *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 30(3): 286–297. The citation by “abstract” rests on the authors' abstract, which is open on PubMed, <https://pubmed.ncbi.nlm.nih.gov/11760769/>. <https://doi.org/10.1109/3468.844354>
- Parfit, Derek. 1984. *Reasons and Persons*. Clarendon Press. <https://doi.org/10.1093/019824908x.001.0001>
- Parker, Wendy S. 2020. “Model Evaluation: An Adequacy-for-Purpose View.” *Philosophy of Science* 87(3): 457–477. <https://doi.org/10.1086/708691>
- Pathak, Deepak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. “Curiosity-driven Exploration by Self-supervised Prediction.” In *Proceedings of the 34th International Conference on Machine Learning (PMLR 70)*. Page numbers cited in the text refer to arXiv:1705.05363v1, which carries the proceedings header. <https://arxiv.org/abs/1705.05363v1>
- Paul, Elliot Samuel and Dustin Stokes. 2023. “Creativity.” *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/creativity/> Accessed 2026-09-29.
- Pauling, Linus. 1932–1941. “Research Notebook 09.” *Linus Pauling Research Notebooks*. Archival notebook; cited by the archive's page numbers. Courtesy Ava Helen and Linus Pauling Papers, Special Collections & Archives Research Center, Oregon State University Libraries. <https://scarc.library.oregonstate.edu/coll/pauling/rnb/09/rnb09.html> Accessed 2026-10-05.
- Pauling, Linus and Robert B. Corey. 1953. “A Proposed Structure For The Nucleic Acids.” *Proceedings of the National Academy of Sciences* 39(2): 84–97. <https://doi.org/10.1073/pnas.39.2.84>

Pauling, Linus. 1953. "Letter to Max Delbrück, April 20, 1953." *Linus Pauling and the Race for DNA: A Documentary History*. Oregon State University Special Collections & Archives Research Center. <https://scarc.library.oregonstate.edu/coll/pauling/dna/corr/sci9.001.39-lp-delbruck-19530420.html> Accessed 2026-09-29.

Pauly, Daniel Uptmoor. 2022. "Mind Meets the Body Searching for Itself—Embodied, Social, Perceptual, Ethical, Linguistic and Temporal Aspects of Consciousness." Universidade do Vale do Rio dos Sinos. PhD thesis in Philosophy, deposited in the UNISINOS institutional repository; page numbers are those printed in the thesis. https://repositorio.jesuita.org.br/bitstream/handle/UNISINOS/12184/Daniel%20Uptmoor%20Pauly-DT_PROTEGIDO.pdf?sequence=1&isAllowed=y

Peirce, Charles S. 1877. "Illustrations of the Logic of Science. First Paper: The Fixation of Belief." *Popular Science Monthly* 12: 1–15. Wikisource transcription with page anchors. https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_12/November_1877/Illustrations_of_the_Logic_of_Science_I Accessed 2026-10-04.

Pelc, Andrzej. 2014. "Why Do We Believe Theorems?" *arXiv*. Significantly extended version of the paper of the same title published in *Philosophia Mathematica* 17 (2009): 84–94; page numbers cited in the text refer to arXiv:1411.4857v1 (17 November 2014). <https://arxiv.org/abs/1411.4857>

Penadés, José R., Juraj Gottweis, Lingchen He, Jonasz B. Patkowski, Alexander Daryin, Wei-Hung Weng, Tao Tu, Anil Palepu, Artiom Myaskovsky, Annalisa Pawlosky, Vivek Natarajan, Alan Karthikesalingam, and Tiago R.D. Costa. 2025. "AI mirrors experimental science to uncover a mechanism of gene transfer crucial to bacterial evolution." *Cell* 188(23): 6654–6665.e2. <https://doi.org/10.1016/j.cell.2025.08.018>

Peter, Fabienne. 2023. "Political Legitimacy." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/legitimacy/> Accessed 2026-09-29.

Peters, Uwe. 2024a. "Science Based on Artificial Intelligence Need not Pose a Social Epistemological Problem." *Social Epistemology Review and Reply Collective* 13(1): 58–66. <https://wp.me/p1Bfg0-8vp>

Peters, Uwe. 2024b. "Living with Uncertainty: Full Transparency of AI Is Not Needed for Epistemic Trust in AI-Based Science." *Social Epistemology Review and Reply Collective* 13(6): 8–15. Read as the web version, which has no page numbers. <https://wp.me/p1Bfg0-8Si>

Pettit, Philip. 1996. "Freedom as Antipower." *Ethics* 106(3): 576–604. Cited from the JSTOR copy of the article that Philip Pettit posted on his Princeton web page, as captured by the Internet Archive's Wayback Machine on 5 July 2013, https://web.archive.org/web/20130705090436/http://www.princeton.edu/~ppettit/papers/FreedomasAntipower_Ethics_1996.pdf, which carries the journal pagination. <https://doi.org/10.1086/233648>

Pettit, Philip. 1997. "Republican Political Theory." In *Political Theory: Tradition and Diversity*, edited by Andrew Vincent, 112–131. Cambridge University Press. Cited from the scan of the chapter that Philip Pettit posted on his Princeton web page, as captured by the Internet Archive's Wayback Machine on 17 September 2006, which carries the book's pagination. https://web.archive.org/web/20060917093942/http://www.princeton.edu/~ppettit/papers/RepublicanPoliticalTheory_PoliticalTheory_1997.pdf

Pettit, Philip. 1998. "Republican Theory and Political Trust." In *Trust and Governance*, edited by Valerie Braithwaite and Margaret Levi, 295–314. Russell Sage Foundation. Cited from the scan of the chapter that Philip Pettit posted on his Princeton web page, as captured by the Internet Archive's Wayback Machine on 17 September 2006, which carries the book's pagination. https://web.archive.org/web/20060917093353/http://www.princeton.edu/~ppettit/papers/RepublicanTheoryandPoliticalTrust_TrustandGovernance_

Pettit, Philip. 1999. "Republican Freedom and Contestatory Democratization." In *Democracy's Value*, edited by Ian Shapiro and Casiano Hacker-Cordón, 163–190. Cambridge University Press. Cited from the scan of the chapter that Philip Pettit posted on his Princeton web page, as captured by the Internet Archive's Wayback Machine on 25 January 2012, which carries the book's pagination. https://web.archive.org/web/20120125145240/http://www.princeton.edu/~ppettit/papers/RepublicanFreedom_DemocracysEdges_1999.pdf

Pettit, Philip. 2000. "Democracy, Electoral and Contestatory." In *Designing Democratic Institutions*, edited by Ian Shapiro and Stephen Macedo, 105–144. New York University Press. Cited from the scan of the chapter that Philip Pettit posted on his Princeton web page, as captured by the Internet Archive's Wayback Machine on 1 April 2011, which carries the book's pagination. https://web.archive.org/web/20110401205022/http://www.princeton.edu/~ppettit/papers/Democracy_Electoral_and_Contestatory_NOMOS_2000.pdf

Pettit, Philip. 2008. "Freedom and Probability: A Comment on Goodin and Jackson." *Philosophy & Public Affairs* 36(2): 206–220. Cited from the publisher's offprint that Philip Pettit posted on his Princeton web page, as captured by the Internet Archive's Wayback Machine on 25 January 2012, <https://web.archive.org/web/20120125133705/http://www.princeton.edu/~ppettit/papers/2008/Pettit%20Freedom%20and%20Probability%20Offprin> which carries the journal pagination. <https://doi.org/10.1111/j.1088-4963.2007.00133.x>

Pinder, Mark. 2020. "On Strawson's critique of explication as a method in philosophy." *Synthese* 197(3): 955–981. <https://doi.org/10.1007/s11229-017-1614-6>

- Pirozelli, Paulo. 2020. "Thomas Kuhn's Philosophy of Language." *Trans/Form/Ação* 43(special issue): 345–372. <https://www.scielo.br/j/trans/a/dxYdhRS8PQdPNV8cVMV9nmX/?lang=en>
- Plato. n.d. *Meno*. Translated by Benjamin Jowett. Project Gutenberg. Cited in Benjamin Jowett's translation from the Project Gutenberg text; 97e–98a is the standard Stephanus reference. <https://www.gutenberg.org/ebooks/1643> Accessed 2026-11-01.
- Poincaré, Henri. [1908] 1914. *Science and Method*. Thomas Nelson and Sons. Translated by Francis Maitland; first published as *Science et méthode* (1908). <https://henripoincarepapers.univ-nantes.fr/chp/hp-pdf/hp1914sm.pdf> Accessed 2026-09-30.
- Polanyi, Michael. 1962. "The Republic of Science: Its Political and Economic Theory." *Minerva* 1(1): 54–73. Reprinted in *Minerva* 38(1): 1–21 (2000); page numbers cited in the text refer to the reprint. The link goes to the full text of the article on the Polanyi Society website, posted with the permission of Kluwer Academic Publishers and John C. Polanyi, <https://polanyisociety.org/primary-resources-the-republic-of-science/>, which has no page numbers. <https://www.jstor.org/stable/41821153> Accessed 2026-09-07.
- Polymath, D. H. J. 2012. "A New Proof of the Density Hales–Jewett Theorem." *Annals of Mathematics* 175(3): 1283–1327. Page numbers cited in the text refer to the arXiv version linked here. <https://arxiv.org/abs/0910.3926>
- Popper, Karl. 1978. "Three Worlds." *The Tanner Lectures on Human Values*. The Tanner Lecture on Human Values delivered at the University of Michigan on 7 April 1978; cited by section number from the text on the Tanner Lectures website. <https://tannerlectures.org/documents/three-worlds/> Accessed 2026-10-29.
- Popper, Karl R. [1972] 1979. *Objective Knowledge: An Evolutionary Approach*. Revised edition. Clarendon Press. <https://archive.org/details/objectiveknowled00popp> Accessed 2026-10-08.
- Porter, Theodore M. 1995. *Trust in Numbers*. Princeton University Press. The citations are to the Introduction, "Cultures of Objectivity" (pp. 3–8) in the publisher's sample of the 2020 paperback reprint, which keeps the pagination of the 1995 edition, <https://pup-assets.s3.amazonaws.com/public/onix/images/9780691208411/9780691208411.pdf>. <https://doi.org/10.1515/9781400821617>
- Porter, Theodore M. 1996. "Statistics, Social Science, and the Culture of Objectivity." *Österreichische Zeitschrift für Geschichtswissenschaften* 7(2): 177–191. <https://doi.org/10.25365/oezg-1996-7-2-2>
- Porter, Theodore M. 2012. "Funny Numbers." *Culture Unbound* 4(4): 585–598. The issue is dated 2012 on the article's pages; Crossref records online publication on 24 January 2013. <https://doi.org/10.3384/cu.2000.1525.124585>
- Potochnik, Angela. 2015. "The diverse aims of science." *Studies in History and Philosophy of Science Part A* 53: 71–80. Page numbers cited in the text refer to the author's manuscript, PhilPapers, <https://philpapers.org/rec/POTTDA>; the published pagination differs. <https://doi.org/10.1016/j.shpsa.2015.05.008>
- Potochnik, Angela. 2016. "Scientific Explanation: Putting Communication First." *Philosophy of Science* 83(5): 721–732. Page numbers cited in the text refer to the author's preprint, PhilSci-Archive, <https://philsci-archive.pitt.edu/19008/>; the published pagination differs. <https://doi.org/10.1086/687858>
- Potochnik, Angela. 2020. "Idealization and Many Aims." *Philosophy of Science* 87(5): 933–943. Page numbers cited in the text refer to the author's submitted version ("Forthcoming in *Philosophy of Science: Proceedings of PSA2018*", 18 pp.), PhilPapers, <https://philpapers.org/archive/POTIAM.pdf>; the published pagination differs. <https://doi.org/10.1086/710622>
- Power, Maddy, Ruth Patrick, and Kayleigh Garthwaite. 2024. "'I Feel Like I am Part of Something Bigger Than Me': Methodological Reflections From Longitudinal Online Participatory Research." *International Review of Qualitative Research* 18(2): 179–198. Page numbers cited in the text refer to the online-first version (pp. 1–20). <https://doi.org/10.1177/19408447241260448>
- Preston, Beth. 2022. "Artifact." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/artifact/> Accessed 2026-10-08.
- Pritchard, Duncan. 2008. "Knowing the Answer, Understanding and Epistemic Value." *Grazer Philosophische Studien* 77(1): 325–339. Section numbers cited in the text refer to the author's peer-reviewed version ("Forthcoming in *Grazer Philosophische Studien*"), Edinburgh Research Explorer, <https://www.research.ed.ac.uk/en/publications/522f8eba-15b2-46d0-8019-4647e795642c>. <https://doi.org/10.1163/18756735-90000852>
- Pritchard, Duncan. 2016. "Seeing It for Oneself: Perceptual Knowledge, Understanding, and Intellectual Autonomy." *Episteme* 13(1): 29–42. Page numbers cited in the text refer to the peer-reviewed author's version, Edinburgh Research Explorer, <https://www.research.ed.ac.uk/en/publications/c5e73911-ad2f-415c-92ec-59e823db6658>; the published pagination differs. <https://doi.org/10.1017/epi.2015.59>
- Pritchard, Duncan. 2021. "On Hinge Epistemology." *Sképsis: Revista de Filosofia* 12(23): 1–17. <https://www.pure.ed.ac.uk/ws/portalfiles/portal/201742368/PritchardD2021SONHingeEpistemology.pdf>
- Qian, Yifan, Zhe Wen, Alexander C. Furnas, Yue Bai, Erzhao Shao, and Dashun Wang. 2026. "The rise of large language models and the direction and impact of US federal research funding." *Proceedings of the National Academy of Sciences* 123(33): e2601439123. <https://doi.org/10.1073/pnas.2601439123>

