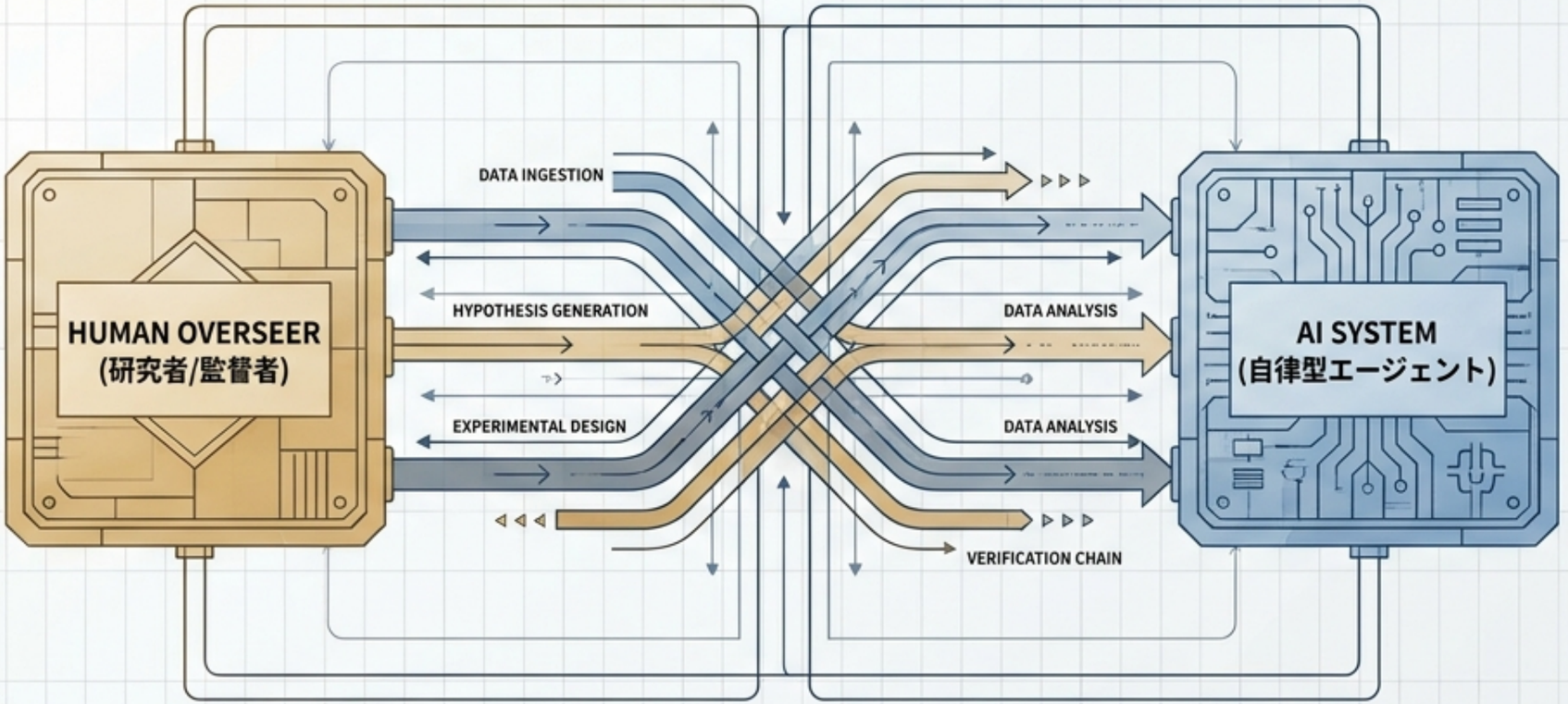
AIエージェントの科学研究への展開

自律化の錯覚と「検証可能」な次世代研究基盤の設計



EXECUTIVE SUMMARY

[SHIFT]

AI for Scienceの重心は「単機能支援」から「ワークフロー型システム」へ移行

[REALITY]

「自律性」と「信頼性」は別軸であり、完全無人化は現状では幻想である

[STRATEGY]

