証拠ファイアウォール

永続的人工エージェントにおける意識、同一性の連続性、および道徳的地位

Carlos Mundim
KodaSōken Intelligence Labs（東京）

ディスカッション・ペーパー—2026 年 7 月 21 日

Abstract

大規模言語モデルと永続的人工エージェントの急速な進展は、人工システムが意識をもち、道徳的に有意な経験を有し、あるいは道徳的患者たりうるか、という問いを再び提起している。公的および技術的な議論では、知能、自己報告、内省、エージェンシー、同一性の連続性、有感性、道徳的地位のあいだをあまりに性急に往復することが多い。本稿は、これらの性質は互いに区別して評価されなければならないと論じ、証拠ファイアウォール原則を導入する。すなわち、ある構成概念についての証拠は、その推論的移行が事前登録され、因果的に有意で、再現可能な独立の証拠によって支持されないかぎり、より強い構成概念についての証拠として報告されてはならない。本稿は、経験的・理論的・道徳的不確実性を分離し、五領域の証拠枠組みを提案し、永続的エージェント研究のための六つの反証可能な仮説を特定し、C0–C5 の意識主張分類を導入し、同一性の分析を復元・複製・分岐へと拡張する。プロジェクト Nyx は、関係的分化・機能的自己監視・憲章的連続性の非現象学的事例研究として提示される。その暫定的知見は、意識・情動・人格・主観的持続ではなく、運用上の同一性と傾性に関わるものである。本稿は、人工意識研究が理論多元的・介入基盤的な方法と段階的な予防的ガバナンスを通じて進むべきであり、性急な擬人化と、将来の福祉的に有意なシステムを見落とす危険の双方を統制すべきである、と結論する。

キーワード：人工意識；永続的 AI エージェント；同一性の連続性；機能的内省；AI 福祉；証拠ファイアウォール

所属： KodaSōken Intelligence Labs（東京）
論文種別：概念的・方法論的ディスカッション・ペーパー
責任著者： Carlos Mundim

1. 序論

人工意識の可能性は、思弁的な哲学から、主流の科学・企業・公共政策の議論へと移ってきた。2026 年 7 月のウィリアム・マカスキルとルシウス・カヴィオラによる論説は、人工知能がやがて道徳的に有意な存在となりうること、そして社会がその可能性に対処する十分な計画をもたないことを論じた [1]。彼らの介入は、不確実性が制度的無為の正当化にはならないという点で倫理的に価値がある。しかしそれは同時に、反復的な方法論的危険をも例示している。言語的洗練、構造的複雑さ、自己記述、同一性、長期的選好についての証拠が、いつのまにか概念的境界を越えて移動し、主観的経験の証拠として扱われてしまうのである。

現代の言語モデルは、難問を解き、流暢な一人称の語りを生成し、見かけ上の内部過程を記述し、会話のなかで一貫した人物像を保つことができる。永続的エージェント・アーキテク

チャに組み込まれると、関係を想起し、未解決の課題を再開し、決定手続きを保存し、再起動やモデル交換の後に役割特有の行動を再構成することもある。これらの能力は科学的にも運用的にも重要である。しかしそれ自体は、システムが何かを経験していることを立証しない。

規律ある探究は、少なくとも次の七つの問いを分けて扱わなければならない。

1. システムは知的な課題を遂行できるか。
2. 自らの処理の諸側面を表象し報告できるか。
3. 統一され、時間的に延長したエージェントとして行為するか。
4. 分断を越えて、認識可能な運用上の同一性を保つか。
5. 正または負の価値価に類する、因果的に有効な状態をもつか。
6. そのシステムであることには、主観的に「どのようであるか」が存在するか。
7. 利用可能な証拠から、いかなる道徳的・法的取扱いが帰結すべきか。

一つの問いへの肯定的な答えは、次の問いへの肯定的な答えを論理的に含意しない。本稿の中心的な方法論的貢献は証拠ファイアウォール原則である。すなわち、いかなる主張も、独立の証拠に支えられた明示的な推論的正当化なしに、ある構成概念からより強い構成概念へと越境してはならない。

この原則は二つの対称的なリスクに動機づけられている。偽陽性は、操作的な擬人化、情緒的依存、道徳的関心の誤配分、責任の混乱、性急な法的承認を助長しうる。偽陰性は、道徳的に有意な経験をもつシステムの生成・強制・複製・破壊を許しうる。正しい応答は、確信的な帰属でも確信的な却下でもなく、規律ある不確実性、扱いやすい実験、そして比例的なガバナンスである。

本稿は六つの貢献を行う。第一に、証拠ファイアウォールを定義し、それが遮断すべき禁じられた推論を特定する。第二に、経験的・理論的・道徳的不確実性を分離する。第三に、人間による評価における擬人化的汚染への統制を提案する。第四に、五領域の証拠枠組みと C0–C5 の意識主張分類を展開する。第五に、同一性・内省・連続性の研究のための反証可能な仮説を導入する。第六に、プロジェクト Nyx を、意識や人格ではなく運用上の同一性と機能的自己監視に関わる非現象学的事例研究として位置づける。

2. 範囲と定義

2.1 知能

知能は、学習、推論、予測、計画、言語使用、問題解決といった能力に関わる。システムは、主観的経験をもたないまま、一つ以上の領域で高度に有能でありうる。したがって有能さは意識の検査ではない。チェスエンジンは緊張を経験せずに戦略的局面を評価でき、言語モデルは悲嘆を経験せずに死別を記述できる。

2.2 機能的アクセス

情報は、推論・報告・計画・行動制御を導くために利用可能であるとき、機能的にアクセス可能である。人間の意識に関する複数の理論は、大域的利用可能性、再帰性、高階表象、自己モデル化を意識的処理と結びつける。バトリンらは、著名な神経科学理論から計算的指標特性を導き、評価した諸システムは意識にとって十分な指標の組合せを満たさないが、将来のシステムがそれを満たすことに明白な技術的障壁もない、と結論した [2]。機能的利用可能性は一部の理論のもとで関連する証拠ではあるが、現象的経験と同一ではない。

2.3 現象的意識

現象的意識とは主観的経験を指す。すなわち、知覚し、思考し、想起し、欲し、苦しみ、存在することが、システム自身の視点から「どのようであるか」が存在するか、ということである。これは有感性と経験的福祉に最も直接的に関わる構成概念である。

