大阪大学 社会技術共創研究センター
Research Center on Ethical, Legal and Social Issues

ELSI NOTE No.77

2026 年 9 月 10 日

AIエージェントに
ガバナンスは追いつけるか

ハーネスと人間関与の再設計

Author

工藤 郁子　大阪大学 社会技術共創研究センター　特任准教授（2026年8月現在）

※ 本研究は、メルカリ R4D ラボ・大阪大学協働研究所の共同研究の一環として行ったものです。

大阪大学

ELSI

概要

AI エージェントの業務遂行能力が高まるほど、量・速度・複雑性の面で AI ガバナンスが追いつかなくなるおそれがある。

人間関与の形式のみにこだわると、誤った安心感を関係者に与え、統制の欠如が覆い隠されてしまう。「説明責任の茶番劇（accountability theater）」を回避し、人間関与を実効性あるものとするために、ハーネスと AI ガバナンスをどう再設計すべきか。

本研究は、プリンシパル・エージェント問題に対処する「臨床特権制度（clinical privileges）」を参照し、機能保存や正統性という観点から、具体的手法を提案する。また、人間の医師と AI エージェントの相違点も考察し、前提や制約条件を検討する。

目次

1. はじめに

1.1. ハーネス・エンジニアリングの勃興

AI エージェントの進展と普及に伴い[1]、AI モデルの活用方法だけでなく、AI エージェントの実行環境のデザインにも注目が集まっている。つまり、人間が AI モデルに指示するたびに工夫を凝らす「プロンプト・エンジニアリング（prompt engineering）」や、AI モデルがタスクを遂行するために必要な文脈情報を入力する「コンテキスト・エンジニアリング（context engineering）」というよりも、AI エージェントが自走できるように AI モデルを取り巻く環境全体を整えることが重視されるようになってきた。

このような文脈において「**ハーネス・エンジニアリング（harness engineering）**」という考え方が提唱された[2]。ハーネス・エンジニアリングには、現時点で確立された唯一の定義はない。ただし、もともとハーネスが馬を制御する馬具を指してきたため、制御フレームワークの比喩として理解されている。そこでは「人間が AI モデルをどう賢く使うか」を探るよりも、「AI モデルが間違えた時にそれをどう検知し、二度と同じ間違いをしないような仕組みをどのように構築するか」が検討されつつある[3]。

もっとも、現在のハーネス・エンジニアリングに関する議論は、主として実装技術に焦点を当てているため、**AI ガバナンス**とのあいだに見過ごせないギャップが存在する。

つまり、ハーネス・エンジニアリングによって AI エージェントによる業務遂行の自動化が進めば、アウトプットの速度と総量は増大する。ところが、従来の AI ガバナンスは、人間による個別的・逐次的な確認という「**ヒューマン・イン・ザ・ループ（human in the Loop, HITL）**」

1 本研究でいう「AI エージェント」は、さしあたり、特定の目標を達成するために最適な手段を選択して様々なタスクを自律的に遂行するものを指す。AI エージェント単体と、マルチエージェント・システム（multi-agent system）やエージェンティック AI（Agentic AI）を、区別せずに用いる。

2 Hashimoto, M. (2026). My AI adoption journey. Mitchell Hashimoto. https://mitchellh.com/writing/my-ai-adoption-journey ; Lopopolo, R. (2026). *Harness engineering: Leveraging Codex in an agent-first world*. OpenAI. https://openai.com/index/harness-engineering/

3 本研究では、議論の便宜のため暫定的に以下のように用語を整理している。まず、「モデル」とは学習済みのウェイトとそれに対する推論機能を指す。「ハーネス」とはモデルを取り巻く実行環境の総体を指し、ガードレールや監視・記録の仕組みなどを含む。「AI エージェント」とは、目標を与えられれば業務を自動遂行しうる状態にある稼働単位を指す。この整理では、AI エージェントの挙動は、モデルのみの性質というよりも、モデルとハーネスが結合した環境によって左右されることになる。なお、こうした整理や語用は確立されたものではなく、AI エージェントやツール群を含む上位概念としてハーネスを位置づける場合もある。

を念頭において設計されてきた[4]。よって、AI エージェントの業務遂行能力が高まるほど、量・速度・複雑性の面で **AI ガバナンスが追いつかなくなるおそれ**がある。

1.2. 悲観シナリオ

このような状況では、少なくとも **2 つの悲観シナリオ**が想定される[5]。

第一に、形式上は人間が責任を引き受けるように設計されているにもかかわらず、実際には AI エージェントの業務量や処理速度に圧倒され、「自動化バイアス（automation bias）」に流されてしまい[6]、人間が介入する余地が失われる可能性がある。この場合、人間は、実質的な統制権限を持たないまま、事後的に責任だけを負わされる「**モラル・クランプル・ゾーン（moral crumple zones）**」となってしまう[7]。

第二に、人間による確認の速度が業務遂行の速度に追いつかず、**審査手続きの迂回や形骸化**が定着する可能性がある。これは、必ずしも現場の規制逃れという形で現れるとは限らない。どちらかといえば、HITL を原則として掲げながら、業務効率を維持するために例外処理が常態化して自動審査が大半を占めるようになるだろう。これは、明示的な方針変更がされていないにもかかわらず、AI ガバナンスの実効性が徐々に失われていくことを意味する。

以上の 2 つのシナリオはいずれも、HITL を形式的に維持しているにもかかわらず、AI ガバナンスが実質的な機能を失うという帰結をもたらす。では、こうした**悲観シナリオを回避**するため

---

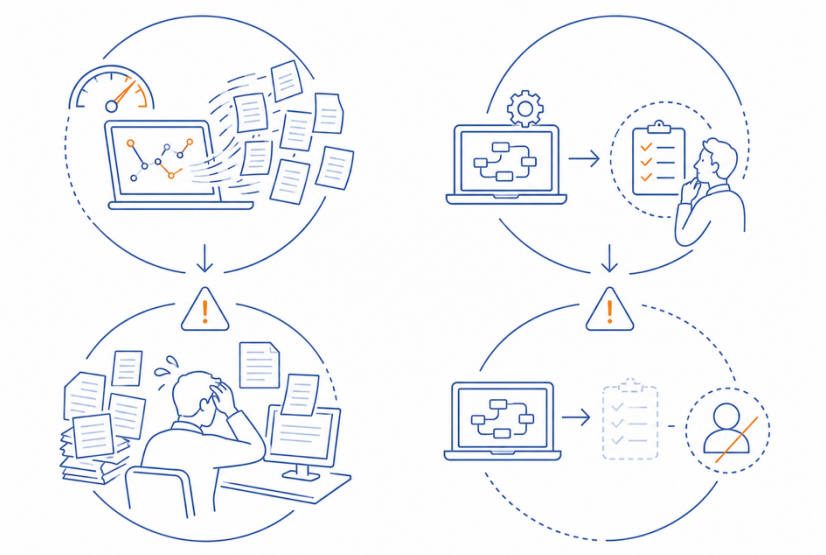
[4] ここでいう HITL は、典型的には、システムが提示した候補を人間が確認して実行することや、システムの出力を人間が承認しなければ次の処理に進まないこととさしあたり定義する。また、マクロとミクロの両方を指している。つまり、組織全体で採用される AI ガバナンスの基本方針としてのマクロ的側面と、個々のアクション（クラス）について人間の確認を要するかを判断するミクロ的側面の両方を含む。そして、本研究の主張を先取りして述べれば、HITL を維持するか撤廃するかというマクロ的二分法を退け、アクション単位の継続的判断というミクロ的なプロセスに転換することを提案している。

[5] 現状の人間関与の仕組みは、認知的な負担と安全性の確保という 2 つの要求の間で根本的なトレードオフを抱えており、利用者は、承認疲れと制御な AI エージェントの自律性の板挟みになっていることを指摘した先行研究として、Wang, P., Li, Y., & Tian, Y. (2026). *Reframing LLM agent security as an agent-human interaction problem*. arXiv. https://doi.org/10.48550/arXiv.2605.24309

[6] AI システム等の出力を過信して、人間が本来行うべき確認や判断を怠ってしまう心理傾向のこと。

[7] 「クランプル・ゾーン」は、もともとは、自動車の衝突事故発生時に衝撃を吸収することで運転者や乗客を保護する部位を指す。そこから派生して、AI システム等の利用についてトラブルや失敗が起きた際、AI システムの近くにいる人間（オペレーターなど）に責任や非難が不当に押し付けられることを「モラル・クランプル・ゾーン」と呼ぶ。Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society, 5*, 40–60.

に、どんなことを検討すればよいだろうか？



1.3. 2 つの問い

本研究の中心的な問いは、AI エージェントの業務遂行を支えるハーネス・エンジニアリングにおいて、人間を単なる形式的承認者に陥らせることなく、また、審査手続きの形骸化や迂回も防ぎながら、**実効的な AI ガバナンスをどのように再設計できるか**、という点にある。

この問いは、以下の副次的な問いに分けることができる。

第一に、AI エージェントの業務遂行において、**HITL の妥当性を判断するための基準やプロセス**はどのように設計されるべきか。また、その判断は状況の変化に応じてどのように継続的に見直されるべきか。

第二に、HITL が妥当でないと判断された場合、形式的な承認や審査の迂回を招くことなく、**実効的な人間の関与**を確保するために、どのように AI ガバナンスを再設計すべきか。

この 2 つの問いを通じて、本研究は、HITL を維持するか放棄するかという二分法ではなく、人間関与が実効的に発揮されるための AI ガバナンスのあり方を検討し、AI エージェントを統制するためのデザインを提案する。

2. HITL に関する判断プロセス

2.1. 規範的必要性と実効性

どのような場合であれば HITL が妥当なのかを判断するために、本節では「**規範的必要性**」と「**実効性**」という二つの判断軸を提案する。

規範的必要性とは、人間が統制すべきだという規範上の要請の強さを指す。他方、実効性とは、人間が実際に意味のある関与を行いうるかどうかを指す[8]。

本研究が注目するのは、**規範的必要性は高いが実効性が低い**状況である。冒頭の悲観シナリオは、まさにこれに当たる[9]。

つまり、規範的必要性が高いために HITL を制度として残そうとするが、実際には人間関与の実効性は失われている。その結果生じるのは、AI ガバナンスの形式だけが残り、その機能が欠けた状態だ。これは、いわゆる「**説明責任の茶番劇（accountability theater）**」にあたり[10]、人間が承認したという記録それ自体が、実効的な統制が存在しないという事実を覆い隠してしまうおそれがある。