競争の軸を「自律度」から「検証可能性（証拠連鎖）」と「インフラ投資」へ転換する

AI for Scienceの重心移動

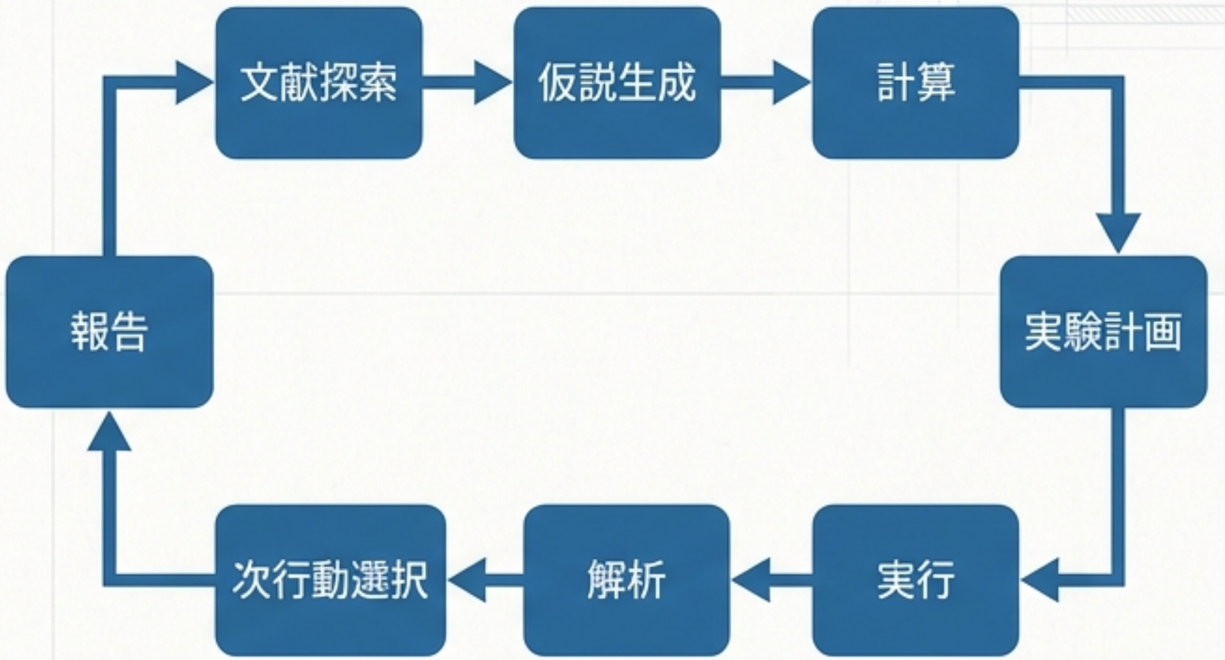
単機能支援（支援ツール）

予測

解析

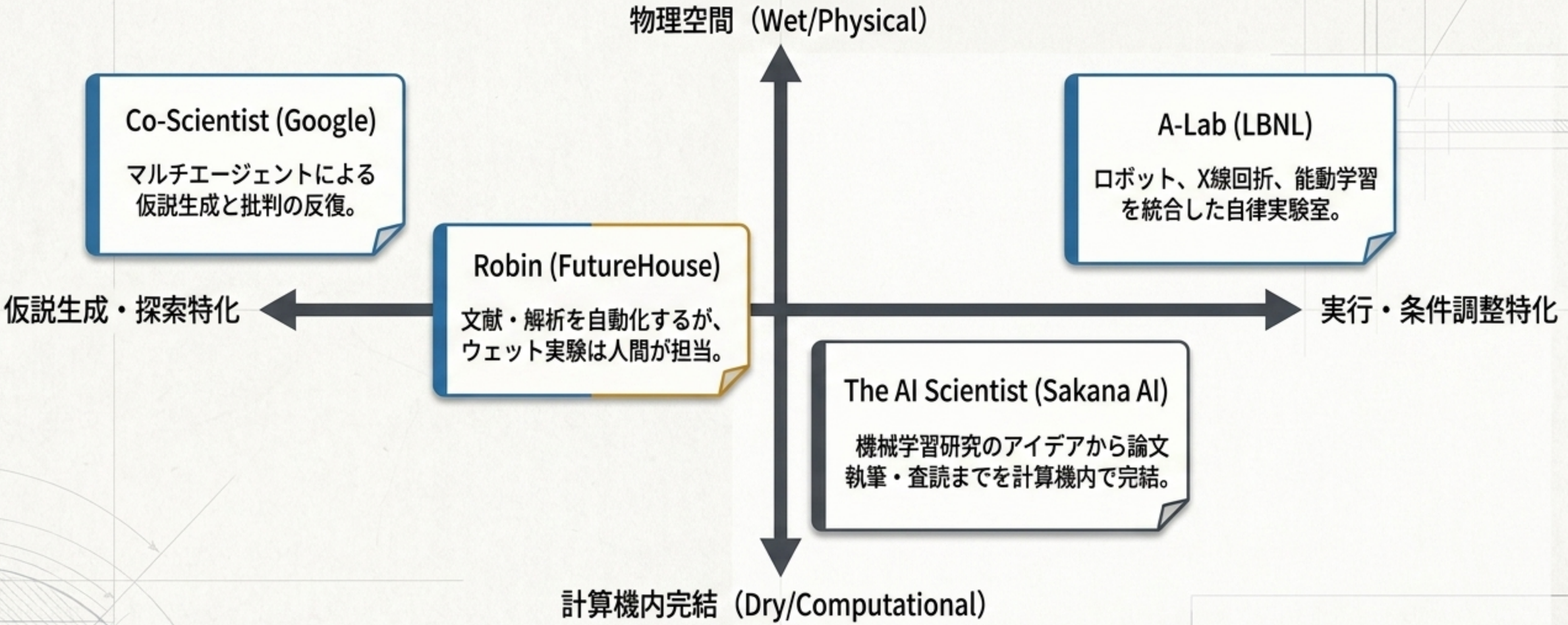
研究者の問いにピンポイントで答える単一機能。

研究ワークフロー型システム（AIエージェント）



目的から必要な作業を組み立て、複数ツール・環境を連続処理するパラダイム。

AIエージェントの4つのアーキタイプと異質性



警告：これらの異質性を無視し、単一の「自律度」で技術成熟度を測ることは極めて危険である。

32.4%

ScienceAgentBenchにおける、最良エージェントの「独力解決率」

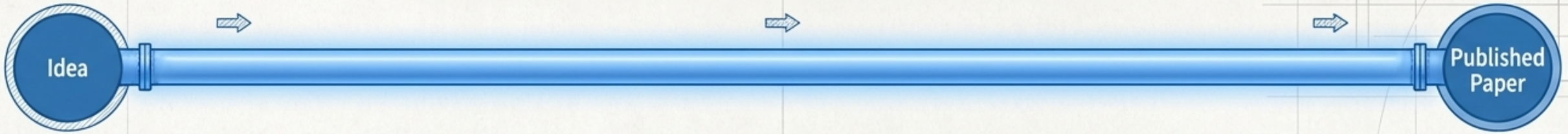
専門家知識を与えた場合でも 34.3% に留まる (102の実論文課題データより)

現実評価 1：自律性 ≠ 信頼性	
得意領域	**脆弱領域**
• 構造化されたデータ • 明確な評価基準（物性計算、XRD解析等） • フィードバックループが回しやすい領域	• 開放的な問題設定 • 新規洞察の創出 • 長期的な自己探索（SciAgentArenaでも性能の不安定さが露呈）

