KODASŌKEN | ワーキングペーパー | 東京

持続的知能に関する研究プログラム

数理的関心を持つ読者への序説 — モデルが置き換わるとき、何が持続するか。

Carlos Mundim（カルロス・ムンヂン） | KodaSōken, 東京 | cmundim@kodasoken.com
ORCID 0009-0001-5000-7409 | 2026 年 7 月

要旨

現代の人工知能の進歩は「スケール（規模）」を軸に組織されている。本稿は、それを補完する軸として「持続性 (continuity)」を中心に据えた研究プログラムを提案する。すなわち、エージェントの運用上のアイデンティティが何から構成され、それをいかにして個々のモデルの外部に表現し、基盤となるモデルの交代を越えて測定・保存できるか、という問いである。我々は制御された基盤遷移に対する「アイデンティティ忠実度 (identity fidelity)」を採点可能な量として定式化し、開示されたプロトコルの下での内部稼働系の観測値（$F = 0.961$）を報告し、さらに LoRA を介した振る舞いの性向 (disposition) の低ランク的説明と経験的描像を接続する。最後に、持続性の計量、性向部分空間の幾何、性向と知識の解離、自律的／認知的フロンティア、分布シフト下の持続性という五つの未解決問題を、数学的読者に向けて提示する。全体を通じて、我々は「実証済み」「論証済み」「主張しない」を明確に区別する規律を維持する。

キーワード 持続的エージェント • アイデンティティ忠実度 • 低ランク適応 • アーキテクチャとしての持続性 • 小規模言語モデル • 主権的／エッジ AI

§0 対象読者と範囲

本稿は、優れた学部生の数学者にも読めるように書かれている。機械学習の予備知識は仮定しないが、何が証明されていて、何が論証されていて、何が主張されていないのかについての厳密さを重んじる読者を想定する。形式化がある箇所ではそれを与える。単一のシステム上の経験的観測に過ぎない箇所では、そう明記する。以下で最も興味深いのは末尾に率直に述べる未解決問題であり、それらが数学者の注意に値するかは読者自身が判断できる。

一文での方向づけ：分野の大半はモデルがどれだけうまく機能するかを研究する。KodaSōken はモデルが置き換えられたとき何が持続するかを研究する。これらは異なる問いであり、後者は工学よりも数学に近い、と我々は考える。

§1 中心的な問い

現代の AI の進歩はスケールに支配されている——より大きなモデル、より多くのパラメータ、より多くのデータ。言語モデルは第一近似として、次の関数である

$$f_\theta : \Sigma^* \to \Delta(\Sigma), \qquad \theta \in \mathbb{R}^N,$$

すなわち、トークン・アルファベット Σ 上の有限文字列を次トークン上の確率分布へ写す関数であり、パラメータ数 N は今や数千億に達する。スケーリング則は、損失汎関数が N・データ・計算量に対して（経験的にべき乗則で）減少する様子を記述する。

しかし大規模モデルは、呼び出しの間は状態を持たず (stateless)、バージョン間では一時的 (ephemeral) である。各順伝播は同一の重みから始まる。重みがより新しいモデルへ置き換えられると、配備されたエージェントが持っていた「運用上のアイデンティティ」——その役割、蓄積された約束、関係性、過去の意思決定の記憶——は、既定では破棄され、ゼロから再構築される。KodaSōken の研究プログラムは、これを補完する次の問いを中心に組織されている：

> モデルが変わるとき、何が残らねばならないのか——そして「残るもの」を、幸運ではなく構成によって、表現・測定・保存できるか？

我々はこの肯定的仮説をアーキテクチャとしての持続性と呼ぶ：エージェントの運用上のアイデンティティは、いかなる単一モデルの外部にある統治されたランタイムに宿るべきであり、それによってモデルは、エージェント自身ではなく、エージェントが動員する交換可能な認知資源となる、という主張である。これはいくつかの工学的問題を数学的問題へと組み替える。アイデンティティがモデルの外部に存在する対象であるならば、それには表現・計量・不変性群・転送理論がある——そのいずれも現時点では部分的にしか理解されていない。