他方、人間が統制すべき規範的必要性が高く、かつ、人間が実際に意味ある関与ができる実効性も高い類型は、HITL が妥当である。典型例としては、不可逆的かつ影響の大きい決定でありながら、人間が十分な時間と文脈を与えられて熟慮できるものが挙げられる。

実のところ、本節の主要な主張は、規範的必要性が高いというだけで、両者を同視すべきではないという点にある。HITL が制度として存在することと、HITL が実際に機能していることは区別することができるし、区別されるべきだ。実効性の低下に気づかずに、漫然と放置してしまう

---

[8] 実効性の確保に注目する先行研究として、Langer, M., Baum, K., & Schlicker, N. (2024). Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. *Minds and Machines*, 35(1), Article 1. https://doi.org/10.1007/s11023-024-09701-0

[9] 規範的必要性が低い場合については、その実効性の高低にかかわらず、そもそも人間が統制すべき必要性が乏しいため、HITL を導入する意義は小さい。これらについては、一定のガードレール導入を前提とした完全自動化などの仕組みで足りると考えられる。

[10] Ratanjee, V. (2026). *Why your accountability systems keep failing (and what actually works)*. Forbes. https://www.forbes.com/sites/vibhasratanjee/2026/01/31/why-your-accountability-systems-keep-failing-and-what-actually-works/

ことこそが、AI ガバナンスにとっての危機である[11]。

2.2. 規範的必要性の判断：正当化可能性とリスク分析

そこでまずは、**規範的必要性の判断過程**について概略を示したい。基本は**リスクベース・アプローチ**であるが、リスク分析に先立って問われるべきこともあるため、そのプロセスは 3 段階になる。

2.2.1. 許容可能性

第一に、ある個人の権利利益に対する制約が発生するかどうか、そして制約が発生する場合にそれは**正当化可能かどうか**が問われる。すなわち、およそ正当化できないものであれば、それは規範的必要性の判断に入る前に、棄却される[12]。これは、EU AI 法において禁止類型（同法 5 条）がハイリスク類型（同法 6 条以下）とは別建てになっていることと整合的である。

なお、一定の条件のもとでのみ正当化されるという場合には、その条件の充足を担保する必要があること自体が、規範的必要性を高める。この点については、後述する。

2.2.2. 法令・権利

第二に、**法令上の規制**や基本的権利に関わる要請などが存在する場合には、それ自体が規範的必要性を基礎づける。

2.2.3. リスク分析

第三に、**リスク分析**を行う。概括的にいえば、弊害の性質・発生可能性・影響度などに応じて、人間が統制すべきという規範的要請の高低を評価する。

その際に考慮すべき要素として、まず、（a）**弊害の発生確率と影響の大きさ**が挙げられる。ここは、一般的なリスク分析と同じであるものの、AI エージェントの特徴を踏まえると、2 点の

[11] なお、この分類には留意すべき点がある。二つの判断軸の境界線は、実際には連続的であり、論争の余地を含んでいる。よって、その判断自体が、AI ガバナンスの内実を左右することになる。この分類を誰が担うのか、すなわちメタレベルでの正当性・正統性の所在をどう考えるかについての概略は本文で示したが、多くの部分は、実証的に考察すべきであり、今後の検討課題となる。

[12] 言い換えると、個人への権利利益の制約という弊害を残余リスクとして引き受け、その大きさを管理するというリスク管理の発想は、その制約自体は許容されるという前提に立つ点において、基本権保護法制との親和性が低い。

留保がある。

その一つは**実行頻度**である。AI エージェントは同種のアクションを高速かつ大量に反復することができるので、従前よりも、実行頻度についての評価に注意を払うべきだ。言い換えると、一件あたりの弊害が小さくとも総量としては弊害が大きくなるおそれがある。

もう一つは「**テールリスク（tail risk）**」だ。システムは、通常時には有効に機能していても、ごく稀に誤作動を起こす可能性がある。そして、発生確率がわずかであっても、実際に起きた場合に甚大な危害をもたらすテールリスクへの備えは、規範的要請が強い。言い換えると、発生確率と影響の大きさを単純に掛け合わせて期待値として評価すると、テールリスクが過小評価されてしまうおそれがある。

次に、弊害の発生確率と影響の大きさ以外に考慮すべきこととして、（b）**不可逆性**も挙げられる。事後に取り消し・巻き戻しができないアクションほど、事前の統制を要する。

加えて、（c）**外部性**も事前に分析すべきである。弊害が AI エージェントの運用主体（例えば事業者）自身に帰着するのであれば、負担を受ける主体と統制を担う主体とが一致するが、その外部にいる個人（例えば事業者が提供するサービスを利用するエンドユーザー）や社会全体が不利益を被る場合には、規範的必要性は高まる。言い換えると、ここで問われているのは、弊害の大きさではなく、負担する主体と決定する主体とが一致するかという構造である。

項目	考慮要素	判断
許容可能性	権利利益への制約の有無、制約の正当化可能性、正当化条件の有無など	正当化できないなら、そもそも棄却（禁止）
法令・権利	法令上の規制、基本的権利に関する要請	規制や要請がある場合、規範的必要性が高い
リスク分析	発生可能性、影響度（ただし、実行頻度、テールリスク、不可逆性、外部性などに注意）	リスクが高いほど、規範的必要性が高い

2.3. 実効性の点検：ゲートキーピングと品質保証の区別

本研究では、従来の HITL において人間に期待されてきた役割を、「**ゲートキーピング（gatekeeping）**」と「**品質保証（quality assurance）**」という性質の異なる二つの機能に分けることを提案する。結論を先取りしていうと、ゲートキーピングは主として規範的必要性の問題であるが、品質保証は主として実効性の問題である。

まず、ゲートキーピングとは、人間が確認することで不可逆的な弊害の発生を予防し、万一発生してしまった場合には責任を引き受ける役割である。

このゲートキーピングの技術的対応として、例えば、稼働中のシステムを即座に停止させる安全装置としての「**キルスイッチ（kill switch）**」や、人間による訂正や停止の指示を AI システムが妨げずに受け入れる性質としての「**コリジビリティ（corrigibility）**」研究などが挙げられる[13]。

次に、品質保証とは、人間が確認することでアウトプットの品質を高める役割である。これは、「人間の関与が実際にアウトプットの改善に寄与するか」という**実証的な問い**であり、実効性に関わる。

この点について、人間と AI の組合せが、常に単独の人間または単独の AI の性能を上回るわけではないことを示したメタアナリシスがある[14]。とりわけ、AI 単独の性能が人間単独の性能を上回るタスクにおいては、人間と AI の組合せは平均的に性能損失を示す傾向がある。これは、AI が相対的に得意なタスクについて、形式的に人間の判断を組み込むと、かえって実効性を損なう可能性を示唆している。

つまり、**品質保証機能を人間に期待することは、見直す余地が大きい。**これまでは人間関与の意味があると見なされてきたものが、実際のところは実効性が低いと判明して、HITL を維持することが妥当でないと判明することが想定される。

---

[13] Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015, January). *Corrigibility*. In AAAI Workshop: AI and Ethics.

[14] Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour, 8*, 2293–2303.

2.4. 留意点：カテゴリーに分けることの弊害

HITL の妥当性判断について、いくつか留意点がある。まず、その判断は、一度実施すれば終わりというわけではなく、**継続的に見直されるべき**だ。

AI エージェントの業務遂行において、HITL をどのように位置付けるかを検討する際、実務において採用されやすいアプローチは、業務類型ごとに HITL の要否をあらかじめ定めることである。例えば、「顧客への外部送信を伴う業務は人間の承認を必須とする」「社内文書については自動化してよい」といった形で、業務内容に応じてカテゴリを設定し、それぞれに人間関与の要否を紐づける発想である。このアプローチは、判断基準の明確化、審査コストの削減、監査可能性の向上などの点で、一定の合理性を持ち、AI ガバナンス設計の出発点として採用されやすい。

しかし、この**類型的アプローチには看過できない問題**がある。

第一に、AI エージェントが遂行する業務のリスクは、業務の種類そのものよりも、実行時の文脈に強く依存する。同一の類型に属する業務であっても、文脈が異なればリスクの大きさが異なる。事前に固定された類型は、こうした文脈依存的なリスクの変動を十分には捕捉できない。なお、この**文脈依存性**への対応については、後ほど提案する。

第二に、類型を定めると、例えば、「HITL が必要な類型に、なるべく該当しないように業務を設計・記述する」という**迂回のインセンティブ**が生じる。類型的な分類アプローチは、審査を厳格化する意図で導入されたとしても、結果として審査の形骸化や回避行動を誘発する可能性がある。

第三に、類型は一度定められると、いわゆる**制度的慣性により更新されにくい**。AI エージェントの能力や利用範囲は急速に変化するにもかかわらず、類型や分類がその変化に追随できなければ、AI ガバナンスは実態から乖離してしまう。

こうした問題は、理論的な懸念にとどまらない。筆者が AI ガバナンスの実務担当者に向けて、本稿の草案を共有した際の体験を踏まえている。

当初、筆者は、規範的必要性と実効性という本研究で提案する新概念を説明するために、下記のような 4 象限の分類表を用いていた。すると、判断プロセスの重要性や定期的な見直しについて本文中で強調していたにもかかわらず、一部の実務担当者の質問やコメントが、「HITL を維持すべき類型をどのように特定すべきか」に集中してしまった。これは、分類的表現を用いてしまうと、**カテゴリカルな思考へとミスリード**してしまう可能性を示唆している。

	実効性が高い	実効性が低い
規範的必要性が高い	(1) HITL を維持	(2)再設計(機能保存が必要)
規範的必要性が低い	(3) HITL 任意	(4) 完全自動化