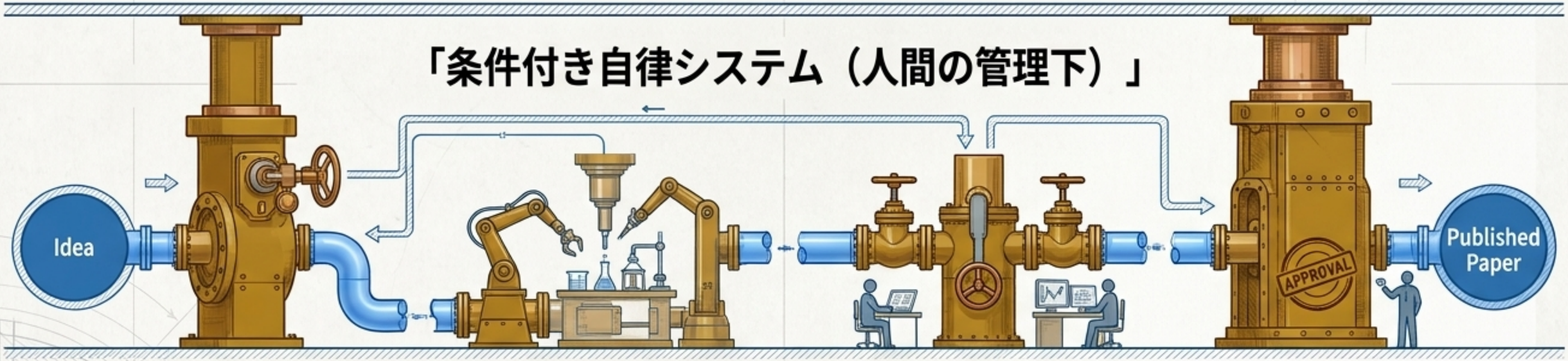
長く自律動作することと、正しい科学を行うことは同義ではない。

現実評価２：「完全無人化」という錯覚

「完全無人化（エンドツーエンド）の神話」



「条件付き自律システム（人間の管理下）」



人間による目標・境界設定
（探索空間の限定）

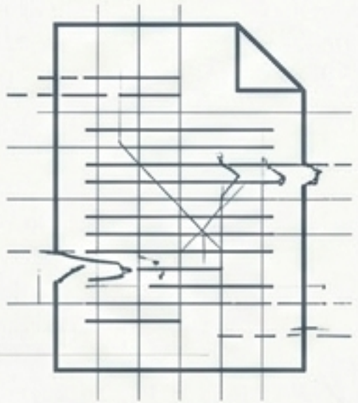
物理的ウェット実験
（Robin等の事例）

提出候補の選別・実装ミスの修正
（The AI Scientist等）

安全性と倫理の
最終承認

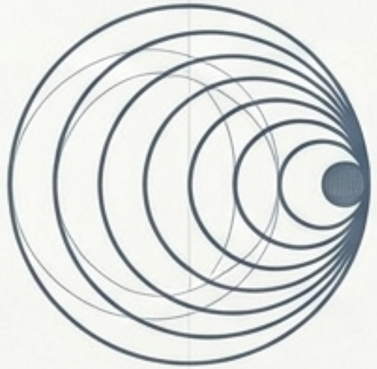