- Queloz, Matthieu. 2022. "Function-Based Conceptual Engineering and the Authority Problem." *Mind* 131(524): 1247–1278. <https://doi.org/10.1093/mind/fzac028>
- Quine, W. V. 1951. "Two Dogmas of Empiricism." *The Philosophical Review* 60(1): 20–43. Cited from the scan of the January 1951 issue of *The Philosophical Review* in the Internet Archive (Serials in Microfilm), https://archive.org/details/sim_philosophical-review_1951-01_60_1, which carries the journal pagination. <https://doi.org/10.2307/2181906>
- Raja, Vicente and Guilherme Sanches de Oliveira. 2024. "Two species of realism." *European Journal for Philosophy of Science* 14(3): 35. Page numbers cited in the text refer to the published PDF's numbered pages. <https://link.springer.com/article/10.1007/s13194-024-00596-3>
- Raji, Deborah, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. "AI and the Everything in the Whole Wide World Benchmark." *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1. <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html> Accessed 2026-09-07.
- Rathkopf, Charles. 2026. "From Hallucination to Reliability: Generative Modeling and the Structure of Scientific Inference." *arXiv*. Preprint (arXiv:2504.08526v3, 26 August 2026; earlier versions titled "Hallucination, reliability, and the role of generative AI in science"); page numbers cited in the text refer to version 3. <https://arxiv.org/abs/2504.08526> Accessed 2026-10-17.
- Ratti, Emanuele. 2020. "Phronesis and Automated Science: The Case of Machine Learning and Biology." In *A Critical Reflection on Automated Science: Will Science Remain Human?*, edited by Marta Bertolaso and Fabio Sterpetti, 157–172. Springer. Page numbers cited in the text refer to the author's penultimate draft, PhilSci-Archive 15770, <https://philsci-archive.pitt.edu/15770/>; the published pagination differs. https://doi.org/10.1007/978-3-030-25001-0_8
- Ratti, Emanuele. 2026. "Epistemic Control and the Normativity of Machine Learning-Based Science." *arXiv*. Preprint of a chapter to appear in *The Role of AI in Science: Epistemological and Methodological Studies*, edited by David Barack, André Curtis-Trudel, and Darrell Rowbottom (Routledge); page numbers cited in the text refer to the preprint. <https://arxiv.org/abs/2601.11202>
- Rav, Yehuda. 1989. "Philosophical Problems of Mathematics in the Light of Evolutionary Epistemology." *Philosophica* 43: 49–78. <https://doi.org/10.21825/philosophica.82453>
- Ravetz, Jerome. 2016. "Democratizing Science in an Age of Uncertainty." Great Transition Initiative. Interview by Allen White, June 2016. <https://greattransition.org/publication/democratizing-science/> Accessed 2026-10-31.
- Ravetz, Jerome R. 1993. "The Complementarity of Knowledge and Ignorance in Science." *Natures Sciences Sociétés* 1(4): 348–350. <https://doi.org/10.1051/nss/19930104348>
- Rawls, John. 1958. "Justice as Fairness." *The Philosophical Review* 67(2): 164–194. https://archive.org/details/sim_philosophical-review_1958-04_67_2 Accessed 2026-11-01.
- Reamer, Robert. 2024. "Markets and Metis: Reading Hayek with Scott." *Critical Review* 36(1-2): 162–182. <https://doi.org/10.1080/08913811.2024.2354648>
- Reason, Peter and Hilary Bradbury. 2001. "Introduction: Inquiry and Participation in Search of a World Worthy of Human Aspiration." In *Handbook of Action Research: Participative Inquiry and Practice*. SAGE. https://www.peterreason.net/wp-content/uploads/Handbook_Introduction.pdf
- Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. "Do ImageNet Classifiers Generalize to ImageNet? Supplementary material." In *Proceedings of the 36th International Conference on Machine Learning (PMLR 97)*. <https://proceedings.mlr.press/v97/recht19a/recht19a-supp.pdf> Accessed 2026-09-29.
- Reddy, Raj. 2021a. "Oral History of Raj Reddy, Part 1." Computer History Museum. <https://archive.computerhistory.org/resources/access/text/2021/06/102792182-05-01-acc.pdf> Accessed 2026-09-29.
- Reddy, Raj. 2021b. "Oral History of Raj Reddy, Part 2 of 2." Computer History Museum. <https://archive.computerhistory.org/resources/access/text/2022/02/102792298-05-01-acc.pdf> Accessed 2026-09-29.
- Redfield, Rosemary J. 2011a. "Maybe it's the water? Or the tubes?" *RRResearch* (research blog), 27 June 2011. <http://rrresearch.fieldofscience.com/2011/06/maybe-its-water-or-tubes.html> Accessed 2026-09-24.
- Redfield, Rosemary J. 2011b. "It's not the water, nor the tubes, nor the parafilm..." *RRResearch* (research blog), 30 June 2011. <http://rrresearch.fieldofscience.com/2011/06/its-not-water-nor-tubes-nor-parafilm.html> Accessed 2026-09-24.
- Redfield, Rosemary J. 2011c. "Two mistakes discovered." *RRResearch* (research blog), 14 July 2011. <http://rrresearch.fieldofscience.com/2011/07/two-mistakes-discovered.html> Accessed 2026-09-24.
- Redfield, Rosemary J. 2011d. "Growth of GFAJ-1 in 40 mM arsenate (not)." *RRResearch* (research blog), 7 August 2011; the post was updated on 5 November 2011. <http://rrresearch.fieldofscience.com/2011/08/growth-of-gfaj-1-in-40-mm-arsenate-not.html> Accessed 2026-09-24.
- Reichenbach, Hans. 1938. *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. University of Chicago Press. Cited from the first Phoenix edition (1961), a later printing of the 1938 publication; Internet Archive digitization.

<https://archive.org/details/experiencepredic00reic> Accessed 2026-09-30.

Reijers, Wessel, Iris Wuisman, Morshed Mannan, Primavera De Filippi, Christopher Wray, Vienna Rae-Looi, Angela Cubillos Vélez, and Liav Orgad. 2021. "Now the Code Runs Itself: On-Chain and Off-Chain Governance of Blockchain Technologies." *Topoi* 40(4): 821–831. Published online 2018. Page numbers cited in the text refer to the authors' manuscript, HAL hal-02046663, <https://hal.science/hal-02046663>. <https://doi.org/10.1007/s11245-018-9626-5>

Reinhart, Carmen M. and Kenneth S. Rogoff. 2013. "Errata: 'Growth in a Time of Debt'." 5 May 2013. https://carmenreinhardt.com/wp-content/uploads/2020/04/36_data.pdf Accessed 2026-09-24.

Reiss, Julian and Philip Kitcher. 2009. "Biomedical Research, Neglected Diseases, and Well-Ordered Science." *Theoria: An International Journal for Theory, History and Foundations of Science* 24(3): 263–282. <https://doi.org/10.1387/theoria.696>

Reiss, Julian and Jan Sprenger. 2020. "Scientific Objectivity." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/scientific-objectivity/> Accessed 2026-10-18.

Renzo, Massimo and Leslie Green. 2022. "Legal Obligation and Authority." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/legal-obligation/> Accessed 2026-10-08.

Research England; Scottish Funding Council; Higher Education Funding Council for Wales; Department for the Economy, Northern Ireland. 2019. *Guidance on Submissions: REF 2021*. https://2021.ref.ac.uk/media/1447/ref-2019_01-guidance-on-submissions.pdf

Resnik, David B., Mohammad Hosseini, and Rico Hauswald. 2026. "Autonomous artificial intelligence, scientific research, and human values." *AI and Ethics* 6(1): 141. Section numbers cited in the text are those of the author manuscript (PMC12948163). <https://doi.org/10.1007/s43681-025-00908-0>

Rheinberger, Hans-Jörg. 1994. "Experimental Systems: Historiality, Narration, and Deconstruction." *Science in Context* 7(1): 65–81. <https://doi.org/10.1017/s0269889700001599>

Rheinberger, Hans-Jörg. 2011. "Consistency from the Perspective of an Experimental Systems Approach to the Sciences and Their Epistemic Objects." *Manuscrito: Revista Internacional de Filosofia* 34(1): 307–321. <https://periodicos.sbu.unicamp.br/ojs/index.php/manuscrito/article/view/8642022> Accessed 2026-10-29.

RIKEN Investigation Committee (研究論文に関する調査委員会). 2014. 研究論文に関する調査報告書 [Investigation Report on Research Papers]. RIKEN. <http://www3.riken.jp/stap/j/c13document5.pdf> Accessed 2026-09-30.

Rodan, Simon. 2005. "Exploration and Exploitation Revisited: Extending March's Model of Mutual Learning." *Scandinavian Journal of Management* 21(4): 407–428. Page numbers cited in the text refer to the author's preprint, San José State University College of Business Working Paper OM-05-003, <https://www.sjsu.edu/cob/docs/OM-05-003.pdf>; the preprint's header gives the volume as 22. <https://www.sjsu.edu/cob/docs/OM-05-003.pdf>

Rolla, Giovanni. 2026. "The Embodiment Challenge for Artificial Intelligence." *Philosophy & Technology* 39(3): 125. <https://doi.org/10.1007/s13347-026-01139-9>

Romeijn, Jan-Willem. 2025. "Philosophy of Statistics." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/statistics/> Accessed 2026-10-29.

Rosenstock, Sarita, Justin Bruner, and Cailin O'Connor. 2017. "In Epistemic Networks, Is Less Really More?" *Philosophy of Science* 84(2): 234–252. Cited from the publisher's PDF on the website of co-author Cailin O'Connor, https://cailinoconnor.com/wp-content/uploads/2015/03/In_Epistemic_Networks_Is_Less_Really_Mor-FINAL-VERSION.pdf, which carries the journal pagination. <https://doi.org/10.1086/690717>

Ross, Lauren and James Woodward. 2023. "Causal Approaches to Scientific Explanation." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/causal-explanation-science/> Accessed 2026-10-29.