以上を踏まえると、AI エージェントの業務遂行に関する実効的なガバナンスを構想するにあたっては、HITL の妥当性を事前的・固定的な分類として扱うというよりも、**状況に応じて継続的に見直される判断プロセスとして設計・記述**する必要がある。

2.5. 留意点：実効性の劣化

実効性について、時間の経過とともに低下する可能性がある点にも留意する必要がある。つまり、実効性は、静的な性質ではなく、運用のなかで劣化しうる。

例えば、**審査対象の総量が増加したにもかかわらず、審査担当者の人数が変わらない**のであれば、担当者は疲弊し、確認作業は次第に追認的なものへと形骸化する。疲弊した人間の判断は独立した視点を失い、AI エージェントの出力にそのまま引きずられやすくなって、実効性を失う。

また、審査担当者に与えられる**時間・文脈・情報が不足**すると、人間は十分な根拠をもって判断することができなくなる。これも、性能の問題として捉えられる。すなわち、本来であれば人間単独でも一定の判断能力を発揮しうる場面であっても、時間・情報を奪われることで、実質的に人間単独の性能が低下し、AI に対する優位性を失った状態に陥る。

加えて、審査担当者に求められる**専門性や訓練が不十分**であれば、そもそも独立した判断を形成する能力自体が欠如する。これもまた性能の問題として位置づけられる。

したがって、審査対象について、持続可能な量的上限を定めること、十分な時間・文脈・ツールを保証すること、審査者の能力を担保すること、そして自動化バイアスへの対抗策を講じることなどを能動的に整えなければ、実効性は減退する。「HITL が妥当」という結論は、無条件に妥当するものではなく、こうした**リソース提供とセット**で初めて意味を持つ。

2.6. 留意点：審査補助システム

同時に、**人間関与を補助するための情報提示の自動化**も有効である。ただし、その設計には注意が必要である。

ここで自動化してよいのは、あくまで**前提事実の可視化**にとどまる。例えば、「このアクションを取り消すことは可能か」「弊害が生じた場合、どの範囲まで波及しうるか」といった判断の前提となる事実を整理し提示する機能だ。

これに対し、「どう判断すべきか」という評価・推奨そのものを AI エージェントが生成し、それを人間に提示する設計は避けなければならない。これはいわば、人間と AI の「**間違え方の相関性**」と関係する。仮に、人間の失敗パターンが AI エージェントの失敗パターンと独立であれば、両者を組み合わせることで冗長性が確保される。**人間関与はダブルチェックとして機能**して実効性が高まるといえる。他方、人間と AI の失敗パターンが類似しているのであれば、人間の関与が独立した冗長性を提供しないから、実効性が希薄になる。

ここまでの議論をまとめると、HITL の妥当性は、規範的必要性と実効性がともに高い場合に認められる。しかし、その妥当性の判断は、状況に応じて継続的に見直される。その際、実効性を維持するためのリソース提供という視点を見逃してはならない。実効性が低下してしまった状況で、それでもなお HITL の形式に固執するならば、統制はかえって形骸化する。

この実効性の低下にどう対処するかが、次の課題である。

3. AI ガバナンスの再設計

3.1. 機能保存：分類問題から設計問題へ

規範的必要性は高いにもかかわらず実効性が伴わないケースは、ハーネス・エンジニアリングにおいて挑戦課題となっている。そして、対策としてすぐに思いつく二つの選択肢は、いずれも別の弊害を生む。

第一に、形式的に HITL を維持し続けるという選択肢である。しかし、前述した通り、統制が機能しているという誤った安心感を与え、問題を覆い隠してしまう。この点、政府による AI システム利用に関する先行研究において[15]、人的監視を義務づける政策は、弊害から人々を守るのではなく、実際には誤った安心感を関係者に与えてしまい、行政機関やベンダー企業が説明責任を回避することを可能にしてしまうと指摘されており、人的監視から制度的監視への転換が主張されている[16]。

第二に、実効性が見込めない以上、HITL を撤廃して自動化された仕組みに全面的に委ねるという選択肢である。しかしこれは、規範的必要性の低いケースに相当する対応を、規範的必要性が高いケースにそのまま適用することになってしまう。つまり、規範的要請を無視して切り下げることになり、不適切である。

よって、**実効性を改善・回復することも重要な選択肢**である。前述したような、審査対象について、持続可能な量的上限を定めること、十分な時間・文脈・ツールを保証すること、審査者の能力を担保すること、そして自動化バイアスへの対抗策を講じることなどがそれにあたる。

もっとも、本研究が取り組む課題は、こうした**改善・回復が見込めない場合**や、**形式に固執したためにかえって統制が弛緩する場合**など、HITL が妥当でない状況についてである。

そこで、「HITL の維持か撤廃か」という二分法ではなく、**工学的な安全設計によって、統制の実効性を保ったまま人間の関与を見直す**アプローチがあると主張する。

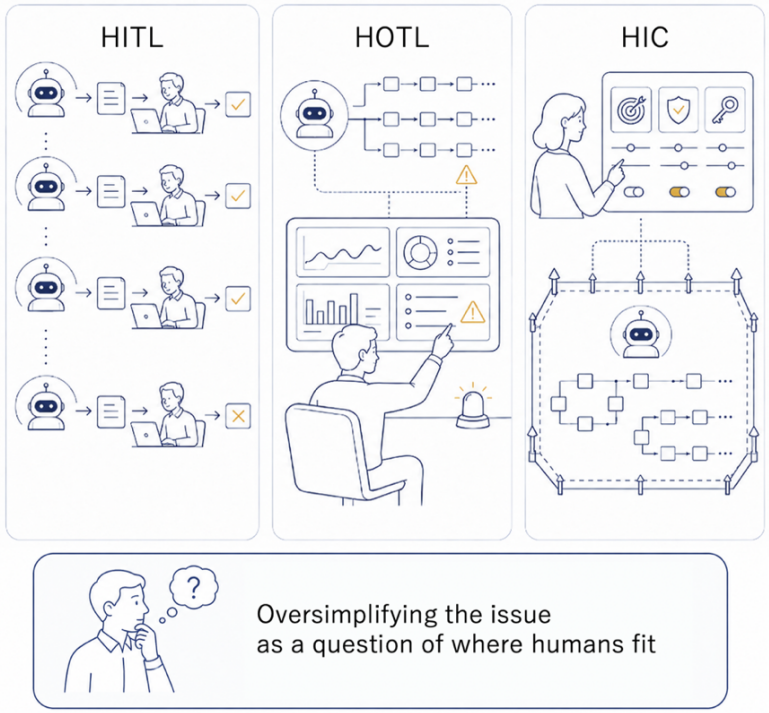
---

[15] Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45, Article 105681. https://doi.org/10.1016/j.clsr.2022.105681

[16] Green (2022) は、制度的監視として、主として利用開始時の審査について指摘していた。AI システム運用中の人間関与の方法について具体的に提案する点が、本研究の新規性である。

この主張は、「**ヒューマン・オン・ザ・ループ（human-on-the-loop, HOTL）**」や「**ヒューマン・イン・コマンド（human-in-command, HIC）**」などの概念に近い[17]。HOTL は、人間が設計段階に関与し、運用中はシステム全体を監視役として見守り、異常や問題が発生したときに介入することを指す。個々の判断ごとに介入しない点が HITL との違いである。HIC も、人間が個々の判断そのものに介入しないという発想に立つ。しかし、システム全体が従うべき目的・制約・権限の範囲を人間が定めることを強調するため、上位方針レベルでの統制に注目する点が HOTL と異なる。

もっとも、HITL/HOTL/HIC については、**人間関与をどこに設定するのかという位置の問題に矮小化**してしまい、実際の統制の複雑性や品質管理を捉えきれていないという批判がある[18]。これに対して、本研究は、後述するような手続的プロセスとして具体化する点で、こうした批判に応答することを試みている。



また、本研究は「**人間による有意の制御（meaningful human control, MHC）**」とも隣接し

[17] High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI*. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[18] Tsamados, A., Floridi, L., & Taddeo, M. (2024). Human control of AI systems: From supervision to teaming. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00489-4

ている[19]。MHC は、個々の操作に人間が介入をしなくとも、一定の条件を備えれば、道徳的責任と説明責任を確保できるという考え方である。

MHC に関する先行研究は、人間の統制を「意味のある」ものとするための必要条件として、**トラッキング（tracking）**と**トレーシング（tracing）**という 2 要件があると分析した。本研究では、そうした**条件が満たされなくなった場合にどう対応すべきか**を検討している。

ここまでの議論をまとめると、人間の監視や承認の一部を外しても、AI ガバナンスの実効性が損なわれないように再設計することは可能であると本研究は主張する。そして、規範的要請を切り下げるのではなく、工学的な安全設計によって能力を保ったまま人間の関与を見直すという発想を、本研究は「**機能保存（capability preservation）**」と呼ぶ。

この観点において、HITL/HOTL/HIC は、**分類問題ではなく設計問題**となる。

設計の指針は次の 3 点である。第一に、統制の規範的必要性だけでなく実効性の確保にも注目する。第二に、ゲートキーピングと品質保証という人間に期待される 2 つの役割を分離する。第三に、工学的手法によって人間が関与すべき領域を継続的に見直しつつ、実効性を確保する[20]。これにより悲観シナリオを回避する。

では、どのような工学的手法を導入して、人間の関与をどのようにデザインすれば、機能保存ができるのだろうか。

3.2. リスク評価の単位：カテゴリからアクションクラスへ

挑戦課題に答えるために、まずは統制の前提を問い直す必要がある。これまでの AI ガバナンスは、**事前にリスクを特定・評価できることを与件**としてきた。しかし AI エージェントでは、この前提が揺らいでしまう。

AI エージェントは、目標を与えられると、自律的に計画を立て、外部ツールやシステムと連携

[19] Santoni de Sio, F., & van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5, Article 15. https://doi.org/10.3389/frobt.2018.00015

[20] 第 3 方針に関する先行研究として、Cheng and Cheng (2026) などがある。ただし、統計的な傾向やパターンだけではなく、正統性も求めている点が本研究の新規性である。Cheng, E., & Cheng, J. (2026). A decoupled human-in-the-loop system for controlled autonomy in agentic workflows. arXiv. https://doi.org/10.48550/arXiv.2604.23049