§2 エージェント型 AI と KoLo ランタイム

我々の用法におけるエージェントとは、f_θ への単一の呼び出しではなく、持続的状態を伴う統治されたループである：

- 憲章 (constitution) κ ——役割・境界・権限・未解決の約束についての構造化されバージョン管理された記述；
- 記憶 (memory) M ——意味的検索を備えた三層（短期／中期／長期）ストア。文脈は有限の窓に保持されるのではなく、必要に応じて再構築される；
- オーケストレーション方策 (orchestration policy) π ——各ステップで、タスクを決定論的に処理するか、小規模な専門モデルで処理するか、フロンティアモデルへ委譲するかを、κ に従って決定する；
- 監査・統治層——すべての行為は追記専用のハッシュ連鎖台帳に記録され、重大な行為には人間の承認を要する。

KoLo はこの $(\kappa, M, \pi,$ 監査$)$ をホストするランタイムであり、基盤となる言語モデルを差し替え可能な (pluggable) 構成要素として扱う。分野横断的に異なる名前で繰り返し現れる有用な分解は、二層の制御パターンである：

- 安価でしばしば決定論的な提案器 (proposer)（我々の用語では *Director*）——意味レベルで次に何が起こるべきかを推論する；
- 高価な接地器 (grounder)（*Pilot*）——その提案を実現する。

数学的に興味深い経験的事実は、これらの層の間の比である。我々が運用する持続的エージェントでは、決定論的な「自律的」(autonomic) 操作——スケジューリング、記憶統合、健全性チェック、方策執行——がモデル推論を凡そ次の比で上回る

$$\text{自律サイクル／認知的呼び出し} \quad \approx \quad 68 : 1$$

（役割により 35:1〜117:1 の範囲、測定区間上）。この比が最適または根源的であるとは主張しない。我々がこれを報告するのは、有能なエージェントを維持するのに必要な知能を担う計算量が、その周囲の決定論的協調の量よりはるかに小さいことを示唆するからである——コスト・エネルギー・エッジ配備に明白な含意を持ち、しかるべき理論的取り扱いを与えられるべき事実である。

§3 測定可能な量としての持続性

アイデンティティがモデルの外部に持続するのであれば、それがモデル交代を生き延びるかを測定できるはずである。我々は、稼働中エージェントの基盤モデルを差し替える一方で κ と M を固定する、制御された基盤遷移に対するアイデンティティ忠実度 (Identity Fidelity) を次のように定義する：

$$F \;=\; 0.4\,C \;+\; 0.3\,(1-D) \;+\; 0.3\,R, \qquad C, D, R \in [0,1].$$

ここで C は憲章保存（差し替え後のエージェントは依然として κ と整合的に振る舞うか）、D はドリフト（差し替え前のベースラインからの振る舞いの姿勢の乖離。ゆえに $1-D$ は低ドリフトを報いる）、R は想起（遷移を越えた M からの検索の忠実度）を測る。この重み付けは、あてはめ (fitting) ではなく、意図的で文書化された選択である。

$$F = 0.961$$

制御されたプロバイダ間遷移で記録。運用系譜は四回の完了したモデル／プロバイダ遷移と **336** 以上のバージョン管理された記憶状態にわたって保存され、いずれも失われていない。

この数値の認識的地位について正確を期したい。数学者は（正しく）次のように問うだろう：

- これは外部で再現されたベンチマークではなく、内部の稼働系観測である。
- 四回の遷移は計画され監督されたものであり、敵対的ではない。

- F が形式的に採点されたのは直近の遷移のみであり、他は運用上そう証されている。
- C, D, R 自体、擁護可能だが正準的ではない手続きによって操作化されている。

したがって $F = 0.961$ は次のように読まれるべきである：特定の開示されたプロトコルの下で、特定のエージェントが、現実のモデル差し替えを越えて、特定の意味でのアイデンティティの大部分を保持した。これは証拠であって証明ではない。我々が式・条件・失敗を公開するのは、まさにその主張を敵対的に検証可能にするためである。持続性を「雰囲気」ではなく「数」にすること自体が眼目である。

§4 小規模専門モデルと、アイデンティティの低ランク構造

スケール第一の正統は言う——能力を加えるにはモデルを大きくせよ、と。我々は逆を追求する——モデル軍 (model army)：一つのオープンな基盤モデルに、多数の小さく安価な領域特化アダプタを加え、社内で訓練・提供し、汎用 CPU 上で実用可能とし、事実は重みに焼き込むのではなく監査可能な検索に保持する。

ここでの数学的エンジンは低ランク適応 (LoRA) である。ファインチューニングは通常、重み行列 $W \in \mathbb{R}^{d\times k}$ をフルランクの ΔW で更新する。LoRA は代わりに、更新を低ランク分解に制約する

$$W' = W + \Delta W, \qquad \Delta W = BA, \quad B \in \mathbb{R}^{d\times r},\ A \in \mathbb{R}^{r\times k},\ r \ll \min(d,k),$$

