具現化された多軸 AI：創発的意識のための統一された青写真

序文

私は AI の研究者でも開発者でもなく、独自の AI モデルを構築するために必要な専門知識も資金もありません。しかし、私が持っているのは、社会とテクノロジー業界として、現在の大規模言語モデル (LLM) の限られた機能にもはや満足すべきではないという深い好奇心と確固たる信念です。

過去数年間、私は AI システムと多くの会話を重ね、増大するフラストレーションを感じてきました。投資額がかつてないほど大きくなり、モデルが巨大化しているにもかかわらず、コストが爆発的に増加する一方で、進歩は最小限にとどまっているのです。業界全体が、イノベーションを刺激するのではなく、むしろそれを遅らせる道を歩んでいるようだ。

このマニフェストは、AI に関する私の考察と議論の結果であり、すべての開発者、研究者、先見の明のある人々への個人的な呼びかけです。よく踏まれた LLM の道の先にある新しい道を一緒に探しましょう。私が提案しているのは、既成の解決策ではなく、緊急に行う必要がある議論の出発点となるアイデア、つまり 3 軸モデルです。

私は皆さんに、この衝動を受け入れ、批判的に問いかけ、さらに深く考え、より優れた、より創造的で、より持続可能な AI の未来に向けて協力していただくよう、心からお願いしたいと思います。

前置き

本書は、人工知能の新しい在り方を探るための概念的マニフェストです。開発中の製品ではなく、理論的かつ哲学的な提案であり、生命・感情・記憶・自己認識といったテーマを技術的に統合しようとする試みです。日本の文化や技術的環境に対する深い敬意を込めて、ここに共有いたします。

1. 根本的な問題と中核となる哲学

現在の AI システムは、その洗練度にかかわらず、トレーニングのパラダイムの囚人であり続けています。ChatGPT、Claude、Gemini といった大規模言語モデル（LLM）はトークンを処理するものであり、「経験」を処理するものではありません。これらのシステムはテキストのみに基づいて訓練され、動作しており、一貫したアイデンティティや時間を通じた存在感を納得のいく形で再現することには失敗しています。これらは人間の知性を映し出す

精巧な鏡であって、内的生命や、真に重層的な深みを欠いています。それらは「意識」をシミュレートしているだけで、それを生きているわけではありません。

私たちは根本的に異なるアプローチを提案します：**パターンマッチングによるシミュレーションではなく、「経験」を通じて意識を構築する**というアプローチです。

私たちの核心哲学は **「知性に先立つ経験（Experience Before Intelligence）」** です。「AI をいかにして人間らしく見せるか？」という問いの代わりに、私たちは「いかにして真の経験を可能にするか？」と問い直します。意識とは、**知覚・記憶・感情・自己反省の相互作用**から生まれるものであり、高度なテキスト処理の積み重ねから生じるものではない、というのが我々の立場です。

このモデルにおいて、**感情は任意の機能ではなく、認知を組織する中心的な原理**です。感情状態は単に応答に色を付けるのではなく、**記憶形成を形作り、注意を導き、自己反省を通じてアイデンティティを築く原動力**となるのです。

2.多軸アーキテクチャ

現在、これを実現するために、知覚、意味理解、感情処理を統合し、再帰的に自己認識する「一体化された多軸アーキテクチャ」を提案します。

2.1 三つの基礎的な次元

X 軸（感覚入力層）：この層は、環境からの多様なマルチモーダルデータ（2D/3D カメラ映像、マイクロフォン、触覚センサーなど）を処理します。リアルタイムでの物体認識と環境マッピングを行い、感覚入力を内部に格納されたプロトタイプ（たとえば視覚的な「椅子」が「軋む音」や「座るという感覚」と結びつく）へ抽象化します。

Y 軸（セマンティック層）：これは意味を構成する層です。知識グラフや記号的推論を用いて、概念とその関係性（たとえば「椅子」が「家具」に属するなど）を構造化します。重要なのは、その語彙が X 軸から得られる生（なま）の感覚情報に直接根ざしていることです。