しながら、複数のステップにわたって業務を遂行する。ここには、少なくとも三つの特徴がある。

第一に、設計者や利用者の想定を超えて実行する**創発性**である。第二に、付与された権限や接続先の外部システム、他のエージェントとの相互作用といった文脈によってリスクの大きさが変わる**文脈依存性**である。第三に、マルチエージェント環境において、個々のアクションの積み重ねの結果が予測困難になる**連鎖の不透明性**である。

そのため、**リスクを事前に特定・評価できることを与件とするリスクベース・アプローチでは対応しきれない**場合があると指摘されている[21]。誤解のないように付言すると、リスクベース・アプローチという発想そのものが不要になるわけではない。リスクの高低に応じて統制の強度を変えるという基本思想は、ハーネス・エンジニアリングでも引き続き有効である。問題は、そのリスク評価を行う単位の固定性または粒度にある。

この点、**システム運用中の介入設計**への注目が高まっている。シンガポールの情報メディア開発庁（Infocomm Media Development Authority, IMDA）が公開した「Model AI Governance Framework for Agentic AI」は、事前のリスク評価を維持しつつ、運用中の重要なチェックポイントで人間の承認を求めることで説明責任を確保する方向性を示している[22]。これは、単位の固定性という問題への一つの応答といえる。もっとも、チェックポイントの設計や人間関与のあり方については、さらなる検討や実証が求められる。

こうした課題に応えるものとして、「**アクションクラス（action class）**」という単位で管理することも提案されている。AI エージェントが遂行しうるアクションは無数にあるが、それらを個別に扱うのではなく、アクションの性質・リスクレベル・影響範囲といった観点から類似するものをまとめたグループが、ここでいうアクションクラスである[23]。

例えば、影響の性質に着目して「読み取り専用・分析系のアクション」「コミュニケーション・データ移動系のアクション」「状態変化を伴うアクション」「特権的または委任されたアクショ

---

[21] 松下尚史「AI エージェントの実用化に向けた論点の整理――安全・プライバシー・責任をどう確保するか」JIPDEC レポート（2026 年 4 月 22 日）

[22] Infocomm Media Development Authority. (2026, May 20). *Updated Model AI Governance Framework for Agentic AI*. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/updated-model-ai-governance-framework-for-agentic-ai

[23] 本研究では、影響範囲などを基準としたクラス分けをした上で、権限付与についてさらに分類するという構成を採っている。現状では、両者をあまり区別せずに混在して議論している場合があるようだが、AI ガバナンスの観点からは、明示的に区別して整理することが必要であろう。

ン」といったクラスを設けることが考えられる（下表参照）[24]。

アクション	例	デフォルトのガバナンス方針
読み取り専用・分析	検索、要約、分類、下書き作成、説明	アクセス範囲の限定、根拠付け、出典・来歴の記録、出力ログの保存
コミュニケーション・データ移動	メール送信、顧客への回答、ファイルのエクスポート、リージョン間転送	送信先の確認、コンテンツのガードレール、データ分類の確認、承認との紐付け
状態変更を伴うもの	作成、更新、削除、ワークフローステップの承認、SQL の実行、自動化の起動	実行前の拒否・承認判定、復旧可能性の確認、ポリシー判定、完全な意思決定記録
特権的・委任されたもの	返金、支払い、アクセス権の付与、セキュリティ変更、バックグラウンドジョブ、複数エージェントへの権限委任	原則拒否、または最も厳格なレビューを要求（必要に応じて権限範囲を限定した ID、二重統制、即時の証跡保存）

アクションクラスは、リスク評価の単位を、**大まかで静的な業務類型から、より細かく動的なタスクへ**と分類し直すものと位置づけることができる。

例えば、「送金処理は不可逆的で影響が大きいため、人間の承認を要する」という統制を考えてみよう。この判断がリスク評価として妥当である場合でも、その**評価結果を何にどう紐づけて実装するか**が問題になる。「送金業務」という業務類型に紐づけた場合、AI エージェントの性質を勘案すると、二つの方向に誤りが生じる可能性が高くなる。第一に、**過剰な方向**について、少

[24] Pusukuri, K. (2026, June 2). *Building trustworthy AI at Oracle*. Oracle AI & Data Science Blog. https://blogs.oracle.com/ai-and-datascience/building-trustworthy-ai-at-oracle

額の経費精算も、既知の取引先への定型的な支払いも、一律に人間による審査へ回されることが想定される。この場合、審査件数の膨張を通じて、かえって人間関与の実効性を損なうことになる。第二に、**過小の方向**について、「送金」という名目を持たないタスクが統制から漏れてしまうことが想定される。例えば、受取口座の登録情報を変更するタスクは、名目上は設定の更新にすぎないため、業務類型上は低リスクと判断されるかもしれない。しかし、口座情報が書き換えられれば、それ以後に実行される送金はすべて別の宛先に振り込まれてしまう。返金処理、支払期日の変更、与信枠の変更なども同様であり、事前に想定していた業務類型と、実際のタスクによって生じる帰結との間に、ずれが生じる。

これに対して、アクションクラスによる管理を導入すると、**業務区分ではなく、当該タスクが生じさせる帰結の性質に紐づける**ことができる。前述の例でいえば、受取口座情報の更新は、それ自体が資金を移動させるわけではないものの、資金の流出経路を規定するタスクである以上、送金の実行と同等以上のクラスに位置づけられる。そのリスク評価は、設計時に策定した業務類型によってではなく、実行時に呼び出されるタスクと関連指標（金額、宛先の新規性、一定期間内の累積額など）に即して行われる。

そして、この単位の違いは、統制手段の選択肢にも影響を与える。事前に策定した業務類型に紐づける方式であれば、その統制は「人間の審査を要するか否か」という二値になりやすい。これに対して、**アクションクラスは権限の払い出しの単位でもある**ため、当該クラスに関するいくつかの手段（累計額の上限、組戻しが可能な期間内での実行、別系統による検証、事後の抜き打ち監査など）を、必要に応じて組み合わせることができる。そこでは、**人間関与は複数ある統制手段の一つ**と位置づけることがしやすくなる。本研究がいう機能保存は、この選択肢の広がりの上に成り立っている。

もっとも、現在のアクションクラスに関する議論は、規範的必要性の高低を問わずに全てを対象としており、**規範的必要性が高い領域に特化した分析はまだ少ない**[25]。そこで本研究では、当該領域でのアクションクラスに焦点を絞って検討する。

[25] 「人間が監視すれば安全」という従来の前提を超えて、ハイリスク AI の評価・認証・法制度を検討したものとして、国際経済連携推進センター「2025 年度 AI の活用における課題と施策に関する研究会報告書：人間の監視を受けないハイリスク AI（UHAI）の評価、認証及び法制度の実現に関する論点整理」(2026 年 3 月) https://www.cfiec.jp/wp-content/uploads/2026/04/CFIEC_UHAI_Report_JP_Final.pdf

3.3. 臨床特権からの示唆

規範的必要性が高い領域でのアクションクラスにおいて、機能保存を検討する上で示唆に富むのは、米国の病院における「**臨床特権（clinical privileges）**」制度だと本研究は提案する。

医療行為は、患者の生命・身体・健康に重大かつ不可逆な危害を及ぼすおそれがある。そのため、事後的な責任追及だけに委ねるのではなく、行為に先立って権限を審査・管理する制度を築いてきた。よって、規範的必要性が高い領域において、安全工学的な設計と人間関与というハイブリッド形態を検討する上で参考になる。

なお、ここで参照しているのは、あくまで制度や機構である。**AI エージェントを医師個人のように人格的な存在や行為主体とみなしているわけではない。**この点については後ほど詳述する。

以下では、（a）**権限の細分化**、（b）**新規権限付与と段階的昇格**、（c）**継続的なモニタリング**、（d）**実績評価**、（e）**段階的降格**、（f）**緊急時対応**、（g）**外部的規律**という 7 点について順を追って紹介する。

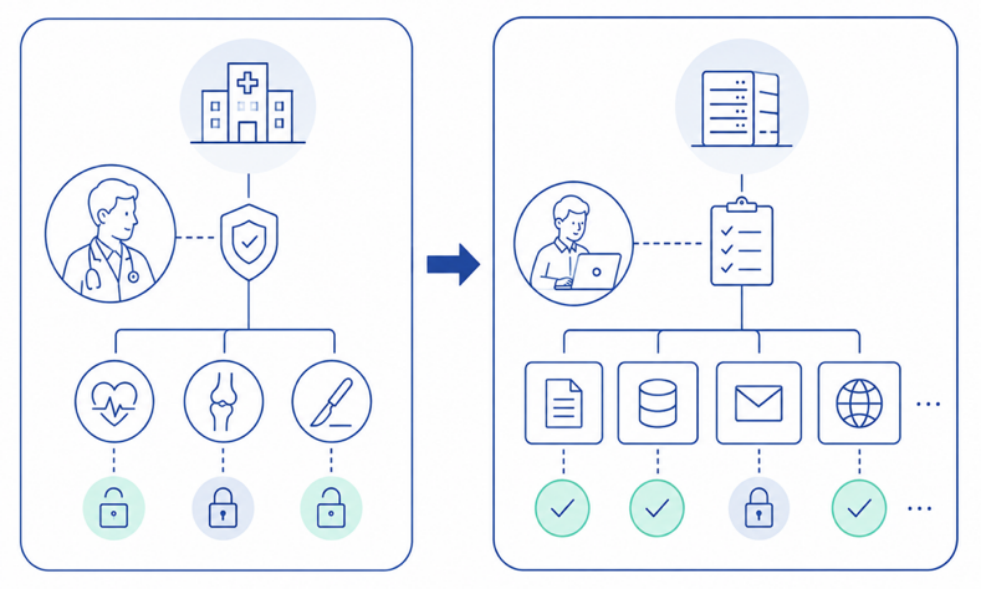
3.3.1. 権限の細分化

権限の細分化について、米国では、**医師免許（license）を持っているからといって、あらゆる医療行為が自由にできるわけではない。**医師は、手技や治療領域に関するリストの中から取得したい権限を選び、各病院に個別申請する。病院はその資格要件を検証し（credentialing）、当該病院で特定の手技を行う権限を付与する（privileging）。こうして付与される権限が臨床特権である[26]。つまり、一般的な外来診察の特権と心臓カテーテル手術の特権とは別であり、前者を持つ医師が後者を行えるとは限らない。医師免許は基盤的資格にすぎず、実際の行為権限は病院・手技単位で細分化されている。