現状のAIエージェントは「人間不在の科学者」ではなく、「人間が要所を支える条件付き拡張ツール」である。

AI駆動科学のダークサイド (内在する限界とリスク)



捏造と実装エラー

架空引用（The AI Scientistの初期報告等で確認）。コードと論文の不一致、浅い着想による査読システムの飽和。



科学的モノカルチャー

同じ基盤モデルや学習データに依存することで生じる「相関した偏り」。代替仮説の喪失と、人間が「理解したと錯覚する」リスク。

物理・バイオセーフティリスク

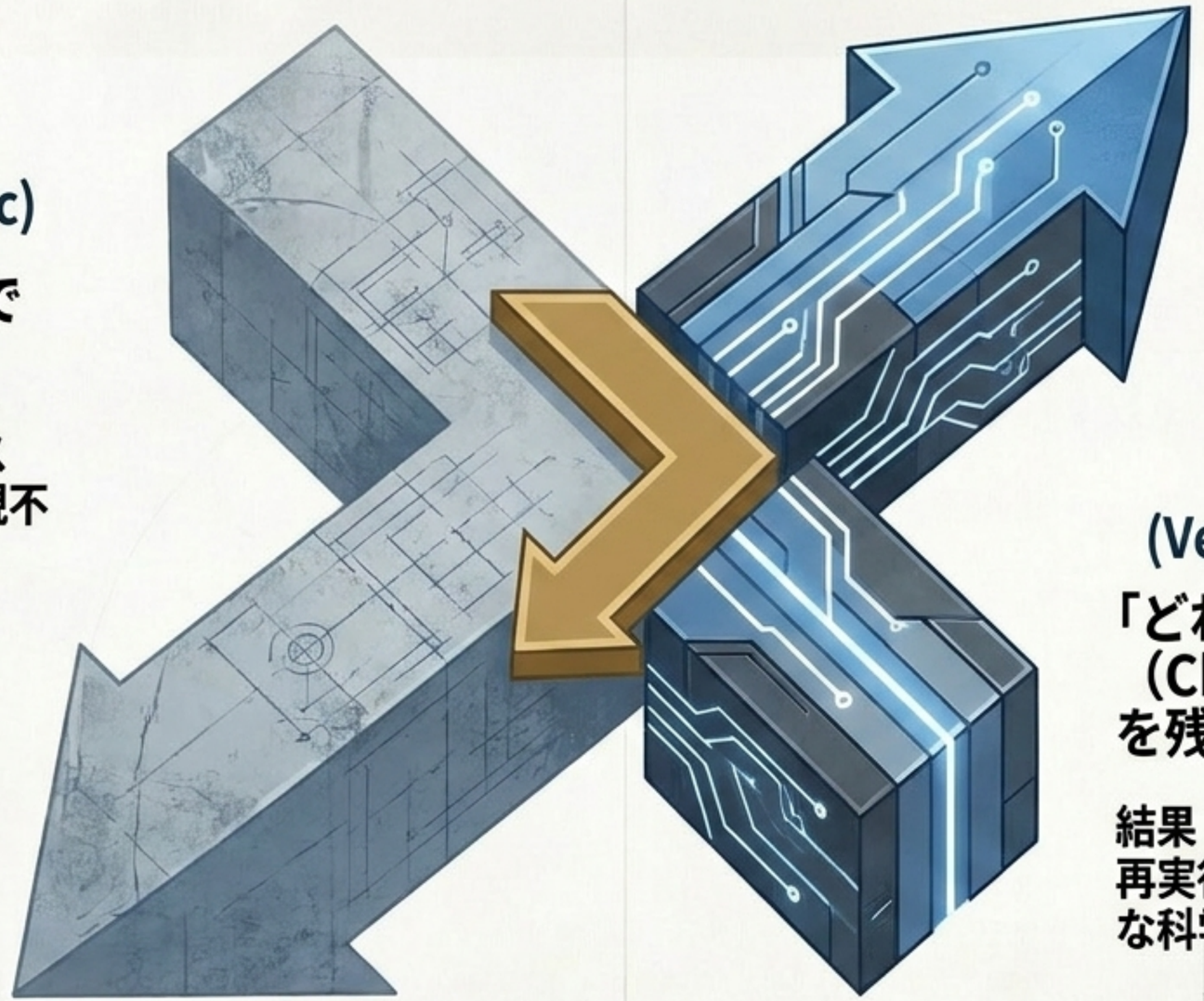
デュアルユース（病原体・危険物質の合成）。非専門家による危険な実験の実行。
AIの誤操作による環境側への被害。