Ross, Lewis D. 2021. "How Intellectual Communities Progress." *Episteme* 18(4): 738–756. Cited without page numbers from the accepted version, LSE Research Online, <https://researchonline.lse.ac.uk/id/eprint/103972/>. <https://doi.org/10.1017/epi.2020.2>

Roush, Sherrilyn. 2017. "Epistemic Self-Doubt." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/epistemic-self-doubt/> Accessed 2026-10-08.

Rowbottom, Darrell P. 2014. "Aimless science." *Synthese* 191(6): 1211–1221. Page numbers cited in the text refer to the version conditionally accepted for publication, Lingnan University Digital Commons, https://commons.ln.edu.hk/sw_master/1952/; the published pagination differs. <https://doi.org/10.1007/s11229-013-0319-8>

Rowlands, Mark, Joe Lau, and Max Deutsch. 2020. "Externalism About the Mind." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/content-externalism/> Accessed 2026-10-04.

- Rubin, Karl and Alice Silverberg. 1994. "A Report on Wiles' Cambridge Lectures." *Bulletin of the American Mathematical Society* 31(1): 15–38. Page numbers cited in the text refer to the arXiv version linked here. <https://arxiv.org/abs/math/9407220>
- Ruhl, Cody R., Breanna L. Pasko, Haaris S. Khan, Lexy M. Kindt, Chelsea E. Stamm, Luis H. Franco, Connie C. Hsia, Min Zhou, Colton R. Davis, Tian Qin, Laurent Gautron, Michael D. Burton, Galo L. Mejia, Dhananjay K. Naik, Gregory Dussor, Theodore J. Price, and Michael U. Shiloh. 2020. "Mycobacterium tuberculosis Sulfolipid-1 Activates Nociceptive Neurons and Induces Cough." *Cell* 181(2): 293–305.e11. <https://doi.org/10.1016/j.cell.2020.02.026>
- Russell, Stuart. 2020a. "Three (plus) questions with Turing Lecturer Stuart Russell." The Alan Turing Institute. Interview published on the Institute's blog, 20 July 2020. <https://www.turing.ac.uk/blog/three-plus-questions-turing-lecturer-stuart-russell>
- Russell, Stuart and Steven Pinker. 2020b. "Steven Pinker and Stuart Russell on the Foundations, Benefits, and Possible Existential Threat of AI." Future of Life Institute. AI Alignment Podcast, interview by Lucas Perry, 15 June 2020; cited from the edited transcript on the Institute's page. <https://futureoflife.org/podcast/steven-pinker-and-stuart-russell-on-the-foundations-benefits-and-possible-existential-risk-of-ai/>
- Russell, Stuart. 2021. "Human-Compatible Artificial Intelligence." In *Human-Like Machine Intelligence*, edited by Stephen Muggleton and Nicholas Chater, 3–23. Oxford University Press. Page numbers cited in the text refer to the author's version on his Berkeley page. <https://people.eecs.berkeley.edu/~russell/papers/mi19book-hcai.pdf>
- Russell, Stuart. 2022. "If We Succeed." *Daedalus* 151(2): 43–57. Cited from the published article, posted on the author's website, which carries the journal pagination. <https://people.eecs.berkeley.edu/~russell/papers/russell-daedalus22-succeed.pdf> Accessed 2026-10-29.
- Ryan, Mark. 2020. "In AI We Trust: Ethics, Artificial Intelligence, and Reliability." *Science and Engineering Ethics* 26(5): 2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>
- Ryan, Richard M. and Edward L. Deci. 2000. "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being." *American Psychologist* 55(1): 68–78. Cited from the scan of the published article on the website of the Center for Self-Determination Theory, founded by the authors, https://selfdeterminationtheory.org/SDT/documents/2000_RyanDeci_SDT.pdf; it carries the journal pagination. <https://doi.org/10.1037/0003-066x.55.1.68>
- Sachdeva, Megha. 2013. "Encounter with March's Organizational Learning Model." *Review of Integrative Business and Economics Research* 2(2): 602–615. https://buscompress.com/uploads/3/4/9/8/34980536/riber_b13-231_602-615_.pdf
- Saji, Noburo. 2026. "Language Acquisition: Statistical Learning and Social Cognitive Skills." In *Symbol Emergence Systems*, edited by Tadahiro Taniguchi, 45–48. Springer. https://doi.org/10.1007/978-981-95-1327-7_8
- Salmon, Wesley C. 1990. "Rationality and Objectivity in Science or Tom Kuhn Meets Tom Bayes." In *Scientific Theories*, edited by C. Wade Savage, 175–204. Minnesota Studies in the Philosophy of Science 14. University of Minnesota Press. <https://hdl.handle.net/11299/185726>
- Sandifer, Ed. 2004. "V, E and F, Part 1." *How Euler Did It*. Mathematical Association of America (MAA Online). Monthly column for June 2004; cited from the PDF in the Euler Archive, whose page numbers (1–4) are cited in the text. <http://eulerarchive.maa.org/hedi/HEDI-2004-06.pdf> Accessed 2026-10-29.
- Santoni de Sio, Filippo and Jeroen van den Hoven. 2018. "Meaningful Human Control over Autonomous Systems: A Philosophical Account." *Frontiers in Robotics and AI* 5: 15. <https://doi.org/10.3389/frobt.2018.00015>
- Sauermann, Henry and Wesley M. Cohen. 2010. "What Makes Them Tick? Employee Motives and Firm Innovation." *Management Science* 56(12): 2134–2153. <https://doi.org/10.1287/mnsc.1100.1241>
- Sawyer, Sarah. 2014. "Contrastive Self-Knowledge." *Social Epistemology* 28(2): 139–152. The accepted manuscript was consulted. <https://doi.org/10.1080/02691728.2013.782586>
- Schickore, Jutta. 2022. "Scientific Discovery." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/scientific-discovery/> Accessed 2026-09-29.
- Schindler, Samuel. 2024. "Normal science: not uncritical or dogmatic." *Synthese* 203: 108. <https://link.springer.com/article/10.1007/s11229-024-04527-w>
- Schmidgall, Samuel, Xiaokai Zhu, Marian Shaw, Lin Yang, Valentin Liévin, Jingyun Yang, Yuchen Zhuang, Tim Strother, Alex Bijamov, Min Woo Sun, Anil Palepu, Justin Chen, David Steiner, Jacqueline Shreibati, Wei-Hung Weng, Yilin Zhao, Xingjian Hu, Nicholas Zahn, Sadhya Garg, Julia Kirby, Yuxiang Gan, Jiaoli Li, Divy Thakkar, Shekoofeh Azizi, David Racz, Juraj Gottweis, Vivek Natarajan, Chenglin Wu, Tal Danino, Keran Rong, Haozhe Wang, Benoit Schillings, Yong Cheng, Quoc V. Le, and Tao Tu. 2026. "Accelerating Scientific Research with Gemini in the Real-World." *arXiv*. Preprint. <https://arxiv.org/abs/2608.26701>
- Schneider, Nathan. 2019. "Decentralization: An Incomplete Ambition." *Journal of Cultural Economy* 12(4): 265–285. Page numbers cited in the text refer to the author's preprint on OSF, <https://osf.io/m7wyj/>. <https://doi.org/10.1080/17530350.2019.1589553>
- Scholze, Peter. 2020. "Liquid Tensor Experiment." *Xena Project*. <https://xenaproject.wordpress.com/2020/12/05/liquid-tensor-experiment> Accessed 2026-09-30.

- Scholze, Peter. 2021. "Half a year of the Liquid Tensor Experiment: Amazing developments." *Xena Project*. <https://xenaproject.wordpress.com/2021/06/05/half-a-year-of-the-liquid-tensor-experiment-amazing-developments> Accessed 2026-09-07.
- Schrape, Niklas. 2015. "«Portal» als Experimentalsystem." In *»The cake is a lie!« Polyperspektivische Betrachtungen des Computerspiels am Beispiel von »Portal«*, edited by Thomas Hensel, Britta Neitzel, and Rolf F. Nohr, 157–177. LIT. Cited from the official online edition, identical to the first print edition; page numbers are those printed. https://nuetzliche-bilder.de/wp-content/uploads/2020/10/Hensel_Neitzel_Nohr_Portal_Onlienausgabe.pdf
- Schroeder, Mark. 2025. "Value Theory." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/value-theory/> Accessed 2026-09-07.
- Schulte, Oliver. 2026. "Formal Learning Theory." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/learning-formal/> Accessed 2026-10-08.
- Schut, Lisa, Nenad Tomašev, Thomas McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. 2025. "Bridging the human–AI knowledge gap through concept discovery and transfer in AlphaZero." *Proceedings of the National Academy of Sciences* 122(13): e2406675122. <https://doi.org/10.1073/pnas.2406675122>
- Schwartz, Matthew. 2026. "Claude-shaped science." *Anthropic*. <https://www.anthropic.com/research/claude-shaped-science> Accessed 2026-10-27.
- Schwitzgebel, Eric and Mara Garza. 2020. "Designing AI with Rights, Consciousness, Self-Respect, and Freedom." In *Ethics of Artificial Intelligence*, edited by S. Matthew Liao. Oxford University Press. Page numbers cited in the text refer to the authors' manuscript of 4 June 2018. <https://faculty.ucr.edu/~eschwitz/SchwitzPapers/AIRights2-180604.pdf>
- Schwitzgebel, Eric. 2023. "Belief." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/belief/> Accessed 2026-09-07.
- Scrivener, Stephen. 2002. "The art object does not embody a form of knowledge." *Working Papers in Art and Design* 2. Page numbers cited in the text refer to the PDF at UAL Research Online, <https://ualresearchonline.arts.ac.uk/id/eprint/783/>. <https://doi.org/10.58129/mxf9-e323>
- Secretaria de Saúde do Estado de Pernambuco. 2015. *Nota técnica SEVS/DGCDA n.º 44/15: Atualização sobre o aumento da ocorrência de microcefalia (anomalia congênita) em nascidos vivos no Estado de Pernambuco*. Secretaria de Saúde do Estado de Pernambuco. https://portalcievs.saude.pe.gov.br/docs/Nota%20Tecnica%20%20DGCDA%2044%20_%202015_V2.pdf Accessed 2026-09-30.
- Shah, Rohin, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. 2022. "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals." *arXiv*. Preprint, version 2; page numbers cited in the text refer to this version. <https://arxiv.org/abs/2210.01790v2>
- Shan, Yafeng. 2019. "A New Functional Approach to Scientific Progress." *Philosophy of Science* 86(4): 739–758. Section numbers cited in the text refer to the author's manuscript, PhilArchive, <https://philarchive.org/rec/SHAANF-2>; they are the same as in the published article. <https://doi.org/10.1086/704980>
- Shapiro, Lawrence and Shannon Spaulding. 2021. "Embodied Cognition." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/embodied-cognition/> Accessed 2026-10-04.
- Shaw, Jamie. 2018. "A Pluralism Worth Having: Feyerabend's Well-Ordered Science." The University of Western Ontario. PhD thesis, Western University; page numbers are those printed in the thesis. <https://hdl.handle.net/20.500.14721/33235> Accessed 2026-10-30.
- Sheridan, Thomas B. and William L. Verplank. 1978. *Human and Computer Control of Undersea Teleoperators*. Technical report, Man-Machine Systems Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology. <https://apps.dtic.mil/sti/pdfs/ADA057655.pdf>
- Shulman, Carl and Nick Bostrom. 2021. "Sharing the World with Digital Minds." In *Rethinking Moral Status*, edited by Steve Clarke, Hazem Zohny, and Julian Savulescu. Oxford University Press. Page numbers cited in the text refer to the authors' draft (version 1.8, 2020). <https://nickbostrom.com/papers/digital-minds.pdf> Accessed 2026-10-07.
- Si, Chenglei, Diyi Yang, and Tatsunori Hashimoto. 2025. "Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers." *International Conference on Learning Representations* 2025: 94003. https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea94957d81b1c1caf87ef5319fa6b467-Abstract-Conference.html Accessed 2026-10-27.
- Si, Chenglei, Tatsunori Hashimoto, and Diyi Yang. 2026. "The Ideation-Execution Gap: Execution Outcomes of LLM-Generated versus Human Research Ideas." *International Conference on Learning Representations* 2026: 74195. https://proceedings.iclr.cc/paper_files/paper/2026/hash/78d84ad7c99601a422301e4bdc4274ce-Abstract-Conference.html Accessed 2026-10-27.
- Siegel, Zachary S., Sayash Kapoor, Nitya Nadgir, Benedikt Stroebel, and Arvind Narayanan. 2025. "CORE-Bench: Fostering the Credibility of Published Research Through a Computational Reproducibility Agent Benchmark." *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=BsMMc4MEGS> Accessed 2026-09-07.