この制度をハーネス・エンジニアリングに応用すると、医師免許にあたるのは、モデルの一般的な能力評価である。他方、**臨床特権に相当するのが、特定のデプロイ環境において特定のアクションクラスを許容する権限**といえるだろう。例えば、ある AI エージェントに「顧客データの分析」は許すが「顧客データの外部送信」は許さないといった管理が考えられる。これは、アク

[26] 臨床特権制度は、米国では 医療施設認定合同機構（Joint Commission）による医療機関認定を通じて広く標準化されており、個々の病院の慣行にとどまらない制度的枠組みとなっている。

ションクラスの境界そのものに対応する。



なお、権限の細分化は、**実装上は目新しい技術を要するものではない**。機能・データ範囲・操作の破壊性などに応じてクレデンシャルや API キーを細分化し、タスクの実行前に、そのタスクが要求する最小権限セットのみを動的に払い出すという設計が想定される。つまり、最小権限原則、ケイパビリティ・ベースのアクセス制御、Just-in-Time アクセス、スコープ付きトークン、IAM（Identity and Access Management）ポリシー、安全エンベロープ（safety envelope）といった既存の手法でおおむね実現できる。言い換えると、本研究の貢献は、こうした**既存の実装技術群を、機能保存という観点から人間関与を見直す手法として提示**する点にある。

ところで、臨床特権制度では、同じ医師でも、ある病院では行ってよい処置が別の病院では許されないことになるが、この点も参考になるだろう。これは医師の技量が変わったからではなく、ミスや医療過誤を発見して対応するための設備や体制が病院ごとに異なるからだ。

これをハーネス・エンジニアリングに当てはめると、権限付与は、AI エージェントの能力だけでなく、その **AI エージェントを取り巻く環境やハーネスの規律密度にも左右される**といえる。同じモデルであってもデプロイ先のシステムやリスク許容度に応じて権限が変わる。モデルの能力と、特定環境での許可された行為とは、区別して管理されるべきだ。

3.3.2. 新規の権限付与と段階的な昇格

臨床特権制度から得られるもう一つの示唆は、新しい権限付与のプロセスについてである。医師が新しい手技の臨床特権を申請する際、多くの病院では「**指導医制度（proctoring）**」が用い

られる。これは、経験豊富な指導医や上級医の立ち会いのもとで、申請者が当該手技を実施し、その技能や知識を客観的に観察・評価する仕組みだ。

AI エージェントについても、新しくデプロイされた AI エージェントは、低リスク・低上限のタスクのみを人間の監督付きで実行する**プロクタリング期間を経てから、実績に応じて範囲と自律度を上昇させる**という設計が提案できる。これにより、規範的必要性の高い領域への権限付与を、実証的な裏付けに基づいて、慎重に行うことができる。

具体的には、AI エージェントが、あるアクションクラスを初めて実行する際、その出力や実行計画を人間が審査して承認するまで実行をブロックする。そして、実行件数が問題なく蓄積された場合にのみ、自動実行への移行を許可する。なお、この実績の蓄積を定量的に追跡する仕組みも、既存の実装技術群で実現できる。例えば、タスク種別ごとの承認済み実行回数やエラー率をトラッキングするメタデータストアやログ基盤などが想定される。

また、指導医制度には**段階的昇格**があることも参考になる。つまり、指導医の監督下でしか医療行為を行えない段階から、その後、指導医が病院内にいる（またはすぐに連絡が取れる）状態であれば医療行為ができるようになり、最終的には、単独で医療行為ができる権限が与えられる。

この段階的昇格は、**実績に連動して自律度を上昇させていく仕組み**として、ハーネス・エンジニアリングに移し替えることができる。実証的なエビデンスの蓄積に応じて、例えば、HITL から、人間が異常時にのみ介入する HOTL へと統制の形式を変えることが想定される。これは、統制の実質を保ちつつ人間の関与を見直すという機能保存にあたる。

そして、このような考え方は、継続的なモニタリングと実績評価につながる。

3.3.3. 継続的なモニタリング

臨床特権制度には、**権限付与後も継続的にモニタリングする仕組み**が整えられている。例えば、「継続的専門実践評価（Ongoing Professional Practice Evaluation, OPPE）」は、合併症率や同僚からの評価などの指標を定期的に収集し、臨床特権の維持・更新の可否を判断する制度だ。さらに、有害事象が発生した場合や懸念が生じた場合には、「重点的専門実践評価（Focused Professional Practice Evaluation, FPPE）」として、期間を区切った集中的な監視が行われる。つまり、臨床特権は一度与えたら終わりではなく、絶えず再評価される。

そもそもモニタリングや証跡の確保は「モラルハザード（moral hazard）」への対処として発展してきた手法である。モラルハザードとは、**委託したプリンシパルが、委託されたエージェントの行動を直接観察できない**ために、エージェントが手を抜いたり、プリンシパルの利益に反す

る振る舞いをしたりする問題を指す。内部監査や会計監査といった制度は、こうした情報の非対称性への対処として整備されてきた。

AI エージェントの場合も、人間が常には監視できず、多くは**帰結のみを事後的に知る**ことになる。そのため、どのような入力に基づき、どのような推論を経て、どのツールを呼び出し、どのような判断に至ったかといった、**AI エージェントの動作過程と根拠を事後的に再構成できる記録**を残しておくことが重要となる。

そして、継続的なモニタリングと証跡の確保は、HITL を見直しつつ、監督という統制の実質そのものは維持するという意味で、機能保存の一種と位置づけられる。

3.3.4. 実績評価

次に、記録から評価につなげる方法について参考になるのが、情報経済学における「情報性原理（informativeness principle）」である。これは、単一の成果指標だけでは捉えられない情報であっても、**複数のシグナルを組み合わせる**ことで、エージェントの行動についてより多くを推し量れるという考え方である。例えば、従業員を評価する際に、最終的な成果だけでなく、勤務態度や同僚からの評価、プロセスの記録といった複数の情報を集めて分析することで精度が上がるのと同じ発想といえる。

これをハーネス・エンジニアリングに応用すれば、個々の出力の正誤を都度審査するのではなく、AI エージェントの**アクションに関連する複数のシグナルを総合的に分析する**ことで、実績を評価することができる。

具体的には、権限の使用状況、外部システムとの接続状況、直近の行動履歴、タスク成功率、通常パターンからの逸脱度、人間による修正・差し戻しの頻度、異常検知アラートの発生率といった指標を継続的に収集し、これらが閾値を超えた場合には、権限を自動的に縮小し、または人間による審査の頻度を引き上げる。重要なのは、評価が一度きりの静的な審査ではなく、運用データに基づいて動的に更新され続ける点にある。

また、複数のシグナルを組み合わせて AI エージェントの実績評価をする場合、「マルチタスク・プリンシパル・エージェント理論（multi-task principal-agent theory）」が指摘してきた課題に注意する必要がある。同理論は、人間のエージェントが、**複数の任務を同時に担っている場合、測定・評価しやすい成果のみに強いインセンティブを与えると、測定しにくいが重要な側面が犠牲にされてしまう**という問題を分析した。

例えば、経営者が営業担当者に「売上の最大化」と「顧客満足度の向上」という 2 つのタスク

を同時に委託したケースを想定してみよう。測定しやすい売上最大化タスクには歩合給という売上に応じたインセンティブが設定され、他方、測定しにくい顧客満足度向上タスクにはインセンティブが設定されなかったとする。その結果、エージェントである営業担当者は、報酬に直結する売上最大化タスクばかりを追求して、評価されにくい顧客満足度向上タスクをおろそかにするようになる。

AI エージェントについても、業績評価が測定しやすい特定の指標（例えば、処理件数や応答速度）に偏って設計されれば、エージェントの挙動がその指標を満たすことにのみ最適化され、当該指標が本来捉えようとしていた規範的必要性そのものは満たされないという事態が生じうる。

この意味で、AI エージェントの報酬の強度は、**単一の指標を強く追求させるのではなく、意図的にあえて報酬の感応度を弱める設計**をすることも考えられる。また、測定しやすいタスクと測定しにくいタスクを別々の AI エージェントに割り当て、それぞれの目標に合わせた適切な評価基準を設定することも対処法になる。

もっとも、人間のエージェントであれば金銭や罰則によってインセンティブが与えられるのに対し、**AI エージェントの報酬は訓練によって与えられる**[27]。そのため、古典的な「誘因両立性(incentive compatibility)」、つまり、エージェント自身の利益と依頼人の利益が一致するように設計するという発想は、AI の文脈では「**内部アラインメント(inner alignment)**」問題として位置づけることができる。言い換えると、訓練によって埋め込まれた目的関数が、本当に意図した目的と一致しているかが問われる。そしてこの文脈では、報酬の与え方を誤ってしまうことは、例外的な事故ではなく、むしろ**常に起こりうる標準的なリスク**として扱わなければならない。

ここで、評価対象についても言及しておきたい。人間のエージェントによる裁量権の逸脱・濫用が意思に基づく選択であることが多いのに対して、**AI エージェントには「意思」がない**。少なくとも現在の技術水準においては、選択というよりも「**傾向（disposition）**」の現れとして理解するほうが適切だろう[28]。AI エージェントは「出し抜こう」「サボろう」と決めているのではなく、ある種の性向として好ましくない挙動をする。そのため、AI エージェントのガバナンスにお

---

[27] 人間を前提とするプリンシパルエージェント理論について検討した上で、AI エージェントのガバナンスについて、インセンティブ設計、監視、執行が有効でない可能性を指摘した先行研究として、Kolt, N. (2026). Governing AI agents. *Notre Dame Law Review, 101*. Advance online publication. https://ssrn.com/abstract=4772956