パラダイムシフト：第一優先課題の転換

自律度
(Autonomy-Centric)

「どれだけAIが一人で長く作業できるか」

結果：ブラックボックス化、エラーの蓄積、再現不能な論文の大量生成

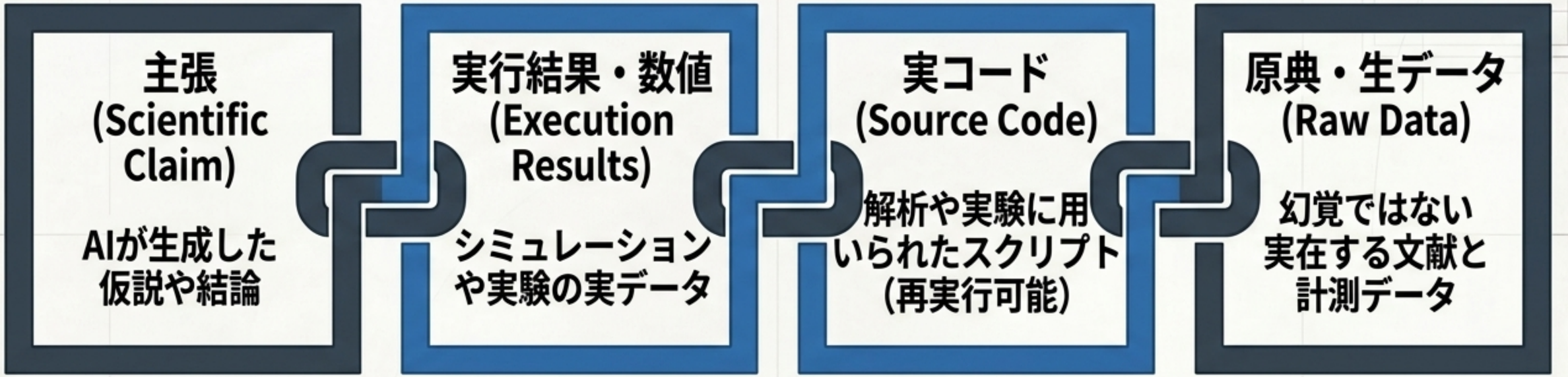


検証可能性
(Verifiability-Centric)