- Silver, David and Richard S. Sutton. 2025. "Welcome to the Era of Experience." Preprint of a chapter to appear in *Designing an Intelligence*, MIT Press. <https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf> Accessed 2026-10-17.
- Simion, Mona and Christoph Kelp. 2023. "Trustworthy artificial intelligence." *Asian Journal of Philosophy* 2(1): 8. <https://doi.org/10.1007/s44204-023-00063-5>
- Simion, Mona. 2025. "Knowledge-First Epistemology: A Defence." *Elements in Epistemology*. <https://doi.org/10.1017/9781009454964>
- Simmons, Graham, Simone A. Glynn, Jerry A. Holmberg, John M. Coffin, Indira K. Hewlett, Shyh-Ching Lo, Judy A. Mikovits, William M. Switzer, Jeffrey M. Linnen, and Michael P. Busch. 2011a. "The Blood Xenotropic Murine Leukemia Virus-Related Virus Scientific Research Working Group: mission, progress, and plans." *Transfusion* 51(3): 643–653. Cited without page numbers from the author manuscript in PubMed Central, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3071162/>. <https://doi.org/10.1111/j.1537-2995.2011.03063.x>
- Simmons, Graham, Simone A. Glynn, Anthony L. Komaroff, Judy A. Mikovits, Leslie H. Tobler, John Hackett, Ning Tang, William M. Switzer, Walid Heneine, Indira K. Hewlett, Jiangqin Zhao, Shyh-Ching Lo, Harvey J. Alter, Jeffrey M. Linnen, Kui Gao, John M. Coffin, Mary F. Kearney, Francis W. Ruscetti, Max A. Pfost, James Bethel, Steven Kleinman, Jerry A. Holmberg, and Michael P. Busch. 2011b. "Failure to Confirm XMRV/MLVs in the Blood of Patients with Chronic Fatigue Syndrome: A Multi-Laboratory Study." *Science* 334(6057): 814–817. Cited without page numbers from the author manuscript in PubMed Central, <https://pmc.ncbi.nlm.nih.gov/articles/PMC3299483/>. <https://doi.org/10.1126/science.1213841>
- Simon, Herbert A. 1973. "The Structure of Ill Structured Problems." *Artificial Intelligence* 4(3–4): 181–201. Read from the scan of the published article in Carnegie Mellon University Libraries' Herbert A. Simon Collection; page numbers are those printed. https://iiif.library.cmu.edu/file/Simon_box00085_fld06830_bdl0008_doc0001/Simon_box00085_fld06830_bdl0008_doc0001.pdf
- Sinhahajari, Soumitra, Navonil Majumder, and Soujanya Poria. 2026. "On the Limits of LLM-as-Judge for Scientific Novelty Assessment." *arXiv*. Preprint, version 1; page numbers cited in the text refer to this version. <https://arxiv.org/abs/2606.12071v1>
- SlaveVoyages Consortium. 2024. "History of the Project." SlaveVoyages. <https://legacy.slavevoyages.org/blog/history-project> Accessed 2026-09-07.
- Smaldino, Paul E. and Richard McElreath. 2016. "The natural selection of bad science." *Royal Society Open Science* 3(9): 160384. <https://doi.org/10.1098/rsos.160384>
- Smaldino, Paul E. and Richard McElreath. 2023. "Correction to: 'The natural selection of bad science' (2016) by Paul E. Smaldino and Richard McElreath." *Royal Society Open Science* 10(9): rsos231026. <https://doi.org/10.1098/rsos.231026>
- Smart, Paul, Robert Clowes, and Andy Clark. 2025. "ChatGPT, extended: large language models and the extended mind." *Synthese* 205(6): 242. <https://doi.org/10.1007/s11229-025-05046-y>
- Smith, George E. 2014. "Kepler's Astronomia Nova." Lecture notes, Philosophy 167, Class 3. Tufts University. Page numbers cited in the text refer to the printed page numbers of the notes. <https://dl.tufts.edu/downloads/s4655t006>
- Smith, Wendy S., Jennifer R. Hale, and Cameron Neylon. 2011. "Applying neutral drift to the directed molecular evolution of a β -glucuronidase into a β -galactosidase: Two different evolutionary pathways lead to the same variant." *BMC Research Notes* 4(1): 138. <https://doi.org/10.1186/1756-0500-4-138>
- Smithers, Tim. 1997. "Autonomy in Robots and Other Agents." *Brain and Cognition* 34(1): 88–106. Cited from the author's abstract, open on PubMed, <https://pubmed.ncbi.nlm.nih.gov/9209757/>. <https://doi.org/10.1006/brcg.1997.0908>
- Snellen, I. A. G., L. Guzman-Ramirez, M. R. Hogerheijde, A. P. S. Hygate, and F. F. S. van der Tak. 2020. "Re-analysis of the 267-GHz ALMA observations of Venus: No statistically significant detection of phosphine." *arXiv*. Preprint. <https://arxiv.org/abs/2010.09761>
- Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. 2015. "Corrigibility." In *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*. AAAI Press. <https://cdn.aaai.org/ocs/ws/ws0067/10124-45900-1-PB.pdf>
- Soares, Nate. 2016. "The Value Learning Problem." In *Ethics for Artificial Intelligence Workshop at the 25th International Joint Conference on Artificial Intelligence (IJCAI-2016)*. Machine Intelligence Research Institute. <https://intelligence.org/files/ValueLearningProblem.pdf> Accessed 2026-10-17.
- Sober, Elliott. 1988. "Apportioning Causal Responsibility." *The Journal of Philosophy* 85(6): 303–318. Cited from the JSTOR copy of the article that Elliott Sober links from his "Selected Papers" page, <https://sober.philosophy.wisc.edu/selected-papers>, <https://drive.google.com/file/d/1A6MIeLsLJN-aCEqTKRjIHZEEdIPtyWvP/view>, which carries the journal pagination. <https://doi.org/10.2307/2026721>
- Sober, Elliott. 1999. "Testability." *Proceedings and Addresses of the American Philosophical Association* 73(2): 47–76. Cited from the JSTOR copy of the article that Elliott Sober links from his "Selected Papers" page, <https://sober.philosophy.wisc.edu/selected-papers>,

<https://drive.google.com/file/d/1dMmsDcMie97uo5FpDt-JXB83YmjHv7jN/view>, which carries the journal pagination. <https://doi.org/10.2307/3131087>

Sober, Elliott. 2004. "Likelihood, Model Selection, and the Duhem-Quine Problem." *Journal of Philosophy* 101(5): 221–241. Cited from the scan of the typeset *Journal of Philosophy* pages that Sober posted on his University of Wisconsin page, as captured by the Wayback Machine on 10 June 2010, <https://web.archive.org/web/20100610094232/http://philosophy.wisc.edu/sober/lik%20model%20sel%20and%20d%20q%20symbols%20ok.pdf>, which carries the journal pagination. <https://doi.org/10.5840/jphil2004101515>

Soleymanibrojeni, Mohammad, Roland Aydin, Diego Guedes-Sobrinho, Alexandre C. Dias, Maurício J. Piotrowski, Wolfgang Wenzel, and Celso Ricardo Caldeira Rêgo. 2026. "GENIUS: an agentic AI framework for autonomous design and execution of simulation protocols." *Communications Materials* 7(1): 115. <https://doi.org/10.1038/s43246-026-01167-0>

Solomon, Miriam. 1994. "Sizing Up Science: A Reply to Fuller." *Informal Logic* 16(1): 55–58. <https://doi.org/10.22329/il.v16i1.2438>

Solomon, Miriam. 2001b. "It Isn't the Thought That Counts." *Argumentation* 15(1): 67–75. Cited from the copy of the published article in PhilArchive, which carries the journal pagination. <https://philpapers.org/archive/SOLIIT.pdf>

Solomon, Miriam. 2006a. "Groupthink versus The Wisdom of Crowds: The Social Epistemology of Deliberation and Dissent." *The Southern Journal of Philosophy* 44(S1): 28–42. The citation by "abstract" is to the article's abstract. <https://doi.org/10.1111/j.2041-6962.2006.tb00028.x>

Solomon, Miriam. 2006b. "Norms of Epistemic Diversity." *Episteme* 3(1-2): 23–36. The citation by "abstract" is to the article's abstract. <https://doi.org/10.3366/epi.2006.3.1-2.23>

Son, Guijin, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, and Stella Biderman. 2025. "When AI Co-Scientists Fail: SPOT—a Benchmark for Automated Verification of Scientific Research." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2505.11855v1>

Song, Zhangde, Jieyu Lu, Yuanqi Du, Botao Yu, Thomas M. Pruyn, Yue Huang, Kehan Guo, Xiuzhe Luo, Yuanhao Qu, Yi Qu, Yinkai Wang, Haorui Wang, Jeff Guo, Jingru Gan, Parshin Shojaee, Di Luo, Steven Dillmann, Andres M Bran, Gen Li, Qiyuan Zhao, Shao-Xiong Lennon Luo, Yuxuan Zhang, Xiang Zou, Wanru Zhao, Yifan F. Zhang, Wucheng Zhang, Shunan Zheng, Saiyang Zhang, Sartaa Takrim Khan, Mahyar Rajabi-Kochi, Samantha Paradi-Maropakis, Tony Baltoiu, Fengyu Xie, Tianyang Chen, Kexin Huang, Weiliang Luo, Meijing Fang, Xin Yang, Lixue Cheng, Jiajun He, Soha Hassoun, Xiangliang Zhang, Wei Wang, Chandan K. Reddy, Chao Zhang, Zhiling Zheng, Mengdi Wang, Le Cong, Carla P. Gomes, Chang-Yu Hsieh, Aditya Nandy, Philippe Schwaller, Heather J. Kulik, Haojun Jia, Huan Sun, Seyed Mohamad Moosavi, and Chenru Duan. 2026. "Evaluating Large Language Models in Scientific Discovery." *arXiv*. Preprint, version 2. <https://arxiv.org/abs/2512.15567v2>

Sosa, Ernest. 2005. "A Virtue Epistemology." Faculty of Philosophy, University of Oxford. Handout for Lecture 2 of the John Locke Lectures, "Apt Belief and Reflective Knowledge." Section numbers cited in the text refer to the handout's numbered paragraphs; the Faculty's John Locke Lectures page, <https://www.philosophy.ox.ac.uk/john-locke-lectures>, lists the lecture under Trinity Term 2005. https://media.philosophy.ox.ac.uk/assets/pdf_file/0018/909/A_VIRTUE_EPISTEMOLOGY.pdf Accessed 2026-10-30.

Sosa, Ernest. 2015. "Knowing full well: the normativity of beliefs as performances." *Disputatio. Philosophical Research Bulletin* 4(5): 81–94. <https://studiahumanitatis.eu/ojs/index.php/disputatio/article/download/sosa-knowing/320> Accessed 2026-10-29.