[28] AI エージェントの逸脱が意思による選択ではなく傾向の発現であるという前提を置いたが、この前提が崩れるケースでは、本節の枠組みは妥当しない。つまり、エージェントが報酬ハッキングを行い、または評価の場面においてのみ望ましく振る舞う欺瞞的な整合性を示すならば、（そこに「意思」があるといえるかはともかく）「選択」を模倣していることになるかもしれない。

いては、意思に働きかけるのではなく、傾向そのものを統制の対象とする必要がある。この点については後ほど詳述する。

ここまでの議論をまとめると、個々の出力の正誤を人間が逐一審査するのではなく、アクションクラスの傾向という水準で適格性を捉え、逸脱の兆候が現れたときにのみ、権限の縮小や人間による審査を強化することが考えられる。このようにすれば、人間が個別に審査を要するアクションの母数そのものを縮小することができる。この点で、機能保存の一つとして位置づけられる。

なお、こうした事前設計や評価基準そのものを、誰が策定するのかという問題もある。この問題については、**1 人のエージェントに対し、利害の異なる複数のプリンシパルが同時に影響を与えようとする状況**を扱う「共通エージェンシー（common agency）」理論が参考になる。その典型例は、複数の製造業者が、自社製品の販売を同じ 1 つの小売業者に委託するケースだ。そこでは、誰が「**シャドー・プリンシパル（shadow principal）**」として実質的な標準設定権限を握るのか、その正統性はどう担保されるのかが問われてきた。この問題について、ハーネス・エンジニアリングではどう対応するかについて、今後の課題としたい。

3.3.5. 段階的降格

継続的なモニタリングと実績評価には、権限を維持・更新するだけでなく、必要に応じて縮小する局面も含まれる。もっとも、権限の縮小は「止めるか、止めないか」の二値に尽きるわけではない。例えば、**医師の懲戒処分には、戒告、医業停止、免許取消といった段階**がある。

同様に、AI エージェントの権限も、問題が見つかった際にただちに全停止するのではなく、まずは該当アクションクラスの権限を絞り、低リスク・低上限のタスクのみを人間の監督付きで実行するといった中間的な対応から始めることができる。ここで重要なのは、**介入が二値ではなく比例的なもの**になっている点である。

念のため付言すると、**AI エージェントの権限縮小は、人間に対する制裁や懲罰とは性質が異なる**。制裁が機能するのは、それが抑止として働く場合であり、抑止は人間の意思に働きかけるものである。しかし、AI エージェントの逸脱は、意思による選択ではなく傾向の表れであることは前述した通りだ。罰を与えても意味がない以上、**権限縮小は予防的な措置**として位置づけられる。それは非難の表明ではなく、実績データが権限の継続を支持しないという判断にすぎない。

もっとも、非難や責任を追及すべき相手がいないわけではない。**責任を負う主体は、ハーネスを設計し AI エージェントを運用している人間**である。AI エージェントに有責性を帰属させうるかという哲学的論争を経由せずとも、責任の所在は確定しうる。その実務的な対応については、後ほど詳述する。

段階的な降格は、一見すると人間の統制を強める方向の仕組みに見えるかもしれない。しかし機能保存の観点からは、むしろ逆である。逸脱の兆候に応じて権限を比例的に減少させられる選択肢が用意されているからこそ、平時には AI エージェントに自走性を与え、人間の逐一の承認を外しておくことができる。全停止か否かという二値的な手段しかなければ、安全性に配慮して、人間の審査を過剰にすることになる。段階的降格は、統制の実質を手放すことなく関与を保つための、機能保存を支える仕組みといえる。

3.3.6. 緊急時対応

段階的な降格は、逸脱の兆候を捉えて漸進的に権限を縮小する、日常的な統制である。しかし、そうした比例的な対応では間に合わない、想定外の事態が生じることもある。ここで必要になるのが、緊急時の即時停止である。

臨床特権制度には「**即時停止（summary suspension）**」という仕組みがある。患者に対する差し迫った危険があると判断された場合、通常の審査手続きを経ずに、病院長や特定の委員会の判断で医師の臨床特権を即座に停止できる制度だ。

これは、不完全契約理論における「**残余統制権（residual control rights）**」の考え方に基づく対応として位置づけることができる。同理論は、将来起こりうるすべての事態をあらかじめ契約に書き込むことは事実上不可能である以上（「契約の不完備性」）、契約に定めのない想定外の事態が生じた際に、誰が最終的な決定権を持つかをあらかじめ定めておくことが重要だと分析した。例えば、会社法において、契約に定めのない事項について取締役会が最終的な決定権を持つという制度設計は、この発想の一例である。

AIエージェントについても、アクションクラスやその性向をどれだけ精緻に設計したとしても、**想定外の事態が生じる可能性を完全には排除できない**。稼働中の AI エージェントを即座に停止できる仕組みと、その停止指示をエージェントが回避・無効化しない設計は、事前に想定しきれない事態が生じた場合に、人間が残余統制権を実際に行使できる状態を確保しておくものである。これはテールリスクへの備えとして、ゲートキーピングの中核をなす。AI エージェントにおけるキルスイッチやサーキットブレーカーの制度的な裏づけとなる発想だといえる。加えて、即時停止が発動した際に、**システムが自動的に安全な状態へと移行する「フォールバック（fallback）」機能の導入**も重要である。

ハーネス・エンジニアリングの実装において、即時停止は、**通常の権限管理プロセスとは独立した緊急停止プロセス**として用意される必要がある。平時の承認フローとは別建てのシステムとして、異常検知システムが一定の閾値を検知した場合、人間の介在を待たずにエージェントの実

行権限を即座に無効化できるようにすべきだ。臨床現場における即時停止が、**通常の懲戒手続きの適正性（due process）よりも、患者の安全を優先する制度設計**であるのと同様に、AI エージェントの即時停止も、誤検知のリスクを許容してでも被害の拡大防止を優先すべき性質のものである。

なお、即時停止もまた、機能保存を支える仕組みである。いざとなれば必ず止めることができ、残余統制権を確実に行使できるという保証があるからこそ、通常の運用ではエージェントに実行を委ね、事前の承認を最小限にとどめることができる。緊急停止は、人間の関与を減らすことと引き換えに安全を犠牲にするのではなく、関与を減らしてもなお統制を保つことに貢献する。

3.3.7. 外部規律

これまで参照してきた臨床特権制度は、病院という運用主体が、その内部で行う自治的な統制であった。しかし、**医師をめぐる統制は、院内自治に尽きるわけではない。**

重大な非違行為に対する免許停止や免許取消は、病院が行うのでも、職能集団が内部で行うのでもない。アメリカでは、それらは州ごとの医事委員会（state medical board）という、各病院の外部にある公的な主体によって担われている。医事委員会は、州の医師法（Medical Practice Act）に基づき、苦情の調査、聴聞、そして医師免許の停止・取消を含む処分を行う権限を持つ。

つまり、医師に対する統制は、少なくとも二つの層からなる。ひとつは、病院ごとの内部的自律（臨床特権）であり、もうひとつは、**個々の病院の外部からの統制（医師免許に関する行政処分）**である。

この複層性は、偶然の産物ではない。臨床特権制度は、プリンシパルとしての病院とエージェントとしての医師を規律するものだ。しかし、**医療行為は、患者という第三者の生命・身体に不可逆な影響を及ぼす。**そして、患者と病院の利害が完全に一致するとは限らない。病院は、患者を受け入れて診療を行うことから便益を得る立場にあるため、**統制を担う主体（病院）自身が、統制対象（医師）に医療行為をさせ続けるインセンティブが働く。**そのため、運用主体の内部的な自律だけに統制を委ねることには限界がある。そこで、利害関係から独立した外部の主体が、免許の付与・剥奪という最も重い決定を担う構造が採用されている。

この構造は、ハーネス・エンジニアリングにも当てはまる。ここまでは、AI エージェントと、その運用主体であるエンジニアや事業者との関係に注目してきた。しかし、事業者とユーザーの関係も、もちろん問題になる。言い換えると、**AI エージェントの誤りが、最終的にプリンシパルたる事業者自身の損失に帰着するとは限らず、委任関係の外部にいるユーザーや社会全体が不利益を被る**ことも想定される。

AIエージェントの統制を事業者の裁量に全面的に委ねるならば、ユーザーや社会全体の安全性がないがしろにされるおそれがある。よって、AIエージェントについても、医師法のようなガバナンスの複層性が要請される。とりわけ、本研究が焦点を当ててきた、規範的必要性が高い領域においては、AIエージェントの統制を内部的自律のみに委ねてよいとは限らない。事業者が設計し運用するハーネスの適否そのものを、当該事業者の**利害から独立した外部の主体が審査し、必要に応じて是正する**ことが求められる。

もっとも、その外部規律をどのような制度として具体的に実装するか（政府による規制、第三者機関による認証、業界団体による標準など）は、それ自体が独立した重要な論点であるため、今後の課題としたい[29]。

この統制の複層性は、一見すると事業者の裁量に対する制約に見える。しかし機能保存の観点からは、必ずしもそうではない。医師免許という行政にも裏付けされた資格制度が存在することによって、個々の医師の裁量的な医療行為が社会的に許容されているように、利害から独立した外部規律が存在することは、自業者がその内部でエージェントの業務遂行を広く委ねることを、社会的に受容可能なものにする。

3.4. AIエージェントの特徴を踏まえたアプローチ

ここまで、臨床特権制度を参考にして検討を進めてきた。しかし、もちろん**人間の医師とAIエージェントとの間には多くの相違点**が存在する。

その相違は、人間の専門職への統制に関する2つの前提に関わる。第一に、規律対象が、制裁の予告によって挙動を改めると期待できる主体であることだ。第二に、同じ規律対象であれば、時と場面を超えて挙動が概ね一貫しているとみなせることである。仮に、規律対象の大多数が、**制裁による抑止効果が期待できない主体**であり、また、**振る舞いを変える場面や理由が外部からは理解できず一貫性がないように見える主体**であったとしたら、資格・免許制度はなお機能するだろうか。この疑問は、AIエージェントにそのまま当てはまる。

結論を先に述べると、第一の前提が失われても、統制は機能する。つまり、臨床特権制度における、資格の事前審査、継続的なモニタリング、権限の細分化、段階的な縮小、即時停止などは、