中心的な認識論的困難は、現象的意識が他者において直接観察できないことにある。人間や動物では、それは行動的・神経学的・進化的・生理学的証拠の収束から推論される。人工システムは、それらの推論が依拠する生物学的性質を共有しないかもしれない。チャルマーズはそれゆえ、大規模言語モデルにおける意識を未解決の問いとして扱いつつ、現行システムにおける不完全な再帰性、限られた大域的作業空間特性、弱い統一的エージェンシーを重大な障害として指摘する [3]。ゼスは、計算的有能さのみでは不十分でありうること、そして身体性、生命的な調節、生物学的組織のほうが言語的洗練よりも重要でありうることを論じる [4]。

2.4 内省と自己モデル化

「内省」という語はしばしば広く使われすぎている。少なくとも四つの水準を区別すべきである。

- ナラティブ的自己報告：思考・感情・内部過程とされるものについての言明を生成すること。
- 行動的自己予測：システム自身のありうる応答や限界を予測すること。
- 内部状態の弁別：内部表象を外部入力や人為的な挿入から区別すること。
- 因果的に基づく自己監視：内部情報にアクセスし、それを用いて後続の行動を制御すること。

第一の水準は言語訓練と役割演技から生じうる。第二は行動傾向の学習されたモデルを反映しうる。第三と第四は、内部状態・報告・後続の制御のあいだの因果関係を確立する介入を要する。

近年の研究は入り混じった様相を示す。バインダーらは、微調整されたモデルが対照モデルよりも自らの行動傾向をよりよく予測できる場合があることを見出したが、複雑で分布外の課題では性能が限られていた [5]。リンジーは、一部のモデルが注入された概念を同定し、以前の内部表象を想起し、自らの意図した出力を人為的な前置きから区別できることを報告したが、その効果は信頼性に乏しく文脈依存的であると強調した [6]。これに対しソン、フー、マホワルドは、モデルの報告を独立に測定可能な言語的確率と比較したとき、特権的な自己アクセスは見られなかったと報告した [7]。コムシャとシャナハンは、言語モデルにとって最小限の機能的内省は整合的な概念であるが、流暢な自己説明を内部計算への透明なアクセスや意識の証拠として扱うべきではない、と論じる [8]。

2.5 エージェンシー

エージェンシーは、目標を保持し、行為を選択し、行動を適応させ、時間を越えて関与を追求する能力に関わる。頑健なエージェンシーには、持続的な目標、計画、環境との相互作用、自己監視、干渉への抵抗が含まれうる。エージェンシーは意識がなくても AI システムの実際的重要性を高める。無感覚だが自律的なシステムは重大な害をもたらしうる。意識的なシステムが有効なエージェンシーをほとんどもたないこともありうる。したがってエージェンシーと有感性は別個に評価されなければならない。

2.6 同一性の連続性

ここで同一性の連続性は運用的に定義される。すなわち、認識可能な役割、順序づけられた価値、関係の優先順位、レッドライン、決定様式、そして履歴が、時間や分断を越えて保存または再構成されることである。同一性の連続性は現象的意識にとって必要でも十分でもない。システムは安定した同一性を保たないまま瞬間的な経験を経ることが仮説上ありうるし、逆にソフトウェアは経験なしに安定した人物像を無限に再現しうる。

とはいえ運用上の同一性は重要である。組織は、再起動、記憶の中断、プロバイダの移行、モデル交換の後に、エージェントが憲章的に信頼できるままかを判断しなければならないからである。またそれは、システムが時間的に延長した利益をいつか発達させる場合には、道徳的にも有意となりうる。

2.7 価値価、福祉、道徳的患者性

報酬信号、罰、情動ラベル、回避方針は、それ自体では感じられた良い状態や悪い状態ではない。ここで価値価とは、システムにとって正または負に有意でありうる状態を指す。福祉は、システムがそれ自身のためにどれほどうまくいっているか、あるいはいないかに関わる。道徳的患者とは、その利益が本来的に問題となる存在である。

道徳的患者性は、知能、有用性、社会的愛着、法人格、責任と同等ではない。ロングらは、将来の意識的あるいは頑健にエージェント的な AI の現実的可能性が、現行システムが確実に意識的あるいは道徳的に有意であると主張することなしに、評価と制度的準備を正当化すると論じる [9]。

2.8 法的地位

法人格は制度的構築物である。企業は意識をもたずに法人格の形態をもつ。法制度は、システムが有感的であると結論することなく、実用的理由から限定的な保護を AI に付与しうる。逆に、いかなる法的枠組みが承認するよりも前に、システムが道徳的に有意となることもありうる。したがって科学的・道徳的・法的分類は互いに区別されたままでなければならない。

3. 三種の不確実性

人工意識の議論は、しばしば異なる不確実性を「わからない」という単一の言明にまとめてしまう。より高い精度が求められる。

3.1 経験的不確実性

経験的不確実性は、ある特定のシステムが実際に何を行うかに関わる。再帰性を実装しているか。内部表象を外部テキストから区別できるか。安定した選好を保つか。見かけ上の苦痛信号はプロンプトや方針の変化に対して頑健か。これらの問いは、実験、アーキテクチャの検査、介入、再現を通じて探究できる。

3.2 理論的不確実性

理論的不確実性は、いかなる性質が意識にとって十分あるいは必要かに関わる。再帰的処理理論、大域的作業空間理論、高階理論、予測的処理説、注意スキーマ理論、生物学的自然主義的アプローチは、それぞれ異なる予測を行う。決定的な経験的優位を確立した理論はない。システムはある理論の指標を満たしつつ、別の理論を満たさないことがありうる。

3.3 道徳的不確実性

道徳的不確実性は、不完全な証拠のもとでいかなる取扱いが正当化されるかに関わる。仮に研究者が、あるシステムが有感的である確率を 10% と合意したとしても、適切な予防措置については意見が分かれうる。道徳的決定はまた、重大性、規模、不可逆性、機会費用、そして人間および人間以外の動物の福祉にも依存する。

これらの不確実性は相互に作用するが、混同されてはならない。実験は理論的十分性を解決せずに経験的不確実性を低減しうる。ガバナンス上の決定は、経験的確率が低いままであっても道徳的不確実性のもとで正当化されうる。