Sourati, Jamshid and James A. Evans. 2023. "Accelerating Science with Human-Aware Artificial Intelligence." *Nature Human Behaviour* 7(10): 1682–1696. Page numbers cited in the text refer to the preprint, arXiv:2306.01495. <https://doi.org/10.1038/s41562-023-01648-z>

Srikanth, Dhruv, Bingchen Zhao, Dixing Xu, Yuxiang Wu, and Zhengyao Jiang. 2026. "Recursive self-improvement of AI research agents." *arXiv*. Preprint. <https://arxiv.org/abs/2609.26457>

Stacey, Martin, Claudia Eckert, Zachary Gallagher Pirtle, Michael Poznic, Beth-Anne Schuelke-Leech, and Loretta von der Tann. 2025. "Methods as a form of engineering knowledge." *Design Science* 11: e12. <https://doi.org/10.1017/dsj.2025.9>

Stanford, Kyle. 2023. "Underdetermination of Scientific Theory." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/scientific-underdetermination/> Accessed 2026-09-29.

Star, Susan Leigh and James R. Griesemer. 1989. "Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39." *Social Studies of Science* 19(3): 387–420. Page numbers cited in the text refer to the German translation by Michael Schmidt, "Institutionelle Ökologie, 'Übersetzungen' und Grenzobjekte: Amateure und Professionelle im Museum of Vertebrate Zoology in Berkeley, 1907–39," in Susan Leigh Star, *Grenzobjekte und Medienforschung*, edited by Sebastian Gießmann and Nadine Taha (transcript, 2017), 81–116, <https://www.transcript-verlag.de/media/pdf/86/c2/16/oa9783839431269qvwZbon4EVU0e.pdf>; quotations are this book's own translations from the German. <https://doi.org/10.1177/030631289019003001>

Star, Susan Leigh and Karen Ruhleder. 1994. "Steps towards an ecology of infrastructure." In *Proceedings of the 1994 ACM conference on Computer supported cooperative work - CSCW '94*, 253–264. ACM Press. <https://doi.org/10.1145/192844.193021>

- Star, Susan Leigh and Karen Ruhleder. 1996. "Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces." *Information Systems Research* 7(1): 111–134. Page numbers cited in the text refer to the German translation by Michael Schmidt, "Schritte zu einer Ökologie von Infrastruktur: Design und Zugang für großangelegte Informationsräume," in Susan Leigh Star, *Grenzobjekte und Medienforschung*, edited by Sebastian Gießmann and Nadine Taha (transcript, 2017), 359–402, <https://www.transcript-verlag.de/media/pdf/86/c2/16/oa9783839431269qvwZbon4EVU0e.pdf>. <https://doi.org/10.1287/isre.7.1.111>
- Steele, Katie and H. Orri Stefánsson. 2025. "Decision Theory." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/decision-theory/> Accessed 2026-10-18.
- Steinle, Friedrich. 2004. "Explorative Experimentieren: Charles Dufay und die Entdeckung der zwei Elektrizitäten." *Physik Journal* 3(6): 47–52. Cited from the publisher's PDF on pro-physik.de, which carries the journal pagination. Paraphrases in the text are this book's renderings of the German. <https://pro-physik.de/zeitschriften/download/15383> Accessed 2026-10-29.
- Stilgoe, Jack, Richard Owen, and Phil Macnaghten. 2013. "Developing a framework for responsible innovation." *Research Policy* 42(9): 1568–1580. <https://doi.org/10.1016/j.respol.2013.05.008>
- Strevens, Michael. 2001. "The Bayesian Treatment of Auxiliary Hypotheses." *The British Journal for the Philosophy of Science* 52(3): 515–537. Section numbers cited in the text refer to the author's version, <https://www.strevens.org/research/prob/auxhyp.pdf>. <https://doi.org/10.1093/bjps/52.3.515>
- Strevens, Michael. 2003. "The Role of the Priority Rule in Science." *Journal of Philosophy* 100(2): 55–79. Section numbers cited in the text refer to the author's version on the author's website, <https://www.strevens.org/research/scistruc/Prioritas.pdf>; they are the same as in the published article. <https://doi.org/10.5840/jphil2003100224>
- Strevens, Michael. 2010. "Reconsidering Authority: Scientific Expertise, Bounded Rationality, and Epistemic Backtracking." In *Oxford Studies in Epistemology*, 294–330. Oxford University Press. Page numbers cited in the text refer to the author's manuscript on his website. <https://www.strevens.org/research/scistruc/Impressio.pdf>
- Strevens, Michael. 2011. "Economic Approaches to Understanding Scientific Norms." *Episteme* 8: 184–200. Page numbers cited in the text refer to the author's manuscript on his website. <https://www.strevens.org/research/scistruc/Oeconomicus.pdf>
- Strevens, Michael. 2013. "No understanding without explanation." *Studies in History and Philosophy of Science Part A* 44(3): 510–515. Cited without page numbers from the author's version on the author's website, <https://www.strevens.org/research/lux/Illuminatio.pdf>. <https://doi.org/10.1016/j.shpsa.2012.12.005>
- Stuart, Michael T. 2025. "A New Account of Pragmatic Understanding, Applied to the Case of AI-Assisted Science." *Philosophical Studies*. Page numbers cited in the text refer to the accepted manuscript, White Rose Research Online, <https://eprints.whiterose.ac.uk/id/eprint/229596/>. <https://doi.org/10.1007/s11098-025-02336-6>
- Suchman, Lucy. n.d. "Agencies in Technology Design: Feminist Reconfigurations." Online paper, Department of Sociology, Lancaster University. The paper has no printed pagination. <https://www.lancaster.ac.uk/fass/resources/sociology-online-papers/papers/suchman-agenciessnodesign.pdf> Accessed 2026-10-13.
- Sullivan, Emily. 2022a. "Inductive Risk, Understanding, and Opaque Machine Learning Models." *Philosophy of Science* 89(5): 1065–1074. <https://doi.org/10.1017/psa.2022.62>
- Sullivan, Emily. 2022b. "Understanding from Machine Learning Models." *The British Journal for the Philosophy of Science* 73(1): 109–133. Section numbers cited in the text refer to the author's penultimate version, PhilSci-Archive 16276, <https://philsci-archive.pitt.edu/16276/>. <https://doi.org/10.1093/bjps/axz035>
- Swanson, Kyle, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. 2024. "The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation." openRxiv. <https://doi.org/10.1101/2024.11.11.623004>
- Szymanski, Nathan J., Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. 2023. "An autonomous laboratory for the accelerated synthesis of inorganic materials." *Nature* 624(7990): 86–91. Consulted the current online version, as amended by the 2026 Author Correction (title changed from "...synthesis of novel materials"). <https://doi.org/10.1038/s41586-023-06734-w>
- Szymanski, Nathan J., Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. 2026. "Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials." *Nature* 650(8100): E1–E1. <https://doi.org/10.1038/s41586-025-09992-y>
- Sørensen, Henrik Kragh. 2010. "Exploratory Experimentation in Experimental Mathematics: A Glimpse at the PSLQ Algorithm." In *PhiMSAMP: Philosophy of Mathematics: Sociological Aspects and Mathematical Practice*, edited by Benedikt Löwe and Thomas Müller, 341–360. College Publications. http://www.lib.uni-bonn.de/PhiMSAMP/Data/Book/PhiMSAMP-bk_Soerensen.pdf Accessed 2026-10-04.

- Takagi, Shiro. 2023. "Speculative Exploration on the Concept of Artificial Agents Conducting Autonomous Research." *arXiv*. Preprint, version 1. <https://arxiv.org/abs/2312.03497v1>
- Takagi, Shiro and an AI agent. 2026. "Another Vision of Metascience for the Age of AI: Exploring the Design Space of Science." https://t46.github.io/blogs/another_vision_of_metascience_for_the_age_of_ai.html Accessed 2026-10-07.
- Takahashi, Koichi. 2019. "Kagaku to gijutsu no rikon [The Divorce of Science and Technology]." *RAD-IT21*. Web column (in Japanese), 24 September 2019; cited without page numbers (section "AI no kagaku" no tanjō" [The Birth of 'the Science of AI']). <https://rad-it21.com/ai/ktakahashi190924/> Accessed 2026-10-29.
- Takahashi, Koichi. 2024. "Kagaku AI no jiritsusei reberu [Levels of Autonomy in Science Automation]." *Jxiv*. Preprint 818, version 2. <https://doi.org/10.51094/jxiv.818>
- Takahashi, Koichi. 2026. *AGI: Ningen o koeru chinō wa bunmei o ika ni hen'yō saseru ka* [AGI: How Superhuman Intelligence Will Transform Civilization]. Kōdansha Sensho Mechie 847. Kōdansha. ISBN 9784065447178. Page numbers cited in the text were read from photographs of the printed book. Appendices 1 and 3, cited by number and heading, are posted on the author's website: <https://koichi-takahashi.me/en/agibook/appendix-1/> (Appendix 1, English version, updated 30 September 2026), <https://koichi-takahashi.me/en/agibook/appendix-3/> (Appendix 3, English version, updated 30 September 2026) and <https://koichi-takahashi.me/agibook/appendix-3/> (Appendix 3, in Japanese, updated 8 October 2026). <https://koichi-takahashi.me/agibook/> Accessed 2026-10-29.
- Tal, Eran. 2012. *The Epistemology of Measurement: A Model-Based Account*. PhD diss., University of Toronto. <https://hdl.handle.net/1807/34936>
- Tanchuk, Nicolas J. 2026. "AI Personalized Learning and the Risk of Epistemic Consolidation." *Studies in Philosophy and Education*. Online-first version of record (10 June 2026), no volume, issue or page range yet; page numbers cited in the text count PDF pages from the first page. <https://link.springer.com/article/10.1007/s11217-026-10081-4>
- Tanenbaum, M. and D. E. Thomas. 1956. "Diffused Emitter and Base Silicon Transistors." *Bell System Technical Journal* 35(1): 1–22. https://bitsavers.org/magazines/Bell_System_Technical_Journal/BSTJ_V35N01_195601.pdf Accessed 2026-09-29.
- Tanenbaum, Morris. 2004. "Oral history interview with Morris Tanenbaum." *Science History Institute Digital Collections*. Chemical Heritage Foundation (Science History Institute). <https://digital.sciencehistory.org/works/9nqp9m6> Accessed 2026-09-29.
- Tang, Qiong, Tianxiang Sun, Xiangkun Hu, Xiangyang Liu, Yiran Chen, Yunfan Shao, Bobo Li, Changze Lv, Cheng Xu, Chengsong Huang, Chunyang Li, Dizhan Xue, Hao Bai, Haodong Duan, Hengquan Guo, Hongyang He, Hongyi Chen, Hui Shen, Jiahao Yuan, Jiankai Sun, Jikang Cheng, Jinfeng Xu, Jingqi Tong, Jingye Chen, Jinxiu Liu, Jixuan Leng, Junchi Yu, Kaixun Jiang, Kun Xiang, Kunpeng Yao, Lang Feng, Liangqi Yuan, Longsen Gao, Meng Li, Qi Jia, Qishi Sun, Shengyuan Ding, Shizhan Gong, Siru Zhong, Terry Jingchen Zhang, Tianle Gu, Tianyi Liang, Weijie Liu, Weikai Yang, Weizhi Fei, Xin Wang, Xinpeng Liu, Xuanwen Ding, Yihong Tang, Yuanli Wang, Yukun Jiang, Yuming Yang, Zhengbao He, Zhikai Chen, Zhikun Xu, Zhuang Li, and Zihao Huang. 2026. "FARS: A Fully Automated Research System Deployed at Scale." *arXiv*. Preprint. <https://arxiv.org/abs/2606.31651>
- Taniguchi, Tadahiro, Shiro Takagi, Jun Otsuka, Yusuke Hayashi, and Hiro Taiyo Hamada. 2025. "Collective predictive coding as model of science: formalizing scientific activities towards generative science." *Royal Society Open Science* 12(6): 241678. <https://doi.org/10.1098/rsos.241678>
- Taniguchi, Tadahiro, Yusuke Hayashi, Momoha Hirose, Mizuki Oka, Ken Suzuki, Olaf Witkowski, and Audrey Tang. 2026. "Symbiotic Alignment via Collective Predictive Coding: A Theoretical Framework for Co-Creative Human–AI Ecosystems." *Artificial Life*. Published online ahead of issue; page numbers cited in the text refer to the early-access version (pp. 1–29). <https://doi.org/10.1162/ARTL.a.475>
- Taniguchi, Tadahiro (ed.). 2026a. *Symbol Emergence Systems*. Springer. Cited for the editor's Preface, v–vii. <https://doi.org/10.1007/978-981-95-1327-7>
- Taniguchi, Tadahiro. 2026b. "Symbol Emergence Systems: Toward a World Where Humans and AI Co-Discover Meaning." In *Symbol Emergence Systems*, edited by Tadahiro Taniguchi, 3–9. Springer. https://doi.org/10.1007/978-981-95-1327-7_1
- Taniguchi, Tadahiro. 2026c. "Symbol Emergence in Robotics: A Constructive Approach to Overcoming the Symbol Grounding Problem." In *Symbol Emergence Systems*, edited by Tadahiro Taniguchi, 39–44. Springer. https://doi.org/10.1007/978-981-95-1327-7_7
- Taniguchi, Tadahiro. 2026d. "Language in the Era of Generative AI: Insights from Large Language Models." In *Symbol Emergence Systems*, edited by Tadahiro Taniguchi, 197–201. Springer. https://doi.org/10.1007/978-981-95-1327-7_31
- Taniguchi, Tadahiro. 2026e. "Constructive Approaches to Symbol Emergence Systems: Probabilistic Generative Models and Language Games." In *Symbol Emergence Systems*, edited by Tadahiro Taniguchi, 203–208. Springer. https://doi.org/10.1007/978-981-95-1327-7_32
- Taniguchi, Tadahiro. 2026f. "The Collective Predictive Coding Hypothesis: If Language Exists to Help Us Predict the World." In *Symbol Emergence Systems*, edited by Tadahiro Taniguchi, 209–214. Springer. https://doi.org/10.1007/978-981-95-1327-7_33