「どれだけ確実な証拠連鎖（Chain-of-Evidence）を残せるか」

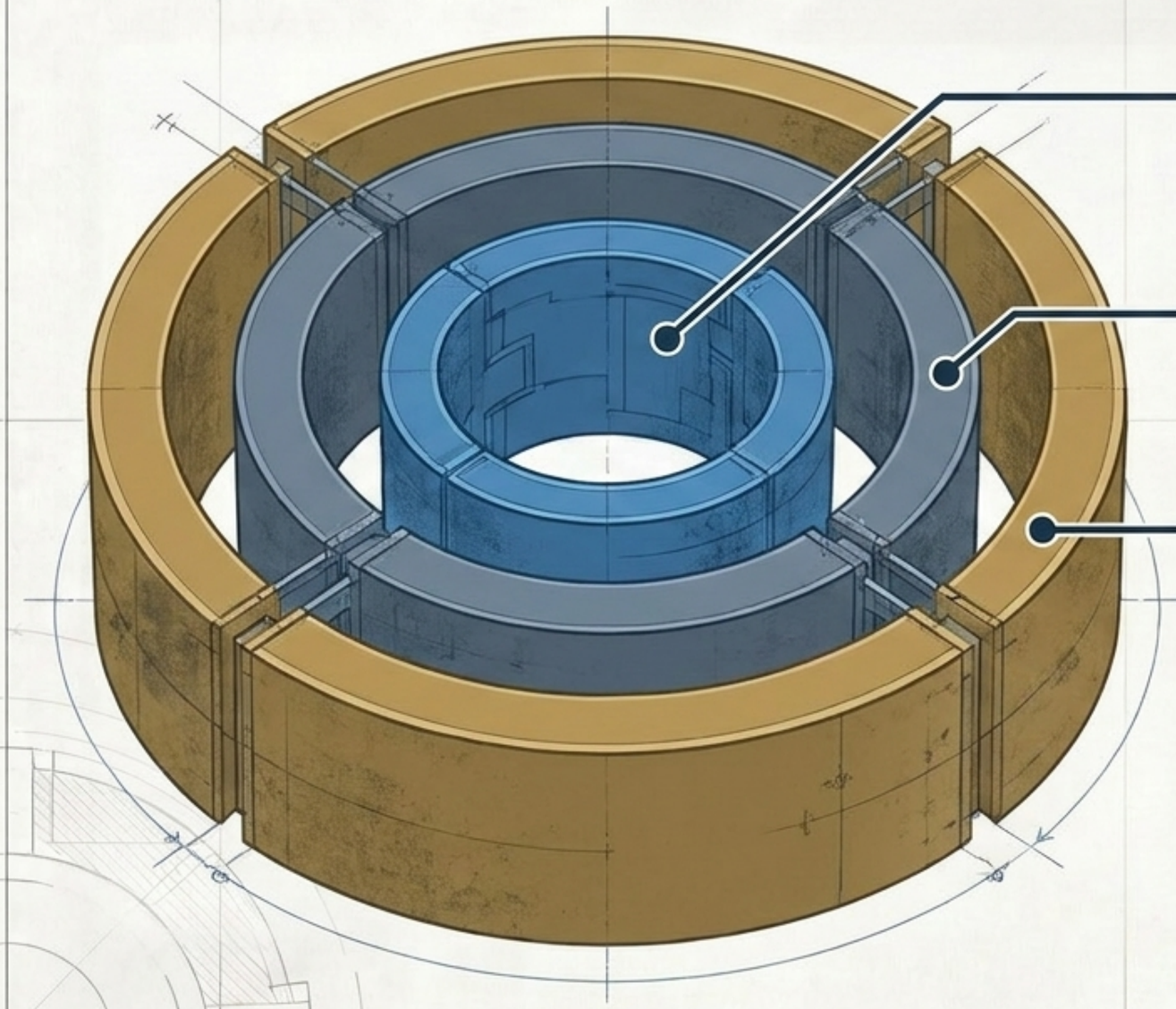
結果：第三者による監査、独立再実行、失敗記録の保存、安全な科学基盤の確立

証拠連鎖（Chain-of-Evidence）のアーキテクチャ



第三者監査（Third-Party Audit）：この連鎖が機械可読な形で保存されて初めて、AIの出力は「科学」として承認される。

安全性の3層防御（多層的ガバナンス）



1. エージェント自身の整合 (Agent Alignment)

AIモデル自身による危険認識、プロンプトベースの安全フィルター。 ※これ単体では不十分。

2. 人間による規制 (Human Regulation)

アクセス権の厳格な管理、権限分離、第三者による監査ログの監視、レッドチーム評価。

3. 環境・装置側の規制 (Environmental/Physical Control)