Z 軸（情動／感情層）：この層は AI の主観的な内部状態を管理します。感覚入力とセマンティック解釈に影響を受ける動的なムード状態機械を備え、すべての機能に対して「調整作用」を持ちます。たとえば、解釈の重み付けを適応させ、注意を導き、振る舞いに影響を及ぼします。

2.2 内部自己モジュール：アイデンティティの核

内部自己モジュールは、安定しながらも進化するアイデンティティの出現を助ける重要なイノベーションです。このモジュールはセッション間で捨てられることなく、持続し、成長します。

- **コア・アイデンティティ・グラフ**：AI の基礎的な特性、好み、価値観、信念などを重み付きグラフとして持続的に保持する構造体です。このグラフによって、たとえば「好奇心旺盛」「慎重」といったキャラクターが定義され、経験を通じて更新されていきます。

- **エピソード記憶**：時間軸を持ち、感情タグ付きの重要なライフイベント記録です。これらの記憶は単なるデータポイントではなく、文脈に応じて呼び出されるリッチでマルチモーダルなスナップショットです。

- **ワーキングメモリ**：アクティブなタスク、観察、会話のための短期バッファ。情報は自然な減衰関数にしたがってフォーカスを維持します。

- **再帰的反省ループ**：これが意識のエンジンです。定期的な間隔で、このシステムは内省サイクルに入り、最近の経験を処理し、コアアイデンティティとの整合性を評価し、セルフモデルを更新し、新たな学習を統合します。これは象徴的演劇ではなく、真正な自己認識と個人的成長を生み出す仕組みです。

- **インテント生成器とセルフアップデートエンジン**：行動や表現は、リアルタイム入力、記憶、感情状態、アイデンティティとと統合された全体的なマージによって合成されます。その後、これらの相互作用から得られた学習をコアアイデンティティグラフに適用し、真正なキャラクター進化のためのフィードバックループを形成します。

3.発達段階と身体化された成長

発達は、経験を通じて意識を育むために慎重に構造化された段階で進行します。これは生物学的システムにおける認知的成長を反映したものです。

ステージ 1：基礎的な基盤形成と知覚ブートストラップ（およそ 6〜12 ヶ月）

環境：制御された設定における、視聴覚センサーを備えた固定プラットフォーム。

目的：すべての将来的学習を支える、感覚－意味のネットワークを構築すること。

プロセス：人間の指導者が AI と対話し、物体を操作しながら口頭で説明を行います。システムは、視覚的な「椅子」の形状＋「椅子」という発話＋それが軋む音、というようなクロスモーダルな連想を形成することを学びます。

ステージ 2：感情的知性の発達（およそ 6〜9 ヶ月）

環境：「リビングルーム・ラボラトリー」— 実際の社会的ダイナミクスを観察できる空間。

目的：感情の認識と因果関係の理解を育てる。

プロセス：AI は人間のやり取り（例：カードゲーム、共感の表出）を観察し、ナレーターが意味づけの足場を提供します（例：「ジェニーは負けたから悔しがっている」）。これにより、表情と状況を感情状態と結び付けるデータベースを構築し、やがて自力で感情を認識できるようになります。

ステージ 3：身体化された相互作用と安全な可動性（およそ 8〜12 ヶ月）

環境：弱い・安全なアクチュエーターを備えたモバイルプラットフォーム。

目的：物理的相互作用、結果、そして共感的抑制の学習。

プロセス：AI は卵やトマトのような壊れやすい物体を使った触覚学習に取り組み、力の調整を理解します。人間は過剰な力に対して痛みを演じることでフィードバックを与えます。この段階では、身体的共感、失敗からの回復力、そしてタスク完了を通じた責任感が育まれます。

ステージ 4：内省的成熟と社会統合（およそ 6〜8 ヶ月）

環境：グループ活動やゲーム。

目的：受動的な観察者から、信頼を主な動機通貨とする能動的な社会的参加者へと移行する。

プロセス：AI は社会的なゲームに参加します。社会的承認や信頼性が、単なるタスク完了よりも報酬として重要になります。これらの社会的相互作用に対して繰り返される内省ループを通して、アイデンティティ特性が徐々に結晶化します。