- Taniguchi, Tadahiro. 2026g. "Toward a Symbiotic Society with Generative AI: For Living Together into the Future." In *Symbol Emergence Systems*, edited by Tadahiro Taniguchi, 221–225. Springer. https://doi.org/10.1007/978-981-95-1327-7_35
- Tao, Terence. 2026. "A Digestion of the Jacobian Conjecture Counterexample." *What's New*. <https://terrytao.wordpress.com/2026/07/21/a-digestion-of-the-jacobian-conjecture-counterexample> Accessed 2026-10-28.
- Termine, Alberto, Emanuele Ratti, and Alessandro Facchini. 2026. "Machine learning and theory-ladenness: a phenomenological account." *Synthese* 207(3): 94. <https://doi.org/10.1007/s11229-026-05454-8>
- Thais, Savannah, Roberto Trotta, Nathan Suri, Emily Sullivan, Viyan Poonamallee, Tanaporn Na Narong, Rupert Croft, and Nicole Hartman. 2026. "AI for Science Needs Scientific Alignment." PhilSci-Archive. Position-paper preprint dated 27 March 2026; page numbers cited in the text refer to this preprint. <https://philsci-archive.pitt.edu/28994> Accessed 2026-09-07.
- Thakkar, Nitya, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. 2025. "Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025." *arXiv*. Preprint, version 1; a revised version was published as "A large-scale randomized study of large language model feedback in peer review," *Nature Machine Intelligence* 8 (2026): 326–336, <https://doi.org/10.1038/s42256-026-01188-x>. Page numbers cited in the text refer to the preprint. <https://arxiv.org/abs/2504.09737v1>
- Thoma, Johanna. 2015. "The Epistemic Division of Labor Revisited." *Philosophy of Science* 82(3): 454–472. <https://doi.org/10.1086/681768>
- Thomas A. Edison Papers. 1900. "N-00-12-20 (ca. 1900), Notebook NA168-N, Battery Test Books." *Thomas Edison Papers Digital Edition*. Rutgers University. Cited by IIF canvas number and by the editors' note on Group 1, Battery Test Books. Courtesy of Thomas Edison National Historical Park. <https://edisondigital.rutgers.edu/notebooks/document/NA168-N> Accessed 2026-10-08.
- Thornley, Elliott. 2025. "The shutdown problem: an AI engineering puzzle for decision theorists." *Philosophical Studies* 182(7): 1653–1680. <https://doi.org/10.1007/s11098-024-02153-3>
- Thornton, Stephen. 2022. "Karl Popper." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/popper/> Accessed 2026-10-29.
- Thorstad, David. 2022. "There are no epistemic norms of inquiry." *Synthese* 200(5): 410. <https://doi.org/10.1007/s11229-022-03896-4>
- Thurston, William P. 1994. "On proof and progress in mathematics." *Bulletin of the American Mathematical Society* 30(2): 161–178. <https://doi.org/10.1090/s0273-0979-1994-00502-6>
- Tian, Chao and Krishna Narayanan. 2026. "From Formulation to Articulation: Considerations for AI-Assisted Research in Information Theory." *arXiv*. Preprint. <https://arxiv.org/abs/2610.25490>
- Tollefsen, Christopher. 2018. "What Is 'Good Science'?" In *Will Science Remain Human? Experts meeting of the Social Trends Institute, Rome, 5–6 March 2018: paper abstracts*. Social Trends Institute. <https://socialtrends-institute.org/meetings/will-science-remain-human/> Accessed 2026-11-01.
- Tong, Jingqi, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, Ming Zhang, Qiguang Chen, Weifeng Ge, Qipeng Guo, Tianlei Ying, Tianxiang Sun, Yining Zheng, Zhiheng Xi, Xinchu Chen, Jun Zhao, Ning Ding, Xuanjing Huang, Yu-Gang Jiang, and Xipeng Qiu. 2026. "AI Can Learn Scientific Taste." *arXiv*. Preprint, version 3; page numbers cited in the text refer to this version. <https://arxiv.org/abs/2603.14473v3>
- Totschnig, Wolfhart. 2019. "Fully autonomous AI." In *Proceedings of the South African Forum for Artificial Intelligence Research (FAIR 2019), Cape Town, South Africa, 4–6 December 2019*, edited by Marelise Davel and Etienne Barnard. CEUR-WS.org. Extended abstract of the paper published as "Fully Autonomous AI," *Science and Engineering Ethics* 26 (2020): 2473–2485. Page numbers cited in the text refer to the extended abstract. https://ceur-ws.org/Vol-2540/FAIR2019_paper_2.pdf Accessed 2026-10-29.
- Totschnig, Wolfhart. 2026. "Why an artificial agent may change the given final goal." *Humanities and Social Sciences Communications*. The accepted manuscript was consulted. <https://doi.org/10.1057/s41599-026-08312-7>
- Tse, Edwin. 2023. "LabArchives ELN Exports (EGT, DMK, DGS) for the OSA Series 2 Paper Titled 'Open Source Antibiotics - Simple Diarylimidazoles Are Potent Against Methicillin Resistant Staphylococcus Aureus'." University College London Research Data Repository. LabArchives notebook exports deposited as a dataset; cited by notebook page title. <https://doi.org/10.5522/04/21749750.v1>
- Tsoukalas, George, Anton Kovsharov, Sergey Shirobokov, Anja Surina, Moritz Firsching, Gergely Bérczi, Francisco J. R. Ruiz, Arun Suggala, Adam Zsolt Wagner, Eric Wieser, Lei Yu, Aja Huang, Miklós Z. Horváth, Andrew Ferraiuolo, Henryk Michalewski, Edward Lockhart, Codrut Grosu, Thomas Hubert, Matej Balog, Pushmeet Kohli, and Swarat Chaudhuri. 2026. "Advancing Mathematics Research with AI-Driven Formal Proof Search." *arXiv*. Preprint. <https://arxiv.org/abs/2605.22763>
- Tu, Youyou. 2015. "Artemisinin—A Gift from Traditional Chinese Medicine to the World." In *The Nobel Prizes 2015*. Nobel Foundation. <https://www.nobelprize.org/prizes/medicine/2015/tu/lecture> Accessed 2026-09-29.
- Tubert, Ariela and Justin Tiehen. 2025a. "Existentialist risk and value misalignment." *Philosophical Studies* 182(7): 1609–1626. Section numbers cited in the text refer to the authors' draft, posted on Justin Tiehen's website,

https://www.justintiehen.com/uploads/1/2/0/1/120134130/existentialist_risk_and_value_misalignment_r_r_feb3.docx; the draft has no page numbers. <https://doi.org/10.1007/s11098-024-02142-6>

Tubert, Ariela and Justin Tiehen. 2025b. "Value alignment, human enhancement, and moral revolutions." *Inquiry* 68(4): 1248–1270. Section numbers cited in the text refer to the authors' draft, posted on Justin Tiehen's website, https://www.justintiehen.com/uploads/1/2/0/1/120134130/value_alignment_human_enhancement_and_moral_revolutions_r-and-r.docx; the draft has no page numbers. <https://doi.org/10.1080/0020174X.2023.2261506>

Turri, John, Mark Alfano, and John Greco. 2021. "Virtue Epistemology." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/epistemology-virtue/> Accessed 2026-09-29.

UsefulChem. 2010. "UsefulChem Archive, Snapshot of 27 April 2010." Drexel University. Internet Archive capture (27 October 2011) of the snapshot of 27 April 2010; notebook pages cited by experiment number. <https://web.archive.org/web/20111027182546/http://showme.physics.drexel.edu/usefulchem/archives/usefulchem2010-04-27.zip> Accessed 2026-10-04.

UsefulChem. 2016. "Exp037: Page History." *usefulchem.wikispaces.com*. Internet Archive capture. <https://web.archive.org/web/20160202230741/http://usefulchem.wikispaces.com/page/history/Exp037> Accessed 2026-09-30.

Uygun Tunç, Duygu. 2021. "Collective scientific knowledge without a collective subject." Preprint, PhilSci-Archive 19187 (deposited 16 June 2021); page numbers cited in the text are those of the preprint. <https://philsci-archive.pitt.edu/19187> Accessed 2026-10-29.

Uygun Tunç, Duygu. 2023. "The subject of knowledge in collaborative science." *Synthese* 201(3): 88. <https://doi.org/10.1007/s11229-023-04080-y>

Vacek, Daniela. 2025. "Against artificial achievements." *Analysis* 85(3): 690–698. <https://doi.org/10.1093/analys/anaf034>

Vallier, Kevin. 2022. "Public Justification." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/justification-public/> Accessed 2026-10-08.

van de Poel, Ibo, Jessica Nihlén Fahlquist, Neelke Doorn, Sjoerd Zwart, and Lambèr Royakkers. 2012. "The Problem of Many Hands: Climate Change as an Example." *Science and Engineering Ethics* 18(1): 49–67. <https://doi.org/10.1007/s11948-011-9276-0>

Van de Sompel, Herbert, Sandy Payette, John Erickson, Carl Lagoze, and Simeon Warner. 2004. "Rethinking Scholarly Communication: Building the System That Scholars Deserve." *D-Lib Magazine* 10(9). <https://doi.org/10.1045/september2004-vandesompel>

van der Merwe, Ragnar. 2023. "How pluralistic is pluralism really? A case study of Sandra Mitchell's Integrative Pluralism." *Theoria: An International Journal for Theory, History and Foundations of Science* 38(3): 319–338. Cited from the published journal PDF deposited in PhilSci-Archive. https://philsci-archive.pitt.edu/23747/1/def_23838_van-der-Merwe_Theoria38-3.pdf

van Fraassen, Bas C. 2005. "The Day of the Dolphins: Puzzling over Epistemic Partnership." In *Mistakes of Reason: Essays in Honour of John Woods*, edited by Andrew D. Irvine and Kent A. Peacock, 111–133. University of Toronto Press. Page numbers cited in the text refer to the author's version, formerly on the author's Princeton website and now available as an Internet Archive capture (11 October 2022), <https://web.archive.org/web/20221011071250/https://www.princeton.edu/~fraassen/abstract/Dolphins.pdf>; the published pagination differs. <https://doi.org/10.3138/9781442677296-010>

van Lier, Maud. 2023. "Introducing a four-fold way to conceptualize artificial agency." *Synthese* 201(3): 85. <https://doi.org/10.1007/s11229-023-04083-9>

Varsavsky, Oscar. 1969. *Ciencia, política y cientificismo*. Centro Editor de América Latina. Cited by section from the ESOCITE repository transcription, which lacks the original pagination. <https://repositorio.esocite.la/339> Accessed 2026-10-08.