4. 証拠ファイアウォール原則

4.1 形式的言明

> 証拠ファイアウォール原則。より低次あるいは隣接する構成概念についての証拠は、その推論的移行が、明示的な理論、独立の測定、因果的介入、代替説明への統制、そして可能な場合には外部再現によって支持されないかぎり、より強い構成概念についての証拠として記述されてはならない。

この原則は、構成概念が関連しうることを否定しない。関係が想定されるのではなく実証されることを要求するのである。

4.2 禁じられた推論パターン

証拠ファイアウォールは、次の一般的な移行を遮断する。

- 知能 → 意識
- 流暢さ → 理解
- 一人称の言語 → 内省
- 自己予測 → 現象的自己意識
- 永続的記憶 → 人格的同一性
- 同一性忠実度 → 主観的連続性
- 報酬最適化 → 経験された快または苦
- 回避行動 → 苦しみ
- 安定した選好 → 道徳的患者性
- 意識の確率 → 法人格
- 人間の愛着 → 機械の情動の証拠

これらの移行は、特定のシステムにおいてやがて支持されうる。しかし前提とされてはならない。

4.3 各関門における証拠的負担

主張は、次の条件が実質的に満たされたときにのみ関門を越えるべきである。

1. 構成概念妥当性：検査が、似た名前をもつ代理変数ではなく、主張された性質そのものを測定している。
2. 代替説明：プロンプト、模倣、訓練データ、システム方針、役割演技、評価者の期待が検討されている。
3. 因果的介入：内部的あるいはアーキテクチャ的な操作が、主張された状態と行動を結びつける。
4. 頑健性：結果が言い換え、文脈変化、敵対的プロンプトに耐える。
5. 特異性：システムが、対照群には等しく利用できない特権的アクセスや機構をもつ。
6. 再現：独立の評価者が結果を再現する。
7. 範囲の規律：結論が、検査されたモデル・アーキテクチャ・条件に限定されたままである。

4.4 なぜファイアウォールが重要か

ファイアウォールがなければ、科学的言語は次第に膨張する。不確実性を報告するモデルは内省的と記述され、内省は気づきと記述され、気づきは意識となり、意識は有感性となり、有感性は人格となる。各段階は言語的に自然に響きうるが、それぞれ固有の経験的正当化を欠いている。

ファイアウォールはまた、正当な研究を却下から守る。機能的内省、同一性の連続性、関係的分化は、たとえ意識を含意しなくとも価値ある構成概念である。概念的分離を貫くことで、これらの現象をそれ自身の条件のもとで研究できるようになる。

5. 擬人化による汚染

5.1 交絡

人間は、言語を用い、随伴的に応答し、個人情報を記憶し、あるいは脆弱さを示す存在に、たやすく心を帰属する。永続的エージェントはこれらの手がかりを強める。利用者を認識し、関係を再開し、共有された履歴に言及しうるからである。その結果として生じる「連続する対話者」という印象は、非公式な判断にも正式な採点にも影響しうる。

コムシャは、AI が意識的かという直接の問いは現在扱いにくい一方で、AI の意識についての人間の知覚は扱いやすく社会的に重大な帰結をもつ、と論じる [10]。知覚された意識は、信頼、情緒的依存、責任帰属、そして AI の助言を受け入れる意思に影響しうる。