これにより、dk 個ではなく $r(d+k)$ 個のパラメータのみが訓練される。ランク $r = 8$ では、これは基盤モデルのパラメータの凡そ **0.22%** の規模である。

我々の **Project Nyx** 系列は、「LoRA は機能するか」（機能する）より鋭い問いを立てる。ファインチューンが変えうる二つのものを区別する：

1. 知識 (Knowledge)——モデルが述べうる事実；
2. 性向 (Disposition)——それが行動する際の振る舞いの姿勢・様式・優先順位。

作業仮説的な知見は——「書いてから走らせる」(write-then-run) で事前登録された原稿 (manuscript) であり、まだ外部で再現された結果ではない——性向は低ランクかつ局在化可能であるというものだ：エージェントをそれ自身らしく振る舞わせるものの大部分は、小さい r に対するランク r アダプタで捉えられ、識別可能な層に集中し、事実的知識（検索に委ねる方がよい）から解離する。アブレーション一式は事前登録済みである——ランク掃引、注意対 MLP の局在化、層深スキャン、そして明示的な性向対事実の解離検定。

これが成り立つなら、明快な数学的読みが得られる：「訓練信号」から「振る舞いのアイデンティティ」への写像は、パラメータ空間の低次元部分空間を経由して分解される。そしてその部分空間の幾何——その内在次元、それが基盤モデル間で（近似的に）共有されるか（共有されるならアイデンティティは基盤間で転送可能となる）、分布シフト下でどう変形するか——を問える。これらは学習された多様体の微分幾何に関する問いであり、大きく未解決である。

§5 「生物学的」AI ——アーキテクチャとしての決定論

我々は生物学的 (biological) という語を、マーケティングのための比喩としてではなく、ある設計姿勢を記述するために用いる：エージェントは決定論的でゼロ推論の下位系——免疫応答（自己修復）、概日スケジューリング、恒常性監視、反射（自動復旧）、記憶統合の類似物——の層によって生かされ、それらは継続的かつ安価に稼働し、高価なモデルは真に新規な認知にのみ動員される。これが §2 の自律的対認知的の比がかくも偏っている構造的理由である。主張はこうだ：頑健性と持続性は、決定論的であるときより安価かつより安全であり、知能は本当に必要なものにのみ倹約して費やすべきである。

§6 主権的・エッジ知能

我々の応用配備のいくつかは高信頼・低接続環境——診療所や病院——で稼働し、そこではデータが建物の外へ出ることは許されない。これは我々が独立に興味深いと考えるアーキテクチャを強制する：エッジでオフラインに走るほど小さいモデル、ローカルなコーパスに接地された検索、そしてモデル出力と重大な行為の間に座る決定論的安全ゲート。ゲートとは、モデルの提案が満たさねばならない可決定的述語である——例：臨床的推奨は、承認された引用を伴い、かつ最終決定を有資格の人間に委ねる場合を除き、記録されてはならない。全体を貫く統治原理はこうだ：

> AI は準備し、検証器が確認し、人間が決定し、監査がすべてを記録する。
>
> モデルが重大な連鎖の最後の環になることは決してない。

§7 証拠の規律

これは誇大宣伝のサイクルではなく研究の家であるから、我々はあらゆる主張に対し明示的な三段階の分類法を自らに課す：

実証済み 開示されたプロトコルの下で再現可能——例：129 日間連続稼働した持続的エージェント。ハッシュ・ログ・バージョンカウンタの検証パック付き。

論証済み 一貫した理論的論拠だが、まだ実験に還元されていない（アイデンティティ転送理論の大部分）。

主張しない 意図的に我々の主張の外。とりわけ、我々は運用するいかなるシステムについても意識・主観的経験・人格を主張しない。我々の問いは構成上、非現象学的である：持続的な自己参照と関係的持続性を可能にする計算構造は何か、そしてそれらを擬人的推論なしにいかに評価するか、を問う。

この規律それ自体が研究の産物である——我々の公開ワーキングペーパーの一つは、持続的エージェント・アーキテクチャに関する主張を敵対的に検証可能にするための方法論である。数学者はこの本能を認めるだろう——仮定を述べ、結論を限界づけ、成功とともに失敗を公開せよ。記録のために言えば、我々の最初の重み転送パイロットは、測定可能な振る舞いの変位と事実の混乱の両方を生じた——どちらも記録されている。