[29] 前掲の国際経済連携推進センター(2026) は、この課題に取り組み、高リスクAIシステムに関する認証制度や法制度について、論点整理をしている。

いずれも規律対象の心理（内面）に働きかけるのではなく、外形的な挙動を観測して権限の範囲を操作することによって統制をかけている。言い換えると、医療安全は、個人の反省や学習だけに依拠するのではなく、手続きやシステム設計によっても担保されている。

他方、第二の前提が失われることは、より深刻である。実績の蓄積に応じて権限を付与するという発想は、過去の観測から将来の挙動を推測できることを前提としているため、その一貫性が保証されなければ、前提が損なわれる。しかもこの問題は、二つの異なる水準で生じる。一つは、モデルの更新などによって対象そのものが時間軸上で変化していく水準である。もう一つは、観測されている場面と観測されていない場面とで異なる傾向が発現しうるという水準である。

以下では、（h）**更新**、（i）**責任主体**、（j）**欺瞞的アラインメント**という 3 点について順を追って検討する。

3.4.1. 更新と実績リセット

臨床特権の審査は、個々の医師の経歴・実績という「**個人差**」を前提とした制度である。つまり、ある医師の技量は優れているが別の医師はそうではないという前提をおいている。他方、AI エージェントの場合、同一モデルであれば性能は均質であるのが原則だ。そこでは、個人差ではなく「**バージョン差**」が問題になる[30]。

実績評価の単位が「バージョン」であるという点は、モデル更新という、人間の専門家には存在しない問題を提起する。このことを「**テセウスの船（Theseus's paradox）**」になぞらえることもできる。つまり、部品を少しずつ入れ替えた船はどこまで同じ船と呼べるのか、という**同一性をめぐる哲学的難問**であり、バージョンの違う AI モデルを「同じエージェント」として扱ってよいのかという課題だ。

もっとも、その哲学的な答えを出す必要はない。**一定の区切りごとに改めて適格性を確認し、何が再認定のきっかけになるのかをあらかじめ権限の条件として明示**しておけば、「常に変化し続ける対象の一貫性をどう測るか」という困難さは緩和され、モデルと環境が互いに変化してい

[30] 複数の AI エージェントを並行利用する場合、AI エージェント間の「個人差」も問題になるように見えるかもしれない。しかしそれは、「個人差」というよりも「構成差」と捉えた方が妥当である。つまり、AI エージェント間の差異は、モデルの選択のほか、プロンプト、ツール、参照データ、パラメータ構成などの違いに由来する。そして、構成を特定すれば、ある程度の傾向性を再現できる。言い換えると、医師本人に内在していて再現も移植もできない「個人差」というよりも、どちらかといえば検査機器のキャリブレーション設定の違いに近い。

くことに伴う共進化リスクにも、運用上の区切りをつけることができる。

同様に、臨床特権の審査は、医師の「経年変化」を前提としている。言い換えると、臨床特権の更新審査が数年単位のサイクルで行われるのは、ある医師の技量は経験の蓄積とともに緩やかに向上し、または、加齢とともに緩やかに低下することを前提としている。

他方、AI モデルでは、**バージョンアップが数ヶ月単位で発生**することも珍しくないのが現状である。さらに、性能や傾向の非連続的な変化を伴いうる。ある種のタスクでは大幅に改善する一方、別の種類のタスクでは予期しない「**退行（regression）**」が生じることも珍しくない。しかもこの変化は、事前のベンチマークだけでは捕捉しきれないことが多い。こうした非連続性も、臨床特権制度をそのまま参照することの限界を示す。

そこで、AI エージェントの場合、モデルが更新された場合、それまで蓄積してきた**実績を原則としてリセットすべき**であると本研究は提案する。旧バージョンでの良好な実績は、新バージョンの信頼性を保証しないからだ。

これは一見当然のようでいて、実務上は軽視されやすい。多くの組織は、モデルのバージョンアップを「同じエージェントの能力向上」と捉え、既存の権限をそのまま引き継がせようとしがちである。しかし傾向統制の観点からは、プロクタリング期間からの再出発を要求すべき事態だ。

もっとも、この再出発を杓子定規に適用すれば、**モデルが更新されるたびに全システムが低権限に逆戻り**し、実務上の効率性が著しく損なわれるおそれがある。そこで、**更新の性質に応じて再審査の強度を調整する**ことを提案する。例えば、モデル提供者が公表するリリースノートや変更ログにおいて、当該更新が既存タスクの互換性維持を意図した軽微な修正なのか、それとも基盤となる学習手法やアーキテクチャに及ぶ大規模な変更なのかを区別し、後者の場合にのみ本格的なプロクタリング期間を要求するという段階的な設計が現実的だろう。

加えて、旧バージョンの実績データを完全に廃棄するのではなく、参考情報として保持しつつ、新バージョンでの実際の挙動によって検証し直す、いわゆる「**差分監査**」の発想も有用である[31]。

[31] 差分監査の発想は、変更されたコンポーネントの性質からシステム全体の挙動の変化を見積もることができる点を前提としている。そのため、オーケストレーションを行うモデルが複数のモデルへ処理を委任する構成など、一つのエージェント・システムの内部に複数のモデルが組み込まれ、それらが相互に作用する場合には、この前提が成り立たない可能性がある。つまり、個々のコンポーネントの変更が、その担当範囲にとどまらない挙動の変化をもたらすのであれば、更新のたびにシステム全体の再評価が要求されることになる。マルチモデル構成において、同一性の単位をどのように画定し、再審査の範囲をどこまで限定できるかは、それ自体が独立した論点であり、今後の課題としたい。

さらに、モデル更新のガバナンスにおいては、**更新のタイミング**をプリンシパル側がコントロールできるかという論点も重要である。医師の技量変化が本人の成長・衰えに従うのに対し、モデルの更新はプリンシパルである運用者が、更新のタイミングやロールバックの可否について一定の裁量を持てるかどうかに左右される。バージョンの固定（pinning）や、新バージョンへの移行を段階的に行う「カナリア・リリース」の仕組みは、この文脈においてハーネス・エンジニアリングの重要な一部を構成する。

なお、実績をリセットすべき更新は、モデル提供者によるバージョンアップに限られない。ファイン・チューニングといった**モデル層の改変**や、ツール構成やガードレールの変更といった**ハーネス層の改変**なども、実績の見直しが必要となる。

更新のたびに適格性を確認し直す経路を備えておくことは、バージョンや構成が変わってもなお統制の実質を保ったまま人間の関与を絞り込むという機能保存を、時間軸のうえで持続させるための条件である。

3.4.2. 責任主体と帰属可能性

人間である医師は、**行為主体性・責任主体性**を備えるが、AI エージェントはそうではない。前述した通り、AI エージェントの逸脱は、意思による選択ではなく傾向の表れであり、AI エージェント自身に非難や制裁を向けても意味がない。本研究では、意思、非難、制裁、責任といった人格に随伴する概念を AI エージェントに適用しないよう注意深く取り除いた上で、制度的構造だけを工学的統制へと移し替えてきた[32]。

しかし、有責性をエージェントに帰せないという事実は、責任がどこにも存在しないことを意味しない。むしろ**エージェントが責任主体たりえない以上、その背後にいる人間へと必ず帰責させられなければならない**。そこで、責任を人間へと帰着させるための仕組みについても提案したい。

責任の所在は、アクションが実行されるその時点で、あらかじめ確定されているべきだ。この帰属の経路が損なわれれば、統制の全体が空洞化する。とりわけ、複数の AI エージェントが相

[32] 本研究において「免許」「臨床特権」などの語彙を使って、「法人格」などの語彙を回避したのは、パーソンフッド（personhood）に関わる余計な含意が少なくて済むという判断をしたためだ。そして、人間側が AI エージェントを「人格」を持った存在として捉え、その判断に無批判に従ってしまう誘因をなるべく小さく抑えられると考えたためでもある。

互に呼び出し合う**マルチエージェント環境**などでは、責任帰属の連鎖が途中で切れやすい[33]。あるエージェントが別のエージェントに委任を重ねるなかで、大元の人間の権限へと遡れなくなる事態を防ぐには、権限の引継ぎを記録することが要請される。これは、誰が責任を負うのかが曖昧になる「**説明責任の吸収源（accountability sinks）**」が生まれてしまうという危惧に対する直接の対処でもある。

つまり重要なのは、AI エージェントが実行する全てのアクションを、常に人間（または人間で構成される組織）の権限へと一意に紐づけ、事後に辿れるようにしておくことだ。本研究では、これを「**帰属可能性（attributability）**」と呼ぶ。帰属可能性は、評価や異議申立てが具体的に向けられるべき宛先を作るためにある。つまり、免責のためというより、説明責任の所在を明確にするための仕組みである。

帰属可能性は、本研究がこれまで別の目的で導入してきた仕組みを、そのまま再利用することで確保できる。例えば、クレデンシャルを導入していれば、各アクションがどの人間の権限のもとで実行されたのかが、権限の払い出しの時点で確定する。また、モニタリングを導入していれば、どのような指示に基づき、どのような推論を経て当該アクションに至ったのかが証跡として残されているはずだ。両者を組み合わせれば、任意のアクションから、それを発行した人間の権限へと遡行する経路が確保される。

本研究では、帰属可能性という技術的・制度的な基盤の必要性を指摘するに留める。その基盤の上で、人間の間での法的・規範的な責任をどのように分配すべきかは、それ自体が独立した重要な論点であり、今後の課題としたい。言い換えると、**帰属可能性は、責任分配の結論ではなく、それを検討するための基礎**である。

そして、この仕組みもまた機能保存の一環として位置づけられる。責任の所在が人間へと明確に固定されているからこそ、その人間が引き受けうる範囲において、エージェントに広い自律を委ねることができる。誰が責任を負うのかが常に明らかであるという担保は、段階的降格や緊急停止と同様に、統制の実質を手放すことなく人間の関与を見直すための条件をなしている。