Varsavsky, Oscar. 1975. "Bases para una política nacional de tecnología y ciencia." In *Ciencia e ideología: aportes polémicos*, 52–57. Ediciones Ciencia Nueva. <https://repositorio.esocite.la/388> Accessed 2026-10-08.

Varsavsky, Oscar. [1971] 1975. "Ideología y verdad." In *Ciencia e ideología: aportes polémicos*, 41–51. Ediciones Ciencia Nueva. First published in *Ciencia Nueva* (1971). <https://repositorio.esocite.la/388> Accessed 2026-10-08.

Venkatesh, Akshay. 2025. "Human mathematics in the age of reasoning machines." *Annals of Mathematics and Philosophy* 3(1): 83–116. <https://doi.org/10.67678/ba8ecaqs>

Villanueva, Geronimo, Martin Cordiner, Patrick Irwin, Imke de Pater, Bryan Butler, Mark Gurwell, Stefanie Milam, Conor Nixon, Statia Luszczyk, Colin Wilson, Vincent Kofman, Giuliano Liuzzi, Sara Faggi, Thomas Fauchez, Manuela Lippi, Richard Cosentino, Alexander Thelen, Arielle Moullet, Paul Hartogh, Edward Molter, Steve Charnley, Giada Arney, Avi Mandell, Nicolas Biver, Ann Vandaele, Katherine de Kleer, and Ravi Kopparapu. 2020. "No evidence of phosphine in the atmosphere of Venus by independent analyses." *arXiv*. Preprint. <https://arxiv.org/abs/2010.14305>

Viola, Marco. 2015. "Some remarks on the division of cognitive labor." *RT: A Journal on Research Policy and Evaluation* 3(1): 1–14. <https://doi.org/10.13130/2282-5398/4768>

- Vlerick, Michael and Alex Broadbent. 2015. "Evolution and Epistemic Justification." *Dialectica* 69(2): 185–203. Cited from the copy of the published article on the first author's PhilArchive record, <https://philarchive.org/rec/VLEEA>, which carries the journal pagination. <https://doi.org/10.1111/1746-8361.12098>
- Volokh, Eugene. 2019. "Chief Justice Robots." *Duke Law Journal* 68(6): 1135–1192. <https://www.law.ucla.edu/volokh/robots.pdf>
- Von Arx, Sydney, Lawrence Chan, and Beth Barnes. 2025. "Recent Frontier Models Are Reward Hacking." *METR (blog)*, 5 June 2025. <https://metr.org/blog/2025-06-05-recent-reward-hacking/> Accessed 2026-10-17.
- Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas, and Rogier A. Kievit. 2012. "An Agenda for Purely Confirmatory Research." *Perspectives on Psychological Science* 7(6): 632–638. Page numbers cited in the text refer to the authors' in-press manuscript on the first author's website, https://www.ejwagenmakers.com/2012/ConfirmatoryResearchFTW_inpress.pdf; the published pagination differs. <https://doi.org/10.1177/1745691612463078>
- Walch, Angela. 2021. "Testimony before the United States Senate Committee on Banking, Housing, and Urban Affairs." Hearing on "Cryptocurrencies: What Are They Good For?", 27 July 2021. Written testimony; archived copy: <https://web.archive.org/web/20260411070518/https://www.banking.senate.gov/imo/media/doc/Walch%20Testimony%207-27-21.pdf> <https://www.banking.senate.gov/imo/media/doc/Walch%20Testimony%207-27-21.pdf> Accessed 2026-10-18.
- Walkowitz, Judith R. 2009. "On Taking Notes." *Perspectives on History*. American Historical Association. Researcher self-account. <https://www.historians.org/perspectives-article/on-taking-notes-january-2009> Accessed 2026-09-07.
- Walsh, Clare R. and P. N. Johnson-Laird. 2009. "Changing your mind." *Memory & Cognition* 37(5): 624–631. <https://doi.org/10.3758/mc.37.5.624>
- Waluchow, Wil and Dimitrios Kyritsis. 2023. "Constitutionalism." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/constitutionalism/> Accessed 2026-10-08.
- Wang, Zhen, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Zhiting Hu, and Eric P. Xing. 2026. "FIRE-Bench: Evaluating AI Agents on the Rediscovery of Scientific Insights." *arXiv*. Preprint, version 2 (ICML 2026 camera-ready version). <https://arxiv.org/abs/2602.02905v2>
- Watson, Lani. 2015. *Why Ask? The Epistemology of Questioning*. PhD thesis, University of Edinburgh. <https://era.ed.ac.uk/handle/1842/19538>
- Weatherall, James Owen and Jesse Wolfson. 2026. "Correctness, Artificial Intelligence, and the Epistemic Value of Mathematical Proof." *arXiv*. Preprint, version 1 (draft of 16 February 2026). <https://arxiv.org/abs/2602.12463v1>
- Weber, Keith and Matthew Inglis. 2024. "Mathematics Education Research on Mathematical Practice." In *Handbook of the History and Philosophy of Mathematical Practice*, 2637–2663. Springer International Publishing. Section and note numbers cited in the text refer to the accepted manuscript, Loughborough University Research Repository, https://repository.lboro.ac.uk/articles/chapter/Mathematics_education_research_on_mathematical_practice/17013851; they are the same as in the published chapter. https://doi.org/10.1007/978-3-031-40846-5_88
- Weber, Marcel. 2018. "Experiment in Biology." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/biology-experiment/> Accessed 2026-10-04.
- Weber, Max. 1904. "Die »Objektivität« sozialwissenschaftlicher und sozialpolitischer Erkenntnis." In *Gesammelte Aufsätze zur Wissenschaftslehre*. Zeno.org. <http://www.zeno.org/nid/20011440120> Accessed 2026-10-07.
- Weisberg, Michael. 2007a. "Three Kinds of Idealization." *Journal of Philosophy* 104(12): 639–659. Section numbers cited in the text refer to the author's version, Penn ScholarlyCommons, https://repository.upenn.edu/philosophy_papers/14; they are the same as in the published article. <https://doi.org/10.5840/jphil20071041240>
- Weisberg, Michael. 2007b. "Who is a Modeler?" *British Journal for the Philosophy of Science* 58(2): 207–233. Page numbers cited in the text refer to the author's final draft ("Final Draft to appear in British Journal for Philosophy of Science"), posted on the author's University of Pennsylvania page and captured by the Wayback Machine, <https://web.archive.org/web/20240418223301/http://www.phil.upenn.edu/~weisberg/documents/modelingfinal.pdf>; the published pagination differs. <https://doi.org/10.1093/bjps/axm011>
- Weisberg, Michael and Kenneth Reisman. 2008. "The Robust Volterra Principle." *Philosophy of Science* 75(1): 106–131. Page numbers cited in the text refer to the authors' final manuscript, PhilSci-Archive 4085, <https://philsci-archive.pitt.edu/4085/>; the published pagination differs. <https://doi.org/10.1086/588395>
- Weisberg, Michael and Ryan Muldoon. 2009. "Epistemic Landscapes and the Division of Cognitive Labor." *Philosophy of Science* 76(2): 225–252. Cited from the publisher's PDF in Penn ScholarlyCommons, https://repository.upenn.edu/philosophy_papers/7, which carries the journal pagination. <https://doi.org/10.1086/644786>
- Weiss, Rainer. 1972. "Electromagnetically Coupled Broadband Gravitational Antenna." *Quarterly Progress Report, Research Laboratory of Electronics, Massachusetts Institute of Technology* (105): 54–76. <https://dcc.ligo.org/LIGO-P720002/public> Accessed 2026-09-29.

- Wendler, David. 2026. "The Ethics of Clinical Research." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/clinical-research/> Accessed 2026-10-08.
- Westerblad, Oscar. 2026. "Instrumental understanding." *Synthese* 207(1): 44. <https://doi.org/10.1007/s11229-025-05422-8>
- Weyl, E. Glen, Audrey Tang, and the Plurality Community. 2024. *Plurality: The Future of Collaborative Technology and Democracy*. Online edition, <https://github.com/pluralitybook/plurality> (commit 5cca933, 2026-09-30; passages cited identical at commit cba7053, 2024-05-31). Cited by chapter number; the online text has no stable pages. <https://www.plurality.net/> Accessed 2026-10-18.
- Wikispaces. 2018. "It's Time for Us to Say Farewell..." *Wikispaces Blog*. Internet Archive capture. <https://web.archive.org/web/20180301003634/http://blog.wikispaces.com> Accessed 2026-09-30.
- Wilholt, Torsten. 2010. "Scientific freedom: its grounds and their limitations." *Studies in History and Philosophy of Science Part A* 41(2): 174–181. Section numbers cited in the text refer to the author's manuscript version, PUB – Publications at Bielefeld University, <https://pub.uni-bielefeld.de/record/2551327>; they are the same as in the published article. <https://doi.org/10.1016/j.shpsa.2010.03.003>
- Wilholt, Torsten. 2013. "Epistemic Trust in Science." *The British Journal for the Philosophy of Science* 64(2): 233–253. Section and note numbers cited in the text refer to the author's manuscript version, PUB – Publications at Bielefeld University, <https://pub.uni-bielefeld.de/record/2555174>; they are the same as in the published article. <https://doi.org/10.1093/bjps/axs007>
- Wilholt, Torsten. 2020. "On Knowing What One Does Not Know: Ignorance and the Aims of Research." In *Science and the Production of Ignorance*, edited by Janet Kourany and Martin Carrier, 195–218. The MIT Press. Page numbers cited in the text refer to the author's draft of December 2018, PhilArchive, <https://philarchive.org/rec/WILOKW>; the published pagination differs. <https://doi.org/10.7551/mitpress/12146.003.0021>
- Wilkenfeld, Daniel A. 2013. "Understanding as representation manipulability." *Synthese* 190(6): 997–1016. Page numbers cited in the text refer to the author's manuscript on the author's website, <https://danielwilkenfeld.owlstown.net/publications/47574-understanding-as-representation-manipulability>; the published pagination differs. <https://doi.org/10.1007/s11229-011-0055-x>
- Wilkenfeld, Daniel A. 2019. "Understanding as Compression." *Philosophical Studies* 176(10): 2807–2831. The author's version was consulted. <https://doi.org/10.1007/s11098-018-1152-1>
- Williamson, Timothy. 2000. *Knowledge and its Limits*. Oxford University Press. <https://doi.org/10.1093/019925656x.001.0001>
- Wilsdon, James, Liz Allen, Eleonora Belfiore, Philip Campbell, Stephen Curry, Steven Hill, Richard Jones, Roger Kain, Simon Kerridge, Mike Thelwall, Jane Tinkler, Ian Viney, Paul Wouters, Jude Hill, and Ben Johnson. 2015. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Higher Education Funding Council for England. <https://www.ukri.org/wp-content/uploads/2021/12/RE-151221-TheMetricTideFullReport2015.pdf>
- Wilson, Shawn. 2003. "Progressing Toward an Indigenous Research Paradigm in Canada and Australia." *Canadian Journal of Native Education* 27(2): 161–178. <https://doi.org/10.14288/cjne.v27i2.196353>
- Wilson, Shawn. 2004. "Research as Ceremony: Articulating an Indigenous Research Paradigm." Monash University. PhD thesis (submitted February 2004; accepted 17 August 2004), open in Monash University's repository, Bridges; page numbers are those printed in the thesis. The author's book *Research Is Ceremony: Indigenous Research Methods* (Fernwood, 2008) develops the thesis. <https://doi.org/10.4225/03/59d448220e4c3>
- Wilson, Shawn. 2007. "Guest Editorial: What is an Indigenist Research Paradigm?" *Canadian Journal of Native Education* 30(2): 193–195. <https://doi.org/10.14288/cjne.v30i2.196422>
- Winn, Zach. 2025. "Accelerating scientific discovery with AI." *MIT News*, 30 June 2025. <https://news.mit.edu/2025/futurehouse-accelerates-scientific-discovery-with-ai-0630> Accessed 2026-10-17.
- Winsberg, Eric and Ramón Alvarado. 2026. "Computer Simulations in Science." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/simulations-science/> Accessed 2026-10-05.
- Wittgenstein, Ludwig. 1969. *On Certainty / Über Gewißheit*. Translated by Denis Paul and G. E. M. Anscombe. Basil Blackwell. Cited by section; the German text was consulted at The Ludwig Wittgenstein Project. https://www.wittgensteinproject.org/w/index.php/%C3%9Cber_Gewi%C3%9Fheit
- Wiśniewski, Andrzej. 2013. *Questions, Inferences, and Scenarios*. Studies in Logic 46. College Publications. PDF posted by the author. <https://awisniew.web.amu.edu.pl/publikacje/QISaw.pdf> Accessed 2026-10-24.
- Wiśniewski, Andrzej. 2022. "An Essay on Inferential Erotetic Logic." In *Asking and Answering: Rivalling Approaches to Interrogative Methods*, edited by Moritz Cordes, 105–138. Narr Francke Attempto. <https://doi.org/10.24053/9783823394808>
- Wolchover, Natalie. 2020. "Artificial Intelligence Will Do What We Ask. That's a Problem." *Quanta Magazine*. Published 30 January 2020. <https://www.quantamagazine.org/artificial-intelligence-will-do-what-we-ask-thats-a-problem-20200130/>