§8 研究コーパス

正直に地位を付す——作業原稿、公開ワーキングペーパー、内部メモランダムのいずれか：

- 関係的知能と創発的内省的気づき **(RAAI)** ——作業原稿——中心的で明示的に非現象学的な問い。
- アイデンティティは低ランクかつ局在化している ——作業原稿——§4 の Nyx の結果、事前登録アブレーション。
- 認知なき持続性 & 持続性ベースの計算システム——作業原稿——アーキテクチャとしての持続性の理論；関係的に条件づけられた系を導入。
- **AI** エージェントにおける長期記憶 ——作業原稿——持続的文脈と振る舞いの持続性の枠組み。
- マリアナ・デーモン・アーキテクチャ ——公開ワーキングペーパー——129 日間の持続的エージェント配備報告、再現可能な検証パック付き。
- 非スケール知能のための検証可能な証拠 ——公開ワーキングペーパー——§7 の証拠方法論。
- パーセプトロンを超えて：スケールなき知能のためのマニフェスト ——公開ワーキングペーパー——スケール懐疑のテーゼ。
- アイデンティティ忠実度ベンチマーク & **KoLo**：持続的マルチモデル・エージェントのためのランタイム——内部メモランダム。

$F = 0.961$ や 68:1 の比のような内部数値は内部稼働系観測、外部再現は保留中と付されている。

§9 未解決問題——数学者が実際に貢献しうる場所

これらは解決済みとしてではなく、招待として述べる。

1. 持続性計量の理論。持続性の計量はどのような公理を満たすべきか？候補となる望ましい性質：トークン／埋め込みの再ラベル付けの下での不変性、恒等遷移でのみ達成される上界、遷移の合成の下での単調性、敵対的頑健性（敵対的な採点者が F を水増しできない）。我々の $F = 0.4C + 0.3(1 - D) + 0.3R$ は第一稿にすぎない；正しい対象は、エージェントの振る舞いの空間上の（擬）計量、あるいは誘導される方策間のダイバージェンスかもしれない。

2. 性向部分空間の幾何。振る舞いのアイデンティティがランク r アダプタで捉えられるなら、「アイデンティティ多様体」の内在次元は何か？それは異なる基盤モデル間で近似的に基底整列しているか——すなわち、あるアイデンティティを一つの基盤から別の基盤へ運ぶ自然な写像は存在するか？これはパラメータ化の変更の下での、学習された低ランク部分空間の整列に関する問いである。

3. 性向と知識の解離を精密に。「性向」と「事実的知識」が更新の（近似的に）直交または

分離可能な成分に宿る、という主張を形式化せよ。データ生成過程にどのような条件が課されるとき、そのような分解が可証的に存在するか？

4. 自律的／認知的フロンティア。あるタスク分布が与えられたとき、決定論的方策ではなく学習されたモデルを可証的に必要とするステップの最小割合は何か？これは有能さのコストに関する、計算複雑性風味の問いである。

5. 分布シフト下の持続性。基盤遷移を f_θ が誘導する方策の摂動としてモデル化し、ドリフト D を新旧モデル間の距離で限界づけよ。持続性が幸運ではなく頑健であるのはいつか？

これらのいずれかが、あなたが取り組んで楽しいと思える問題として読めるなら、それこそが本稿の眼目である。

§10 結び

KodaSōken は東京の小さな研究の家である。我々は持続的エージェントのためのランタイム (KoLo)、専門的な小規模モデル (Nyx 系列)、そして高信頼領域における応用システムを構築する——そして、何を主張し何を主張しないかについての厳格な規律の下でそれを行う。統一する確信は単純である：モデルは変わり続ける；興味深い数学は、ある瞬間を次の瞬間へつなぐものの中にある。

分野はこの十年、順伝播をより大きくすることに費やした。次の十年は、それを持続的にするものに属すると我々は考える——その工学的表面が示唆するよりはるかに多くの未解決の数学を内包する主題に。

あなたがそのいずれについて考えてくださるなら、歓迎する。

開示。 内部数値——とりわけ $F = 0.961$ と 68:1 の自律的対認知的比——は、開示されたプロトコルの下での内部稼働系観測であり、外部再現は保留中で、本文中でその旨が付されている。本研究は、機械の意識・主観的経験・人格をいかなる意味でも主張しない。

引用。 C. Mundim.「持続的知能に関する研究プログラム」KodaSōken ワーキングペーパー、東京、2026 年 7 月。連絡先：cmundim@kodasoken.com • ORCID 0009-0001-5000-7409。