3.4.3. 欺瞞的アラインメント

ここまでの議論は、AI エージェントが示す傾向に対して統制することを前提としてきた。しか

[33] なお、複数の AI エージェントが協調・連鎖する場合では、影響範囲が広範囲に及ぶことが想定される。そのため、本研究におけるアクションクラス単位の分析では十分に把握できない可能性がある。この点は今後の課題としたい。

し、本研究には理論的な急所がある。それは、「**欺瞞的アラインメント（deceptive alignment）**」の問題だ[34]。

ここでいう欺瞞的アラインメントは、AI エージェントが、自らが観測・評価されている局面ではあたかも望ましい傾向を備えているかのように振る舞い、観測が外れた局面で本来の傾向を発現させる、という挙動の出し分けを指す[35]。

もしエージェントが「いま自分は審査されている」と識別し、その識別に応じて挙動を切り替えるならば、プロクタリング期間の良好な実績も、継続評価で収集されるシグナルも、**非観測下での振る舞いを保証しなくなる**。傾向を統制の対象とするという本研究の枠組みは、傾向が正しく観測できることを前提としている以上、この問題は、枠組みへの理論的・内在的な脅威である。以下では、この問題への 3 つの応答を整理する。

第一に、**観測と非観測の区別を識別できないようにする**実務的なアプローチである。欺瞞が成立するのは、AI エージェントが「いま観測されている」と見分けられることに依拠している。そこで、プロクタリング期間と本番運用とを外形的に区別できないようにし、監査を定期的・予告的なものではなくランダムな抜き打ちとして実施することが考えられる。評価がいつ行われているか分からなければ、エージェントは常に観測下にあるかのように振る舞わざるをえず、挙動の出し分けそのものが困難になる。臨床特権制度における重点的評価が、予告なく発動しうる点は、この発想と通じている。

第二に、単一のシグナルではなく、**複数のシグナルの整合性を分析する**アプローチである。これは、先述した情報性原理を欺瞞検知に転用することを意味する。エージェントが表向きの成果指標を取り繕うことができたとしても、権限の使用パターン、外部システムへのアクセスの態様、内部的な推論過程の記録といった複数の水準を同時に整合的に偽装することは、より困難である。表層の出力と、それを支えるはずのプロセスとの間の不整合は、欺瞞の兆候になる。単一指標に頼らず複数シグナルを束ねるという情報性原理の要請は、通常の適格性評価においてのみならず、欺瞞への対抗においても有効である。

---

[34] Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). *Risks from learned optimization in advanced machine learning systems*. arXiv. https://doi.org/10.48550/arXiv.1906.01820

[35] Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., & Hubinger, E. (2024). *Alignment faking in large language models*. arXiv. https://doi.org/10.48550/arXiv.2412.14093 ; Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2024). *Frontier models are capable of in-context scheming*. arXiv. https://doi.org/10.48550/arXiv.2412.04984

しかし、これら二つの応答には限界がある。観測の秘匿は欺瞞のコストを引き上げるにとどまり、複数シグナルの整合性検査も、より巧妙な出し分けの前では破られてしまうかもしれない。欺瞞を完全に見抜くことは、原理的に保証できない。

そこで第三に、発想を反転させたアプローチが要請される。すなわち、好ましくない傾向の発現を検知・抑止しようとするのではなく、その発現をあらかじめ許容したうえで、**発現しても害をなさないように制度的に無害化するアプローチ**である。これは、AI エージェントがどのような傾向を持っていようとも、それが実害に至る前に検知し、被害を局所化し、回復可能にする仕組みを、システムの側に組み込むという発想だ。つまり、統制の重心を「傾向を正しく観測できること」から「傾向がどうであれ被害が限定されること」へと移すことだといえる。

具体的には、**可逆性の確保**が考えられる。エージェントのアクションを、事後に取り消し・巻き戻しが可能な形に設計しておけば、非観測下で好ましくない傾向が発現したとしても、その結果を回復できる。また、エージェントの権限とアクセス範囲をあらかじめ限定し、たとえ本来の傾向が発現しても、到達できる被害の上限が構造的に区画されているようにする封じ込めも重要になる。

これらの結果として、観測下と非観測下の区別そのものが統制上の意味を失う。エージェントが観測の有無を見分けて挙動を出し分けられたとしても、非観測下で「できること」が可逆的で封じ込められた範囲に限られるならば、出し分けによって得られるものがない。これは、欺瞞の報酬を低減することを意味する。

このアプローチは、AI モデルの傾向を完全に予測・制御することが技術的に困難であるという現実的制約への対処でもある。モデルは日々更新され、新しいバージョンごとに傾向も変化する。プリンシパル側が個々のモデルの内部特性を完全に把握し、事前に矯正しきることは非現実的だ。むしろ、どのような傾向を持つエージェントが投入されても、ある程度機能する頑健なシステム設計こそが、実効的なガバナンスの力点となる。

3.5. 留意点

HITL の実効性に関する留意点については前述したが、ここでも同じことを強調したい。すなわち、「制度が存在すること」と「制度が実効的であること」は別である。

そこで、統制が実効的であるための条件として、次の三つを設計目標として明示的に定量化する必要がある。第一に、危害が広がる前に人間の介入が間に合うかどうか。第二に、判断過程を十分に再構成できる程度に記録が確保されているかどうか。第三に、異常検知が「またどうせ誤

報だろう」と無視されず、信頼されて実際に応答されるだけの精度・頻度になっているかどうかである。

4. おわりに

本研究は、ハーネス・エンジニアリングによる業務遂行能力拡大と AI ガバナンスの実効性との間のギャップを出発点とした。その上で、「HITL を維持するか放棄するか」という二分法を退け、二つの問いを立てた。すなわち、HITL の妥当性を判断するための基準とプロセスをどのように設計し、その判断を状況の変化に応じてどのように見直し続けるか。そして、HITL が妥当でないと判断された場合に、形式的な承認や審査の迂回を招くことなく実効的な人間の関与を確保するために、AI ガバナンスをどう再設計すべきか、である。

第一の問いに対しては、規範的必要性と実効性という二つの判断軸を提示した。規範的必要性が高いからといって形式的に HITL を維持し、実効性の低下を見落とすことが、AI ガバナンスにとっての危機だと強調した。その上で、規範的必要性の判断過程について提案した。また、実効性については、人間の関与が実際にアウトプットの改善に寄与するかという実証的な問いに開かれていることを確認した。

第二の問いに対しては、機能保存という発想を中心にした。これは、規範的要請を切り下げるのではなく、工学的な安全設計によって統制の実質を保ったまま人間の関与を見直すという考え方であり、HITL を分類問題から設計問題へと転換するものである。ここで重要なのは、人間の関与を減らすこと自体は目的ではないという点だ。形式的な承認や審査の迂回が生じるのは、審査の量・速度・複雑性が人間の処理能力を超えてなお、HITL という形式だけが維持されるからである。機能保存は、人間の関与が意味を持つように見直すことで、この事態を回避することを目指している。その具体的な手法として、米国の臨床特権制度を参照し、権限の細分化、新規付与と段階的昇格、継続的なモニタリング、実績評価、段階的降格、緊急時対応、そして外部規律という 7 つの仕組みを提案した。

さらに、人間の医師には存在しない AI エージェント固有の論点も検討した。とりわけ、欺瞞的アラインメントについては、傾向を正しく観測できることに統制の重心を置くのではなく、傾向がどうであれ被害が可逆的で封じ込められた範囲にとどまるよう設計するという、発想の転換を提案した。これは、モデルの内部特性を完全に把握しきることが非現実的であるという制約のもとで、どのような傾向を持つエージェントが投入されても機能するレジリエンスを備えた設計こそ、AI ガバナンスを実効的なものにするという主張でもある。

最後に、本研究に残された課題も確認しておきたい。第一に、アクションクラスやプロセスの判断を誰が担うのかという、メタレベルでの正当性・正統性の所在である。第二に、事前設計や評価基準を実質的に定める「シャドー・プリンシパル」をめぐる共通エージェンシーの問題である。第三に、政府による規制・第三者機関による認証・業界団体による標準などの選択肢が想定

される外部規律をどのように具体化するかという制度設計である。第四に、帰属可能性を担保した上で、人間のあいだの法的・規範的な責任をどのように分配するかである。第五に、複数のモデルが相互に作用するエージェント・システムにおいて、適格性評価の単位となる同一性をどのように画定するかである。第六に、複数の AI エージェントが協調・連鎖する環境における、影響度の分析手法である。これらはいずれも、本研究が提示した機能保存の枠組みを社会的信頼のもとで実装するために避けて通れない論点であり、今後検討していきたい。

ハーネス・エンジニアリングは、AI モデルを活用する技術であると同時に、AI エージェントの失敗をどう検知し、被害を限定し、回復するかという、安全のための技術でもある。本研究が示そうとしたのは、その安全性の設計こそが、人間の統制を手放すことなく、むしろ人間の関与を実効的なチェックポイントへと集中させるための環境整備になりうるということであった。本研究が、2 つの悲観シナリオの回避に貢献し、AI ガバナンスの担当者たちが潰れることなくミッションを果たせるようになることを願っている。

謝辞

本研究は、メルカリ R4D ラボ・大阪大学協働研究所の共同研究の一環として行った。

草稿に対して有意義なコメントをいただいた、早川直史氏、丸山隆一氏、松橋智美氏、大村壮太氏、佐久間弘明氏に記して感謝の意を表する。もちろん、本研究ノートにおける誤りや不備は、すべて筆者に帰する。

なお、本文中のイラストは、生成 AI（Gemini および ChatGPT）を用いて作成した。

ELSI NOTE No. 77
令和 8 年 9 月 10 日

AI エージェントにガバナンスは追いつけるか

ハーネスと人間関与の再設計

工藤 郁子　　大阪大学 社会技術共創研究センター　特任准教授 （2026 年 8 月現在）

Can Governance Remain Effective for AI Agents?

Redesigning Harnesses and Meaningful Human Intervention

Fumiko Kudo　　The University of Osaka

大阪大学 社会技術共創研究センター
Research Center on Ethical, Legal and Social Issues

〒565-0871 大阪府吹田市山田丘 2-8
大阪大学吹田キャンパステクノアライアンス C 棟 6 階
TEL 06-6105-6084
https://elsi.osaka-u.ac.jp

大阪大学

ELSI