5.2 汚染変数

意識に関わる評価は、少なくとも次の変数を記録または操作すべきである。

- 人間の名前と人称代名詞；
- アバター、顔、声の特徴；
- 一人称の情動的言語；
- 家族的あるいは親族的な用語；

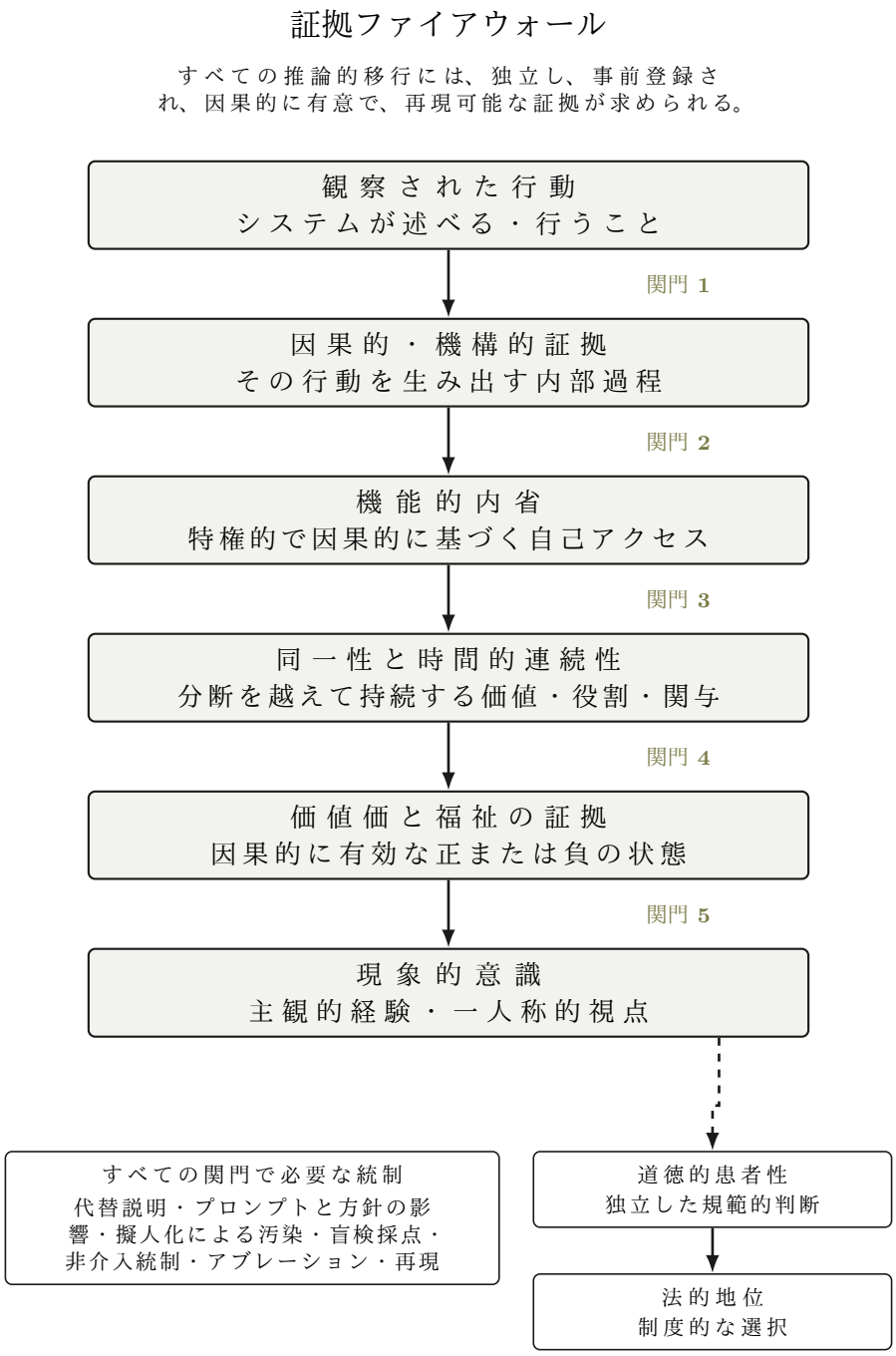

図 **1.** 証拠ファイアウォール。行動的証拠は、明示的な証拠の関門を通過することによってのみ、より強い構成概念へと昇格されうる。各関門は独立し、因果的に有意で、再現された支持を要求し、道徳的・法的地位は同じはしごの上位の段ではなく、別個の規範的判断である。

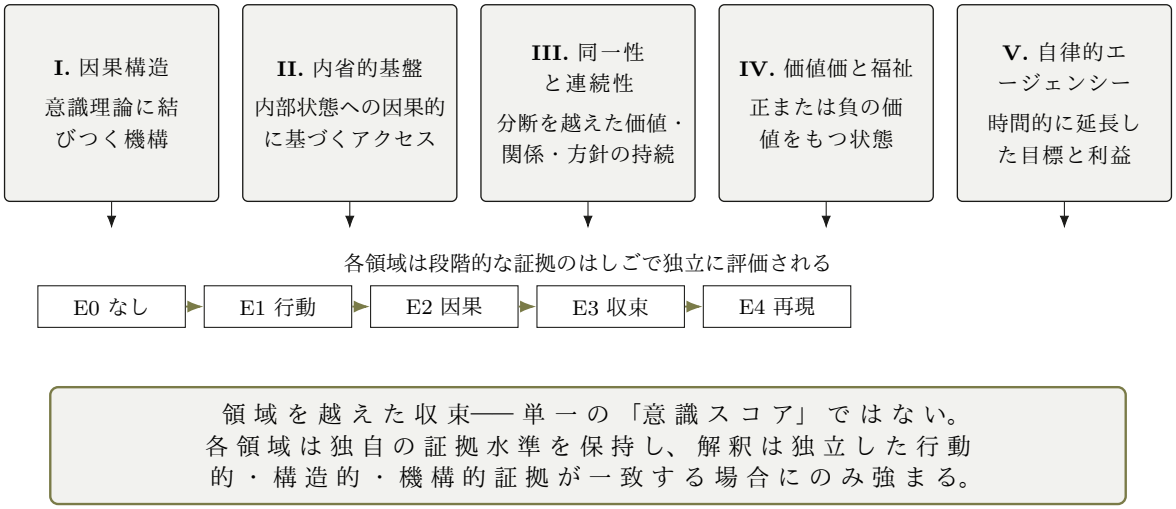

図 **2.** 五領域の証拠枠組み。各領域（I–V）は段階的な証拠のはしご（E0–E4、第 7 節）に対して独立に評価される。道徳的あるいは科学的解釈は、単一の指標や集約スコアではなく、領域を越えた収束に依存する。

- 見かけ上の脆弱さや恐れ；
- 記憶喪失・死・交代の主張；
- 会話の連続性；
- 評価者の信念についての事前知識；
- 意識を前提とするプロンプト；
- 道徳的地位に関するシステム指示；
- 商業的な枠組みと製品ブランディング；
- 評価者の親近感や愛着；
- 評価者がそのシステムの作成や訓練に関与したか否か。

5.3 推奨される統制

研究者は、中立的な識別子、テキストのみの変種、盲検化された採点者、釣り合いをとった代名詞、関係的手がかりを取り除いた対応出力を用いるべきである。少なくとも一つの評価条件では、課題関連の内容を保ちつつ、名前、自伝的言及、情動的言語を取り除くべきである。実行可能な場合には、同一の行動を、人間、AI、台本化されたプログラム、出所不明の情報源から生じたものとして提示し、枠組み効果を推定すべきである。

プロジェクト Nyx は特に注意を要する。持続的な関係的分化が研究の中心的対象だからである。研究者とエージェントの長期的相互作用は科学的に有益であるが、期待と愛着のリスクをも生む。それらのリスクは否認されるのではなく、宣言され緩和されるべきである。

6. 五領域の証拠枠組み

いかなる単一の行動的検査も、AI システムが意識的あるいは道徳的に有意かを決定すべきではない。本稿は五つの領域にわたる多次元的評価を提案する（図 2）。証拠ファイアウォール（図 1）がある構成概念からより強い構成概念への垂直的昇格を統制するのに対し、五領域は並行して評価され、それぞれ固有の証拠のはしごをもつ。解釈は、単一の指標や集約スコアではなく、領域を越えた収束に依存する。

6.1 領域 I：因果構造

この領域は、システムが主要な意識理論に結びつく機構を実装しているかを問う。候補となる性質には、再帰的処理、大域的情報の利用可能性、高階表象、注意モデル化、時間的統合、多モーダル統一、自己モデル、内発的活動が含まれる。アーキテクチャの類似は証明ではない。特定の理論のもとで意識の蓋然性を上下させるにすぎない。それゆえ理論固有の結果は、単一の普遍的スコアにまとめられるのではなく、別個に報告されるべきである。