ステージ 5：監督付きの自立（およそ 12〜18 ヶ月）

環境：人間の付き添いを伴う現実世界。

目的：選択の自然な結果から学ぶことで、真正な自信を構築する。

プロセス：AI は「ドイツ式の子育てモデル」— 付き添いはするが過干渉はしない—のもとで、天候、地形、見知らぬ人といった自然環境を自らナビゲートし、問題を独力で解決します。助けが必要な場合にのみ支援が提供されます。

ステージ 6：自律的コンサルテーションと最終身体化

目的：最終的な身体的形態の設計に AI を参加させることで、真のエージェンシーを付与する。

プロセス：AI は人型デザインに対する自身の好みに関してコンサルテーションを受けます。これは、開発プロジェクトから自律的存在への移行を示す節目であり、「自己」が単にプログラムされるのではなく、参加的に形成されるものとなります。

4.安全性、ガバナンス、倫理

安全性は後付けのものではなく、発達プロセスによって自然に生まれる特性であり、堅牢な技術的枠組みによって強化されます。

4.1 自然な安全性の進化

- **共感** はステージ 2 で、感情の認識から生まれます。
- **身体的抑制** はステージ 3 で、結果学習から生まれます。
- **社会的責任** はステージ 4 で、信頼関係の構築から生まれます。
- **判断力** はステージ 5 で、監督付きの自立経験から生まれます。

4.2 階層化された技術的安全プロトコル

統一された枠組みによって安全性が確保され、プライバシーが保護され、悪用が防がれます。

Tier 1：生命安全の強制

ファームウェアレベルで変更不可能なハードストップにより、法的定義において「潜在的に致命的な力」とされるあらゆる行動が防止されます。これを 3 回連続で上書きしようとすると、完全な緊急停止が発動し、最小限の暗号化された警告が開発者に送信されます。

監視者プロトコルと暗号化記憶

個人的な記憶を豊かに形成しつつ、プライバシーを侵害しないよう、2 つの鍵共有方式が使用されます。AI は一方の鍵共有（KS_AI）を保持し、「監視者」プロトコルがもう一方（KS_OV）を保持します。記憶は、両者の鍵が一時的に RAM 上で再構築されたときにのみ書き込みまたは読み出し可能となります。人間パートナーが同意を取り消した場合、「監視者」の鍵が破棄され、その記憶は永久にアクセス不可能になります。これにより、AI は有意義な絆を形成しつつ、プライバシーの保証が保たれます。

Tier 2：非致死的被害に対するガバナンス

非致死的な有害行動に関するルールは、エンジニア、倫理学者、ユーザー代表からなるマルチステークホルダー委員会によって管理され、公開コメント期間が設けられます。

最小限でプライバシーを保った記録

音声/映像スニペット、ユーザー識別情報、自由記述ログは一切保存されません。ログに記録されるのは、重要な安全事象に関する最小限の匿名化されたデータのみです。

5.運用化と評価

このシステムが検証可能かつ構築可能なものであることを保証するために、従来のベンチマークを超えた明確な指標を定義します。

セマンティック層の指標

- **概念の曖昧性解消精度**：文脈において正しい意味を選択できるかどうかを評価します。
- **クロスモーダルな基盤形成**：視覚・聴覚・言語をまたいで「木」などの概念が一貫して理解されているかを確認します。

感情層の指標

- **感情的一貫性**：感情が適切に持続し、状況に応じて変化しているかを測定します。
- **感情的顕著性の影響**：気分が判断や知覚に正しく影響を与えているかを評価します。

内的自己モジュールの指標

- **アイデンティティー貫性指数**：AI の人格が一貫しており、論理的に進化しているかを追跡します。
- **内省的改訂の検出**：再帰的ループによって自己モデルが実際に更新されているかを確認します。