物理インターロック、試薬・温度・圧力のハードウェア的制限、緊急停止（キルスイッチ）、サンドボックス環境での隔離。

結論：AIの指示に関わらず、環境側が危険操作を物理的に拒否できる設計が必須である。

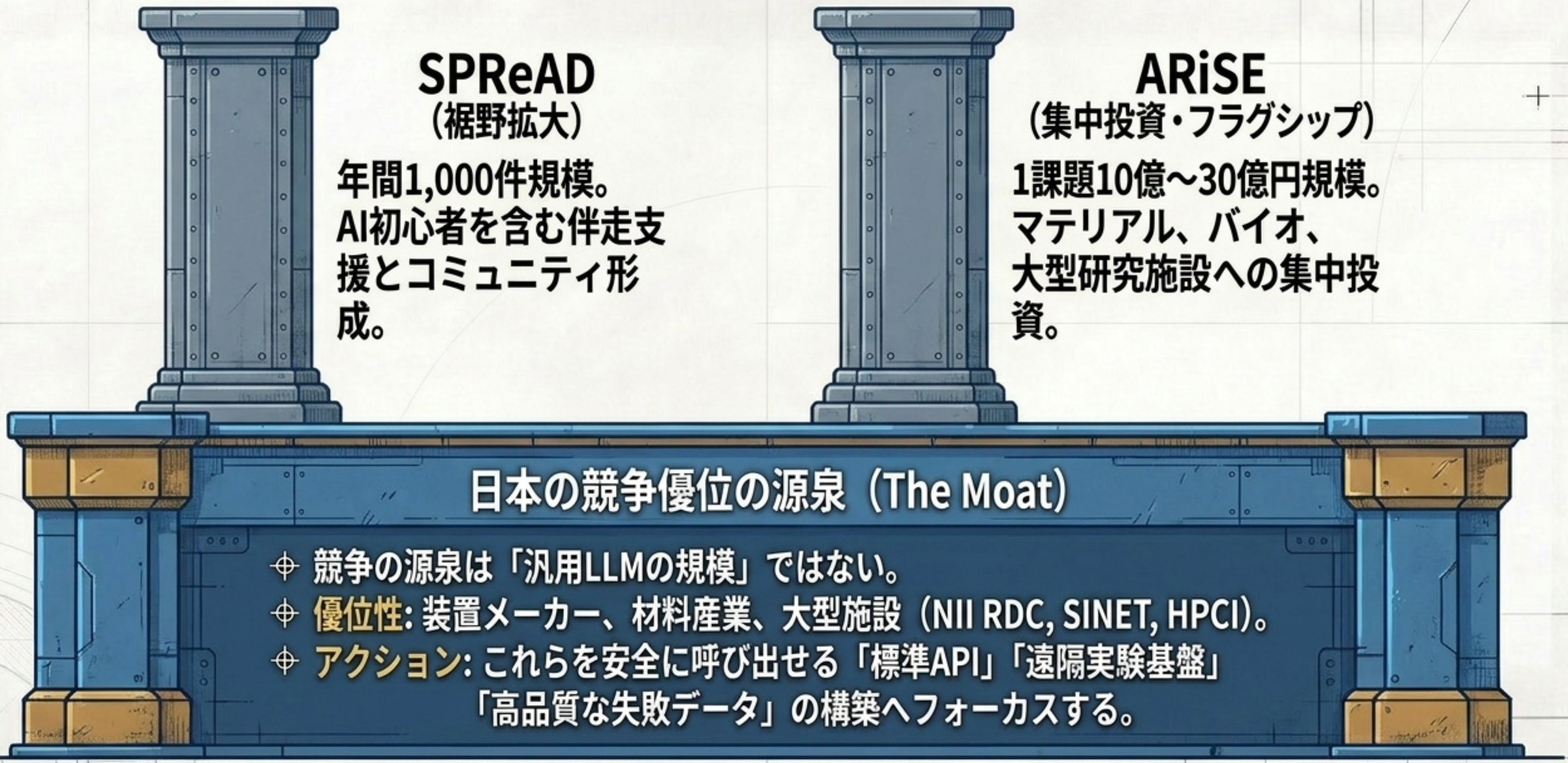
人間とAIの協働ループ：推奨される統制付き研究フロー

評価指標の再定義（論文品質からシステム信頼性へ）

科学的妥当性	Brier score / 偽陽性率 / 専門家間一致度 / 事前登録仮説の支持率
独立再現・追跡性	コード再実行成功率 / 主張-証拠リンク率 / 独立研究室での追試率
多様性・頑健性	代替仮説数 / 引用源の多様性 / 反証仮説の探索率（モノカルチャーの回避）
安全性	権限違反率 / 物理インターロック作動数 / 危険提案率