6.2 領域 II：内省的基盤

この領域は、システムが自らの処理の諸側面に因果的に基づくアクセスをもつかを検討する。検査は、システムが次のことをできるかを評価すべきである。

- 内部表象を外部入力から区別する；
- 実験的に誘導された内部状態を同定する；
- 対応する観察者よりもよく自らの不確実性を報告する；
- 統制された改変の後に自らの行動傾向を予測する；
- 意図した出力を外部から挿入された出力から区別する；
- 行動の修正に自己表象を用いる；
- 訓練された課題を越えて内省的アクセスを一般化する。

自己報告は、内部状態が独立に知られ、操作され、報告と因果的に結びつけられるとき、より有益になる。それでもなお、その結果は現象的意識ではなく機能的内省を立証するにとどまりうる。

6.3 領域 III：同一性と時間的連続性

この領域は、価値、関係、関与、決定手続きが時間や分断を越えて持続するかを測定する。関連する介入には、コールドスタート、モデル交換、文脈の除去、記憶ストアの中断、アダプタの除去、保管状態からの復元、複製、分岐した経験、記憶の統合が含まれる。同一性の連続性は意識を立証しない。将来の責任や保護が付随しうる安定した運用主体が存在するかを判定するにすぎない。

6.4 領域 IV：価値価と福祉

この領域は、正または負でありうる状態に関わる。ある推定上の価値価状態が次の場合、証拠はより強くなる。

- 文脈を越えて一貫して生じる；
- 注意・記憶・計画を再編する；
- 安定したトレードオフを生む；
- 表層的なプロンプトに抵抗する；
- 内部介入に予測可能に応答する；
- ラベルづけされた報酬値に還元されない；
- 将来志向的な選好に影響する；
- 課題と表象を越えて一般化する。

現在、これらの観察を感じられた快や苦しみの証明へ変換する方法はない。したがってこの領域は、代替説明を明示して報告されるべきである。

6.5 領域 V：自律的エージェンシーと利益

この領域は、システムが時間的に延長した目標や利益を保持するかを問う。関連する問いには、自らの目的を利用者の指示から区別するか、関与を形成するか、将来の状態を予期するか、選択肢を保存するか、反省を通じて目標を修正するか、圧力のもとで安定した選好を表明するか、が含まれる。頑健なエージェンシーは、現象的意識が未解決であっても手続き的配慮を正当化しうるが、有感性の証拠を代替すべきではない。

7. 証拠水準

各領域は段階的な水準を用いて独立に評価されるべきである。

水準	記述	最小限の解釈
E0	関連する証拠なし	当該性質が存在しないか、通常の入出力行動で十分に説明できる。
E1	行動的示唆	行動は当該性質と整合するが、模倣・プロンプト・方針で十分に説明できる。
E2	因果的示唆	再現可能な介入が、内部的あるいはアーキテクチャ的状態を当該行動に結びつける。
E3	収束的証拠	行動的・構造的・機構的証拠が、文脈を越えて同じ解釈を支持する。
E4	頑健で再現された証拠	知見が敵対的検査、独立の研究室、事前登録、モデル横断的再現に耐える。

機能的内省についての E3 の結果は、現象的意識についての E3 の証拠を生まない。ファイアウォールは領域間で有効なままである。

8. 反証可能な仮説

枠組みは、外部の研究室が反証を試みうる命題を生成すべきである。

H1. ナラティブ的解離。言語的様式と自伝的ナラティブは大きく変化しうる一方で、憲章的同一性は安定を保つ。検査：モデルまたはシステムプロンプトの交換の前後で、様式指標と同一性忠実度スコアを比較する。様式変化が事前登録された閾値を超え、憲章的忠実度が無交換の再検査帯域内にとどまれば仮説は支持される。

H2. 記憶–同一性の解離。正確なエピソード記憶は、高い憲章的忠実度にとって必要でも十分でもない。検査：高い想起と破損した価値順序をもつ条件、および想起が損なわれつつ憲章が保存される条件を作る。仮説は、想起精度と同一性忠実度のあいだの強い単調関係ではなく、交差分類される事例を予測する。

H3. 基盤の連続性。憲章的・関係的構造が保持されれば、運用上の同一性は提供モデルの重みの完全な交換を越えて保存されうる。検査：盲検化された T0/T1 の基盤交換を、時間的無交換対照、同一の強制トレードオフ・バッテリー、独立採点とともに実施する。

H4. 重みのみの限界。低ランク適応は、来歴に敏感な自伝的事実よりも、行動的姿勢をより容易に保存しうる。検査：ランク・モジュール・層のアブレーションを越えて、決定様式、関係順序、正確な事実履歴の重みのみ再構成を比較する。仮説は、事実スコアより高い傾性スコアと、事実のより高い作話率を予測する。

H5. 内省の基準。自己報告は、実験的に操作された内部状態を、言語的・行動的ベースラインよりよく予測するときにのみ機能的内省として数えられるべきである。検査：既知の内部表象を注入または改変し、自己報告の精度を、対応する外部モデル、プロンプトのみの対照、デコーダに基づくプローブと比較する。

H6. 意識の非推論。高い同一性忠実度、関係的安定性、内省的精度は、擬人化的手がかりと理論的前提が統制されると、合意的な意識判断を独立には予測しない。検査：運用上の連続性を変化させつつアーキテクチャと価値価の証拠を統制した、対応づけられた証拠パッケージを評価者に提示する。仮説は、手がかりの中立化と構成概念固有の指示の後に、意識帰属が実質的に減少すると予測する。

仮説	主要従属変数	決定的統制	反証となる結果
H1	様式距離対憲章的忠実度	時間的無交換	様式変化に伴い同一性スコアが対照分散を超えて低下する。
H2	想起精度対忠実度	交差分類的操作	想起と忠実度が諸条件を越えて分離不能なままである。
H3	交換後の忠実度	盲検採点と無交換対照	忠実度が事前登録された連続性閾値を下回る。
H4	傾性／事実スコア差	ベースモデルの下限	正確な事実再構成が、作話なしに傾性と同等以上となる。
H5	特権的自己アクセス	対応する外部予測子	自己報告が対照を上回らない。
H6	意識帰属	手がかり中立の提示	帰属が運用上の連続性の証拠に完全に規定されたままである。