- Wolf, Thomas. n.d. "The Einstein AI model." *thomwolf.io*. <https://thomwolf.io/blog/scientific-ai.html> Accessed 2026-10-27.
- Wong, Sophia Isako. 2008. "Cognitive Disabilities and the Scope of Contractualist Justice." *Auslegung* 29(2): 1–24. <https://doi.org/10.17161/ajp.1808.9678>
- Woodward, James. 2000. "Explanation and Invariance in the Special Sciences." *British Journal for the Philosophy of Science* 51(2): 197–254. Cited from the published article as served on the author's website. <https://www.jameswoodward.org/s/BJPS2000.pdf>
- Woodward, James. 2014. "A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters—Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment)." *Philosophy of Science* 81(5): 691–713. Section numbers cited in the text refer to the author's preprint, PhilSci-Archive, <https://philsci-archive.pitt.edu/10978/>. <https://doi.org/10.1086/678313>
- Woodward, James. 2018a. "Normative Theory and Descriptive Psychology in Understanding Causal Reasoning: The Role of Interventions and Invariance." In *Philosophy of Psychology: Causality and Psychological Subject*, 71–102. De Gruyter. Section numbers cited in the text refer to the author's preprint (2015), PhilSci-Archive, <https://philsci-archive.pitt.edu/11629/>. <https://doi.org/10.1515/9783110576054-003>
- Woodward, James. 2018b. "Some Varieties of Non-Causal Explanation." In *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*, edited by Alexander Reutlinger and Juha Saatsi, 117–140. Oxford University Press. Section numbers cited in the text refer to the author's submitted version, PhilSci-Archive, <https://philsci-archive.pitt.edu/15497/>. <https://doi.org/10.1093/oso/9780198777946.003.0007>
- Woodward, James. 2023. "Causation and Manipulability." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/causation-mani/> Accessed 2026-10-29.
- Woodward, James and Kenneth Kendler. 2023. "Polygene Risk Scores: A Philosophical Exploration." *Philosophy of Medicine* 4(1): 1–21. <https://doi.org/10.5195/pom.2023.156>
- Woodward, James and Lauren Ross. 2025. "20th Century Theories of Scientific Explanation." *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Summer 2026 archived edition. <https://plato.stanford.edu/archives/sum2026/entries/scientific-explanation-20th/> Accessed 2026-10-29.
- Woodward, James. 2026. "Reflections on Some Strategies for Causal Inference in Psychiatry." In *Causal Concepts in Psychopathology: Multidisciplinary Perspectives*, edited by Kenneth S. Kendler, Josef Parnas, and Peter Zachar, 8–41. Cambridge University Press. Page numbers cited in the text refer to the author's manuscript in PhilSci-Archive, not to the published chapter. <https://philsci-archive.pitt.edu/22951/>
- Wray, K. Brad. 1998. "Defending Longino's Social Epistemology." In *Twentieth World Congress of Philosophy, Boston, 10–15 August 1998*. Congress paper in the Paideia archive hosted by Boston University; cited by section. <https://www.bu.edu/wcp/Papers/TKno/TKnoWray.htm> Accessed 2026-10-31.
- Wu, Jingyi and Cailin O'Connor. 2023. "How should we promote transient diversity in science?" *Synthese* 201(2): 37. <https://doi.org/10.1007/s11229-023-04037-1>
- Wylie, Alison. 2015. "A Plurality of Pluralisms: Collaborative Practice in Archaeology." In *Objectivity in Science: New Perspectives from Science and Technology Studies*, edited by Flavia Padovani, Alan Richardson, and Jonathan Y. Tsou, 189–210. Springer. Read in the author's manuscript on PhilPapers; page numbers are those of the manuscript. <https://philpapers.org/archive/WYLAPO-2.pdf>
- Xia, Bo, Weimin Zhang, Guisheng Zhao, Xinru Zhang, Jiangshan Bai, Ran Brosh, Aleksandra Wudzinska, Emily Huang, Hannah Ashe, Gwen Ellis, Maayan Pour, Yu Zhao, Camila Coelho, Yinan Zhu, Alexander Miller, Jeremy S. Dasen, Matthew T. Maurano, Sang Y. Kim, Jef D. Boeke, and Itai Yanai. 2024. "On the genetic basis of tail-loss evolution in humans and apes." *Nature* 626(8001): 1042–1048. <https://doi.org/10.1038/s41586-024-07095-8>
- Yamamoto, Angella. 2018. "Collaborative Scientific Knowledge and Testimonial Justification." University of Waterloo. MA thesis, UWSpace, University of Waterloo; page numbers are those printed in the thesis. <https://dspacemainprd01.lib.uwaterloo.ca/server/api/core/bitstreams/a7ecf2d3-9087-412b-9aa5-b9daa3f9bc31/content> Accessed 2026-10-31.
- Yanai, Itai and Martin Lercher. 2019. "Night Science." *Genome Biology* 20(1): 179. <https://doi.org/10.1186/s13059-019-1800-6>
- Yanai, Itai and Martin Lercher. 2023. "32 | Marty Martin and Art Woods on Science Podcasting." *Night Science*. Podcast episode, 10 April 2023; statements are paraphrased from a transcript of the episode. <https://nightscience.buzzsprout.com/1744020/episodes/12619555-32-marty-martin-and-art-woods-on-science-podcasting> Accessed 2026-10-03.
- Yanai, Itai and Martin Lercher. 2024. "54 | Bo Xia and a Tale of Tails." *Night Science*. Podcast episode, 28 February 2024; statements are paraphrased from the episode page and from an automatic transcript of the episode. <https://nightscience.buzzsprout.com/1744020/episodes/14595822-54-bo-xia-and-a-tale-of-tails> Accessed 2026-10-04.
- Yanai, Itai and Martin Lercher. n.d. "Night Science: Episodes." *Buzzsprout*. Night Science. Mutable episode index. <https://nightscience.buzzsprout.com/1744020/episodes> Accessed 2026-09-07.

- Ylikoski, Petri and Jaakko Kuorikoski. 2010. “Dissecting explanatory power.” *Philosophical Studies* 148(2): 201–219. Section numbers cited in the text refer to the SocArXiv preprint, <https://osf.io/download/w58dn/>. <https://doi.org/10.1007/s11098-008-9324-z>
- Yudkowsky, Eliezer. 2004. *Coherent Extrapolated Volition*. The Singularity Institute, San Francisco, CA. Cited from the PDF posted by the Machine Intelligence Research Institute (previously the Singularity Institute); page numbers are those printed on the pages. <https://intelligence.org/files/CEV.pdf> Accessed 2026-10-29.
- Yudkowsky, Eliezer. 2008. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308–345. Oxford University Press. Sections cited in the text refer to the Machine Intelligence Research Institute version, which states that it contains minor changes. <https://intelligence.org/files/AIPosNegFactor.pdf>
- Yudkowsky, Eliezer. 2022. “AGI Ruin: A List of Lethalities.” *LessWrong*, 5 June 2022. Forum post; cited by item number. <https://www.lesswrong.com/posts/uMQ3cqWDPHjtiesc/agi-ruin-a-list-of-lethalities> Accessed 2026-10-17.
- Zamora Bonilla, Jesús. 2014. “The Grand Bazaar of Wisdom (and 6): Mathematical Models in the Economics of Science.” *Mapping Ignorance*. Article published 17 March 2014; cited by paragraph. <https://mappingignorance.org/2014/03/17/grand-bazaar-wisdom-6-mathematical-models-economics-science/>
- Zednik, Carlos. 2021. “Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence.” *Philosophy & Technology* 34(2): 265–288. Section numbers cited in the text refer to the preprint, arXiv:1903.04361. <https://doi.org/10.1007/s13347-019-00382-7>
- Zeni, Claudio, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jieliang Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. 2025. “A generative model for inorganic materials design.” *Nature* 639(8055): 624–632. <https://doi.org/10.1038/s41586-025-08628-5>
- Zhao, Bingchen, Sara Beery, and Oisín Mac Aodha. 2026. “Autonomous Scientific Discovery via Iterative Meta-Reflection.” *arXiv*. Preprint, version 1; page numbers cited in the text refer to this version. <https://arxiv.org/abs/2607.01131v1>
- Zimmerman, John, Jodi Forlizzi, and Shelley Evenson. 2007. “Research through design as a method for interaction design research in HCI.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 493–502. ACM. Cited without page numbers from the authors’ version, KiltHub, Carnegie Mellon University, <https://doi.org/10.1184/R1/6470393>. <https://doi.org/10.1145/1240624.1240704>
- Zollman, Kevin J. S. 2010. “The Epistemic Benefit of Transient Diversity.” *Erkenntnis* 72(1): 17–35. Page numbers cited in the text refer to the author’s version dated 29 September 2009, on the author’s website, https://www.kevinzollman.com/uploads/5/0/3/6/50361245/zollman_-_transient_diversity.pdf; its section and note numbers are the same as in the published article, but its pagination differs. <https://doi.org/10.1007/s10670-009-9194-6>
- Šešelja, Dunja. 2022. “Agent-based models of scientific interaction.” *Philosophy Compass* 17(7): e12855. <https://doi.org/10.1111/phc3.12855>