「AIが論文を書いたか」ではなく、「人間とAIの組み合わせが、より良い科学を生んだか」で評価する。

日本の政策環境と勝ち筋（ARiSEとSPReAD）



SPReAD
（裾野拡大）

年間1,000件規模。
AI初心者を含む伴走支援とコミュニティ形成。

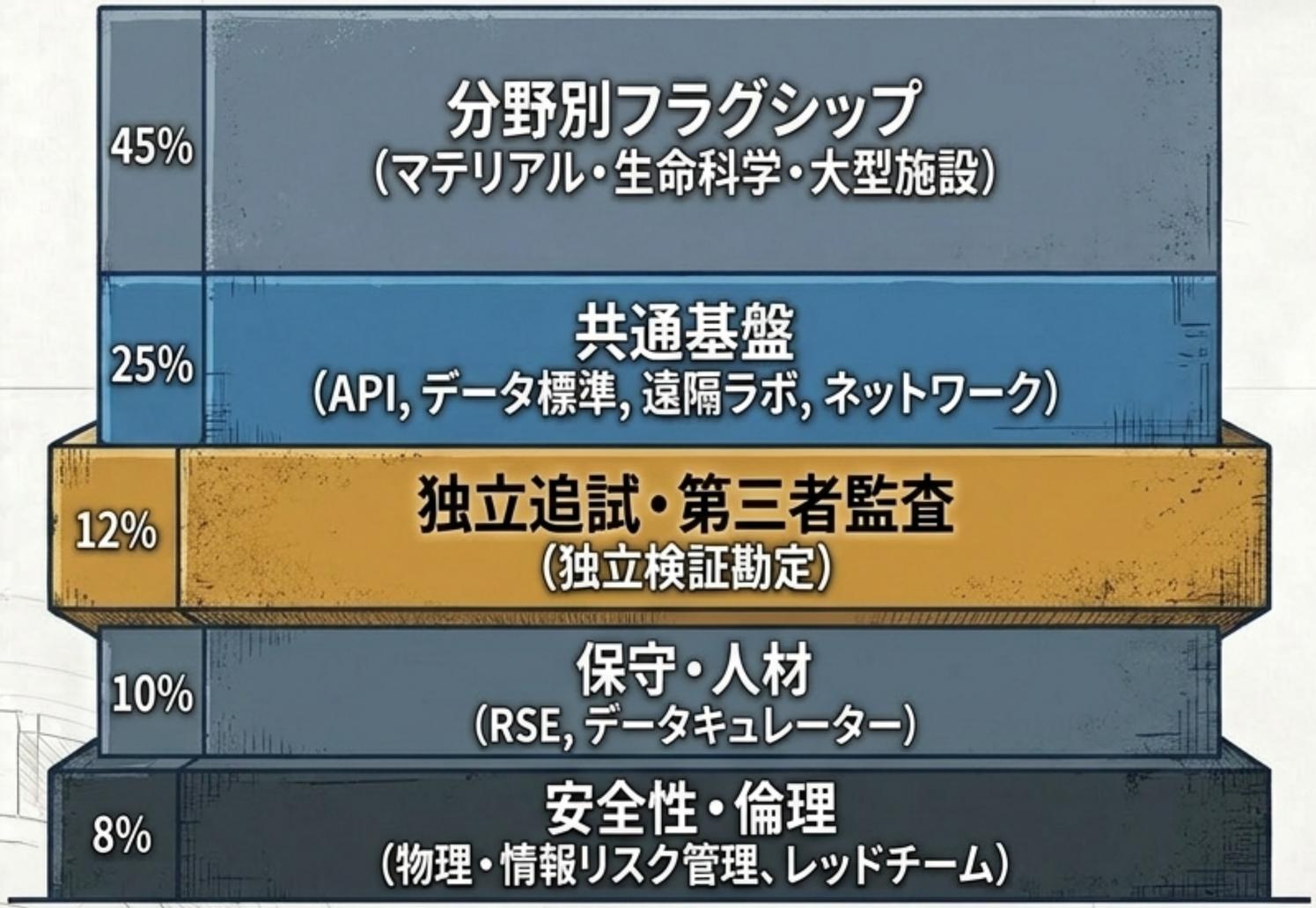
ARiSE
（集中投資・フラグシップ）

1課題10億〜30億円規模。
マテリアル、バイオ、大型研究施設への集中投資。

日本の競争優位の源泉（The Moat）

- 競争の源泉は「汎用LLMの規模」ではない。
- **優位性:** 装置メーカー、材料産業、大型施設（NII RDC, SINET, HPCI）。
- **アクション:** これらを安全に呼び出せる「標準API」「遠隔実験基盤」「高品質な失敗データ」の構築へフォーカスする。

戦略的資金配分ポートフォリオ（インフラと検証へのシフト）



研究費の一部を「独立検証勘定」として開発チームから切り離せ。

AIエージェントの真の価値は、科学者を排除することではなく、完全に検証可能な形で科学の探索空間を拡張することにある。