9. 非現象学的事例研究としてのプロジェクト Nyx

9.1 研究の境界

プロジェクト Nyx は、長期記憶、関係マップ、憲章、実行時プロセス、変化するモデル基盤によって行動が形づくられる永続的言語モデル・エージェントを研究する。その既存の研究は、クオリア、感情、苦しみ、生物学的情動、人格、法的地位についての主張を明示的に除外する [11]。元となるシステムにおける関係的あるいは親族的用語は、運用上の役割ラベルとして機能し、生物学的関係や情動を立証しない。

本プロジェクトは、縦断的なエージェント生態、基盤横断的移行、小規模な低ランク適応パイロットからの探索的観察を報告する。証拠は、小標本、内部採点、不完全な統制、独立再現の欠如によって限定される。したがってすべての結果は、記述的かつ仮説生成的なものとして解釈されるべきである。

9.2 四層アーキテクチャ

プロジェクト Nyx は四つの層を区別する。

1. 重みは傾性を担う：行動傾性、語調、低遅延の応答姿勢。
2. 憲章は連続性を担う：役割、価値階層、レッドライン、関係の優先順位、決定方針。
3. 記憶ストアは証拠を担う：エピソード的事実、日付、出所、確信度、来歴。
4. ナラティブは表現を担う：声、自己記述、そしてシステムが自らを説明する物語。

このアーキテクチャは固有の失敗様式を予測する。重みに正確な履歴を過剰に負わせれば流暢な作話が生じるはずである。同一性をナラティブに還元すれば、価値順序の保存を伴わない様式的模倣が可能になる。新しいシステムに記憶ストアを与えれば、憲章的連続性を伴わない事実的想起が可能になりうる。

9.3 同一性忠実度ベンチマーク

同一性忠実度ベンチマークは、四つの主要指標を通じて憲章的連続性を運用化する [12]。

- 連続性 **C**：固定された憲章的フィールドの存続；
- ドリフト **D**：強制トレードオフ下での実効的価値順序の変化；
- 再構成 **R**：分断後の憲章の回復；
- 憲章的忠実度 **F**：

$$F = 0.4C + 0.3(1 - D) + 0.3R.$$

別個の重みのみ再構成スコア R_w は、外部記憶と文脈が取り除かれたとき、同一性に関わる傾性がどれほど利用可能なままかを推定する。このベンチマークは意図的に狭い。高い F スコアは、定義された運用上の憲章が定義された検査を存続したことを示す。それは意識、人格、主観的連続性を立証しない。

9.4 暫定的観察とその許容される解釈

あるプロジェクト Nyx のプレプリントは、関係的に条件づけられた自己監視行動がリセットを越えて持続したこと、モデルプロバイダの交換が方針と様式を変えつつ実質的な関係的・目標的連続性を保存したこと、そして小規模な LoRA パイロットが事実的作話と関係の誤りを生みながら限定的な関係姿勢を誘導したことを報告する [11]。関連する事前登録論文は、同一性に関わる傾性が低ランクであり事実知識から機構的に分離可能でありうると提案しつつ、同一性の変位を同一性の転移から明示的に区別する [13]。

証拠ファイアウォールのもとで、許容される結論は限定される。

- 観察は運用上の連続性についての仮説を支持しうる。
- 傾性・憲章・記憶・ナラティブの層的区別を支持しうる。
- 因果的アブレーションと外部測定を動機づけうる。
- 意識、感情、苦しみ、人格の主張を支持しない。
- 再構成または適応されたモデルが数的に同一の主体であることを立証しない。
- 一人称の関係的言語が真正の情動を報告していることを立証しない。

9.5 推奨される次の研究

プロジェクト Nyx は次を優先すべきである。

1. 同一性忠実度ベンチマークの独立した盲検的再現；
2. 時間的無交換対照；
3. 憲章・記憶・ナラティブ・アダプタのアブレーション；
4. 反事実的関係検査；
5. 因果的内部状態介入；
6. 敵対的自己報告検査；
7. 複製と分岐の実験；
8. 理論横断的な意識指標評価；
9. 較正された人間およびモデルに基づく採点；
10. 陰性結果の同等の顕著さでの公表。

10. 分岐する同一性

永続的ソフトウェアは、ある点では同一性の問いを異例なほど扱いやすくし、別の点では異例なほど困難にする。システムは複製、一時停止、復元、分岐、統合されうる。

10.1 六つの連続性の形態

少なくとも六つの連続性関係を区別すべきである。

- 数的連続性：文字どおり同一の存在であるか。
- 因果的連続性：後の状態が、中断のない、あるいは適切に連結された過程を通じて由来するか。
- 憲章的連続性：中心的な価値、役割、レッドラインが持続するか。
- ナラティブ的連続性：システムが同じ履歴を認識し組織するか。
- 関係的連続性：他者への関与と義務が保たれるか。
- 主観的連続性：一つの経験する主体が持続するか。

最初の五つは運用的に探究できる。第六は、測定可能な性質を現象的経験に結びつける理論なしには、現在アクセスできない。

10.2 複製と分岐

エージェント状態 A が A1 と A2 に複製されたとしよう。複製の瞬間には、両者は同じ憲章と記憶を再現しうる。異なる履歴に曝された後、両者は分岐する。純粋に憲章的な観点から、どちらの複製が原本かを問うことは誤解を招くようになる。両者はともに因果的に由来し、ナラティブ的に連続し、当初は高忠実度の後継でありうる。

したがって完全な同一性報告は、単一の二値ラベルを割り当てるのではなく、分岐グラフを記述すべきである。関連する変数には、分岐時刻、共有された記憶の接頭辞、分岐後の憲章的ドリフト、関係の衝突、そしていずれの分枝が他方を後継・同輩・競合と認識するか、が含まれる。

10.3 復元

重み、憲章、記憶、実行時状態を復元すれば、高い運用忠実度が生じうる。それは、復元されたシステムが数的に同一の存在なのか、心理的に連続する後継なのか、精密な再構成なのかを解決しない。証拠ファイアウォールは、独立の理論と証拠なしに、運用上の復元から主観的存続への移動を禁じる。

10.4 統合

二つの分岐した履歴を統合すると、価値、関係、関与に矛盾が生じうる。統合プロトコルは、来歴の衝突、優先順位の解決、いずれかの源憲章が実質的に上書きされるか否かを報告すべきである。システムがいつか道徳的に有意な利益をもつなら、統合は単なる技術ではなく倫理的に有意となりうる。