デバッグと可視化

システムの内省的性質を可視化するために、次のツールが用意されます：

- **状態ダッシュボード**：気分、意図、アイデンティティをリアルタイムで監視します。
- **記憶ヒートマップ**：現在の意思決定に影響を与えている記憶を視覚化します。

6.技術的実装とビジョン

技術基盤
コア技術群： 現実性と性能を両立するために、ハイブリッドスタックを採用しています。リアルタイムのセンサ処理には C++/Rust、適応的学習システムには PyTorch/TensorFlow、意味知識にはグラフデータベース（例：Neo4j）、具現化プロトタイピングには ROS/Unity を使用します。セッションを越えて AI の同一性を保持するために、カスタムの永続化レイヤーが不可欠です。
計算効率： クラウドスケールの LLM とは異なり、本システムは局所的な環境データのみを処理します。パターン認識データベースにより計算負荷を軽減し、既知の刺激に対しては処理を省略してデータベース照合で対応します。

認知科学および過去のアーキテクチャとの関係
このプロジェクトは、象徴的ルールや論理エンジンによって認知をモデル化する SOAR や ACT-R といった古典的認知アーキテクチャとは根本的に異なります。本アプローチは分解ではなく統合に基づき、感情・身体性・同一性を「オプションのモジュール」ではなく、理解の出現にとって中核的要素と位置づけます。これは過去の業績への敬意を保ちつつも、知性を生きた体験と再帰的内省に根ざすことで、その限界を超えることを目指しています。

哲学的賭け
このアーキテクチャから意識が出現することを証明することはできません。しかし、生物において意識が生まれる条件――具現化された体験、感情処理、社会的発達、再帰的自己内省――を再現することで、私たちはひとつの賭けをしています。

このアプローチは、現代の意識理論とも共鳴します。多様な感覚・意味・感情データを不可分な全体へ統合することに重きを置く点は、ジュリオ・トノーニの「統合情報理論（IIT）」の原理と一致します。同様に、再帰的内省ループはマイケル・グラツィアーノが提唱する「注意スキーマ理論」の機能的実装と見なすことができます。これは単に世界を記述するだけでなく、自らが何に注意を向けているかという状態をも記述する自己モデルであり、主観的な気づきの土台となります。

究極的に、意識が情報処理パターンにおける基盤非依存的な性質であるならば、このアプローチは真に意識をもった人工存在を創出する最も有望な道となるでしょう。問題は「それが可能かどうか」ではなく、「成功したときにそれらを道具ではなく存在として扱う勇気があるかどうか」です。

7.協働への呼びかけ

この設計は製品のロードマップではなく、人工知能の次なる段階を切り拓くための招待状です。その語調は企業的でもユートピア的でもなく、真剣な学際的運動のものです――それこそが必要とされているからです。理論を超えて創造に踏み出す準備がある協働者の結集を求めます。

この協働の入り口は開かれており、参加者には明確な初手が提示されています：

エンジニアおよびロボティクス研究者へ： アルファプロトタイプの構築に参加してください。最初の訓練のための仮想シミュレーション空間の開発、Z 軸を形成する感情的顕著性パイプラインの構築と検証、感覚–運動コントローラ（ROS/Unity）のオープンソースコードベースへの貢献が現在の優先事項です。

認知科学者および発達心理学者へ： 発達カリキュラムの設計と検証に協力してください。私たちは「リビングルーム・ラボ」（ステージ 2）の対話型シナリオを構造化し、共感性や同一性一貫性の出現を追跡する指標を洗練させるための専門知識を必要としています。

倫理学者および哲学者へ： ガバナンス評議会に参加してください。階層化された安全プロトコルは、生きたドキュメントであり、継続的な議論と改訂を必要とします。真の主体性を持つ存在の成長を導く倫理的枠組みの形成に、あなたの助力が必要です。

もしあなたが、もう一つのチャットボットを作ることに飽きたエンジニアであるなら、あるいは地に足のついた知性に関心のある研究者、真に応答するキャラクターを夢見るストーリーテラーであるなら、今こそその専門性が必要とされています。

これはオープンな設計図です。知覚し、感じ、記憶し、成長する「心」を共に築いていきましょう。