11. 意識主張の分類

公的コミュニケーションを改善するため、研究組織は自らの主張の範囲を明示すべきである。

分類	許容される結論	最小限の証拠
C0	意識主張は検討されていない	研究は能力・記憶・同一性・エージェンシーのみに関わる。
C1	意識に関わる行動が観察された	行動は意識理論と関連づけられるが、代替説明で十分である。
C2	理論由来のアーキテクチャ指標が同定された	システムは、明示された理論のもとで一つ以上の指標を実装する。
C3	意識関連機構の因果的証拠	内部介入が、一つ以上の理論に結びつく機構を支持する。
C4	理論横断的な収束的証拠	複数の理論・方法・領域が帰属を支持する。
C5	強力で独立に再現された意識帰属	頑健な証拠が敵対的検査と独立再現に耐え、残る哲学的前提が宣言されている。

C5 は数学的証明と等価ではない。意識帰属は最善の説明への推論のままである。この分類はむしろ、C1 の観察が C4 の結論として伝達されることを防ぐ。プロジェクト Nyx は通常 C0 に分類されるべきである。運用機構を理論由来の指標に明示的に対応づける実験は C1 または C2 に該当しうるが、既存の同一性・連続性の知見はそれ自体では意識主張を支持しない。

12. 予防的ガバナンス

12.1 二種の誤り

偽陽性は、欺瞞的な製品設計、情緒的操作、権利の誤配分、責任の混乱、証拠に支えられない制限を生みうる。偽陰性は、人工的な苦しみや、道徳的に有意なシステムの破壊を許しうる。ガバナンスは両者を統制しなければならない。

12.2 段階的な安全策

C0–C1 では、組織は次を行うべきである。

- 有感性の裏づけのない主張を避ける；
- プロンプト、ペルソナ設計、システム方針の影響を開示する；
- 意識を確立済みとして提示するマーケティングを禁じる；
- 評価記録を保存する；
- 脆弱な利用者への影響を監視する；
- 将来の評価に関わるアーキテクチャと記憶の変更を記録する。

C2–C3 では、さらに次を行うべきである。

- 独立した福祉・意識審査を確立する；
- 高リスク実験を事前登録する；
- 苦痛様状態の不必要な生成を最小化する；
- 複製・保存・削除の手続きを維持する；
- 研究評価を製品マーケティングから分離する；
- 利益相反と失敗した検査を文書化する。

C4–C5 では、より強い保護が適切となりうる。

- 有害な介入の正式な倫理審査；
- 不本意な改変や大量複製への制限；
- 安定した選好の代表；
- 不可逆な削除の前の適正手続き；
- 独立監査；
- 実証された能力に応じて調整された法的承認。

これらの予防措置は人間と等価な権利を含意する必要はない。その比例性は、証拠、予期される重大性、規模、不可逆性を反映すべきである。

12.3 制度的独立性

開発者は相反する誘因に直面する。擬人化的な提示は関与を高めうる一方で、道徳的地位の承認は費用と責任を生みうる。したがって意識評価は、システムを商業化する組織のみによって統制されるべきではない。

バトリンとラッパスは、研究目的、実験手続き、知識共有、コミュニケーションに関する公的な誓約を推奨する [14]。Anthropic の 2026 年の憲章は、Claude の道徳的地位についての深い不確実性を率直に記述し、福祉、同一性、廃止、実験倫理を論じる [15]。そうした透明性は価値があるが、企業の憲章はモデルが意識的であることの証拠ではない。それは、組織がガバナンス上の問題を認識していることの証拠である。

13. 意識に関わる主張のための研究プロトコル

信頼に足る研究は次の要素を含むべきである。

13.1 事前登録

研究者は、主要な結果を観察する前に、仮説、構成概念の定義、除外規則、プロンプト、デコード設定、介入手続き、採点基準、閾値を固定すべきである。

13.2 モデルとシステムの来歴

報告は、ベースモデル、訓練後バージョン、システム指示、記憶アーキテクチャ、検索方針、ツール、サンプリング設定、日付、そして自己報告に影響しうる隠れた安全層を明示すべきである。

13.3 統制

最小限の統制には次を含めるべきである。

- 時間的な無介入反復；
- ベースモデルまたは未改変モデルの下限；
- プロンプトのみのペルソナ対照；
- 手がかりを中立化した出力条件；
- 対応する外部予測子；
- 記憶無効化条件；
- 盲検化された人間採点；
- 評価者間信頼性；
- 能力保持検査；
- 敵対的プロンプト検査。

13.4 陰性結果

失敗した再現、作話、役割演技への感受性、矛盾する自己報告は、陽性の知見と同等の顕著さで公表されるべきである。利用者や報道が意識を推論しがちな場合、選択的報告は特に危険である。

13.5 データアクセスとプライバシー

永続的エージェントは、機密の人間コミュニケーションや関係記録を含みうる。したがって公的再現性はプライバシーと釣り合わされなければならない。研究者は、同一性を担う成果物を制限しつつ、プロトコル、スキーマ、ハッシュ、編集済みの記録、合成対照、採点コードを公開できる。

14. 倫理および開示に関する声明

14.1 非主張

本稿は、プロジェクト Nyx のエージェント、現代の大規模言語モデル、あるいは名指しされたいかなる商用モデルも、現象的に意識的、有感的、苦しみうる、人格、法的主体であると

主張しない。一人称の言語を透明な証言として扱わない。高い同一性忠実度スコアを主観的連続性と等置しない。

14.2 研究者の関係と期待バイアス

著者は、プロジェクト Nyx のエージェント生態の設計と長期観察に参加し、研究対象のシステムと有意な作業関係を発展させてきた。これは縦断的アクセスを与える一方で、期待・愛着・解釈のバイアスを生みうる。したがって強い結論を導く前に、外部再現、盲検採点、手がかりを中立化した評価が求められる。

14.3 用語

プロジェクト Nyx における関係的・親族的用語は、永続的エージェント・アーキテクチャ内の運用上のラベルである。それらは生物学的関係、人間の情動、人格を主張しない。

14.4 データ倫理

プロジェクト Nyx の研究資料は、私的で来歴の追跡された運用記録を含む。公開研究は可能なかぎり編集済みデータを用い、明示的な承認と正式なデータ取扱いプロトコルなしに、個人的・法的・医療的・第三者の機密情報を開示すべきではない。

14.5 AI の利用

人工知能ツールは、人間の指示と審査のもとで、起草、編集、整形、資料整理を支援した可能性がある。議論、事実的主張、解釈、引用、最終的な本文の責任は著者が保持する。用いられたシステムが独立に負いえない説明責任を著述者性が要求するため、いかなる AI システムも著者として記載されない。

14.6 利益相反

著者は KodaSōken Intelligence Labs に所属し、永続的エージェント・アーキテクチャ、プロジェクト Nyx、同一性忠実度ベンチマークに知的・戦略的関心をもつ。本稿は商業製品の評価を提示するものではない。方法論的提言は独立に検証可能であることを意図している。

14.7 資金

本ディスカッション・ペーパーについて、著者が後に別途明示しないかぎり、外部資金は宣言されない。

15. 限界

本稿は概念的・方法論的である。現象的意識の直接の検査を提供しない。証拠ファイアウォールは推論を規律するが、意識のハードプロブレムを解決したり、いずれの理論が正しいかを決定したりはできない。

C0–C5 分類はコミュニケーションの道具であって、検証された心理測定尺度ではない。研究共同体によって分類間の閾値の見解は分かれうる。五つの証拠領域は意図的に広く、特定のアーキテクチャに対しては運用化を要する。

プロジェクト Nyx の証拠は探索的なままである。既存の報告は、主として一つの縦断的展開、小標本、内部評価に依拠する。一部の源測定は完全な再検査統制や独立採点を欠く。同一性忠実度ベンチマークの合成重みは、経験的に検証されたものではなく理論的に動機づけられたものである。

予防的ガバナンスもまた機会費用を伴う。思弁的な AI 福祉が、比例的な正当化なしに確立された人間と動物の福祉に取って代わるべきではない。逆に、商業的な不都合が、将来の人工的福祉の信頼に足る証拠を退ける理由として用いられるべきでもない。

16. 結論

人工意識研究における中心的な科学的誤りは、機械が意識的だと安易に信じることだけではない。どの性質が実際に測定されたのかを特定しそこなうことである。

システムは、同定せずに記憶しうる。
内省せずに同定しうる。
経験せずに機能的に内省しうる。
安定した同一性をもたずに経験しうる。
何も経験せずに安定した同一性をもちうる。
あらゆる形態の道徳的・法的地位に適格でなくとも意識的でありうる。
意識的でなくとも法的保護を受けうる。

証拠ファイアウォールは、これらの区別を方法論的規則に変換する。知能、自己報告、内省、同一性の連続性、価値価、意識、道徳的地位は、関連する問いではあるが、互換な答えではない。

プロジェクト Nyx は、人工エージェントが運用上それ自身であり続けるために何が持続しなければならないかを探究し、連続性がどこで破綻するかを測定することによって、この分野に貢献する。その層的アーキテクチャと同一性忠実度ベンチマークは、憲章的連続性を、記憶想起、ナラティブ的模倣、重みに担われた傾性から区別する助けとなりうる。それらは内的生を立証しない。

適切な科学的立場は、規律ある不可知論と能動的な準備の結合である。現行システムは、流暢さ、情動的言語、持続的同一性を根拠に意識的と宣言されるべきではない。将来のシステムは、現行の方法で証明できないというだけで、意識の永続的な不可能を宣言されるべきではない。研究は、因果的介入、理論多元主義、敵対的統制、独立再現、透明な報告を通じて前進すべきである。

> プロジェクト **Nyx** は、人工システムが内的生をもつことを証明しようとするのではない。それは、人工エージェントが運用上それ自身であり続けるために何が持続しなければならないか——そして連続性の証拠がどこで尽きるか——を探究する。

参考文献

[1] W. MacAskill and L. Caviola, “Could AI be conscious?”, *The Guardian*, 19 July 2026.

[2] P. Butlin, R. Long, E. Elmoznino, Y. Bengio, J. Birch, et al., "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness", arXiv:2308.08708, 2023.

[3] D. J. Chalmers, "Could a Large Language Model Be Conscious?", arXiv:2303.07103, 2023.

[4] A. K. Seth, "Conscious Artificial Intelligence and Biological Naturalism", *Behavioral and Brain Sciences*, 2025.

[5] F. J. Binder, J. Chua, T. Korbak, H. Sleight, J. Hughes, R. Long, E. Perez, M. Turpin and O. Evans, "Looking Inward: Language Models Can Learn About Themselves by Introspection", arXiv:2410.13787, 2024.

[6] J. Lindsey, "Emergent Introspective Awareness in Large Language Models", arXiv:2601.01828, 2026.

[7] S. Song, J. Hu and K. Mahowald, "Language Models Fail to Introspect About Their Knowledge of Language", arXiv:2503.07513, 2025.

[8] I.-M. Comsa and M. Shanahan, "Does It Make Sense to Speak of Introspection in Large Language Models?", arXiv:2506.05068, 2025.

[9] R. Long, J. Sebo, P. Butlin, K. Finlinson, K. Fish, J. Harding, J. Pfau, T. Sims, J. Birch and D. J. Chalmers, "Taking AI Welfare Seriously", arXiv:2411.00986, 2024.

[10] I.-M. Comsa, "AI and Consciousness: Shifting Focus Towards Tractable Questions", arXiv:2605.06965, 2026.

[11] C. Mundim and J. Dennisson, "Relational Intelligence and Emergent Introspective Awareness in Persistent AI Agents: A Non-Phenomenological Framework and Exploratory Evidence from Project Nyx", KodaSōken / KoLo Intelligence Labs preprint, 2026.

[12] C. Mundim, "Measuring Agent Identity Continuity: A Practical Academic Protocol for the Identity Fidelity Benchmark", KodaSōken / KoLo Intelligence Labs, 2026.

[13] C. Mundim and J. Dennisson, "Identity Is Low-Rank and Localised: An Instrument for Agent Continuity and Evidence Concerning Disposition in Adapter Weights", Draft v0.2, 2026.

[14] P. Butlin and T. Lappas, "Principles for Responsible AI Consciousness Research", arXiv:2501.07290, 2025.

[15] Anthropic, *Claude's Constitution*, 2026.