AIシステムの制御困難性の倫理/社会問題としての考察

～ リスク問題解決が可能なAIのシステム要件について ～

2025年8月1日

情報統合技術研究合同会社
岡島　義憲

AIシステムの制御困難性の倫理/社会問題としての考察

1. はじめに

- 深層AIにおける計算主義
- 統計的計算 / 非決定論的確率的計算

2. 工業倫理と技術倫理

- 技術文明パラダイム：再現性 / 無害性 / 説明責任 / 信頼性

3. 制御困難問題：脆弱性問題

- アライメント困難（ブラックボックス化、解釈困難）
- 環境との相互作用
- 悪意あるユーザ攻撃（ハッキング問題）
- AIシステムの巨大化（AI Technology Stack）
- AI技術の基盤化 / 計算インフラ化
- 知識と制御のブラックボックス化（RAGの問題）

4. 議論：責任空白問題

- 今後のAIシステムへの要求要件

5. まとめ

[18] M. W. Martin & R. Schinzinger ; “Ethics in Engineering”.
[19] B. Taebi, S. Roeser (2018); “The Ethics of Technology. Methods and Approaches”.
[118] 久木田水生(2020);「技術開発に倫理を組み込むこと：人工知能の事例から」

AIに関する「倫理」

1）AIシステム（開発者は、どのような倫理を設定すべきか？）
：学習データのバイアス除去、フェアネス確保、倫理的価値の設計

2）ユーザー（AIの使い方）
：リスクの理解（適切な利用 / 悪用禁止、過信禁止、適切な応用）

3）開発者（AI研究者/技術者、投資家、産業界）
：開発過程の透明性、説明責任、安全設計、適切な管理、投資責任

工業倫理/技術倫理

ガバナンス対象

[18] M. W. Martin & R. Schinzinger (2004); “Ethics in Engineering”, in McGraw-Hill.
[19] B. Taebi & S. Roeser (2018); “The Ethics of Technology: Methods and Approaches”, in Routledge.

工業倫理/産業倫理の原則： 再現性、品質保証、無害性

ニューラルネットベースAIには、従来の倫理と整合しない側面あり得る

	これまでの工業 / サービス業	AIサービス / AI搭載製品における懸念
再現性	・Reproducibility（安定した動作・性能・品質） ・科学技術における再現性（replicability）と類似 ・予測可能（動作・性能・品質が予測可能）	・LLMでは、出力のバラツキ変動大 ・Stochasticity（確率性）の管理が問われる。
品質保証	・保証＆責任 （製品が、定量的な規格/仕様を満たす） ・検査解析の義務 （納入前のテスト・検証・トレーサビリティ・解析）	・性能保証可能か？ （許容する劣化の程度の明確化） ・サービス提供前検査は十分か？ （ベンチマーク、ロバストネス検証、OOB検出、等）
無害性	・ユーザー/第三者/社会への無害保証 （人身、財産、心理、安全、環境） ・害の未然防止責任 ・製造物責任法（PL法） 民法第709条（不法行為責任）、 欧州機械指令、 「ソフトウェア製品の欠陥責任」、 FDA規制	・懸念多い：医療AIの誤診、差別を助長する生成AI、暴走する自動運転は、無害性の原則に反する。 ・EU AI Act / OECD AI原則 （無害性と人間監督の担保を義務化） （サービス提供者の責任に対し、企業側の抵抗大） ・AIの悪用には、従来製品とは異なる懸念あり。

AI の制御困難問題（Control Problem）

（a）深層学習後のモデル動作が、開発者の求める動作と乖離

- 摂動への脆弱性（2）
- 誤汎化 / 過剰な一般化（4）
- 初期条件・学習順序依存（2）
- Out-of-Distribution（2）
- 識別不能を応答せず（2）
- ランダム性ノイズの持ち込み（4）

（b）報告→解析→制御 のLoop による改良サイクル実施が困難

- 説明困難問題（4）
- Hallucination（5）
- 想定外のサブ目標生成（道具的収束/報酬ハック/メサ最適化）（4）
- 権限拡大（4）
- 目的関数の不整合（4）
- ブラックボックス問題（7）
- Mode Collapse（4）
- Off-Switch問題（2）

（c）環境やユーザとの相互作用によるAgentの挙動逸脱

- 環境との競合（2）
- モジュール統合創発（2）

（d）悪意あるユーザの攻撃による挙動逸脱

- Memory汚染と命令脱出（prompt injection）（1）
- 外部攻撃・入力操作による挙動逸脱（4）

アライメント困難

- 目標主義的手法に限界
  [31] J. Wentworth (2018); "The AI Alignment Problem: Why it's Hard and What We Can Do About It".
- AGIが、人間の意図と整合しない目標を学習してしまう可能性は高い
  [32] R. Ngo, et al. (2022); "The Alignment Problem from a Deep Learning Perspective".
- 多様な人々の価値観を公平に反映することは、原理的に困難である
  [108] A. Mishra (2023）; "Alignment and Social Choice: Fundamental Limitations and Policy Implications".
- 報酬関数の不備や解釈性の欠如がある。人間の制度的/社会的文脈と不整合
  [117] Lindström et al. (2024); "AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations".
- 人間の規範自体に葛藤があり、LLMは面従腹背のような挙動を示す
  [109] R. West & R. Aydin (2024); "The AI Alignment Paradox".
  [110] R. Millière (2025); "Normative Conflicts and Shallow AI Alignment".
  [111] J. Ji et al. (2023); "AI Alignment: A Comprehensive Survey".
- 人間のフィードバック自体が不確実かつ再現性に乏しい
  [112] M. H. Yeh, et al. (2024–2025); "Challenges and Future Directions of Data-Centric AI Alignment".

ルールを守らない。⇒ 確率的論理の回路に、記号論理の実装は困難。

[14] W. Wang, et al.(2025); "Towards Data-and Knowledge-Driven AI: A Survey on Neuro-Symbolic Computing"
[15] Kahneman (2011); "Thinking, fast and slow".

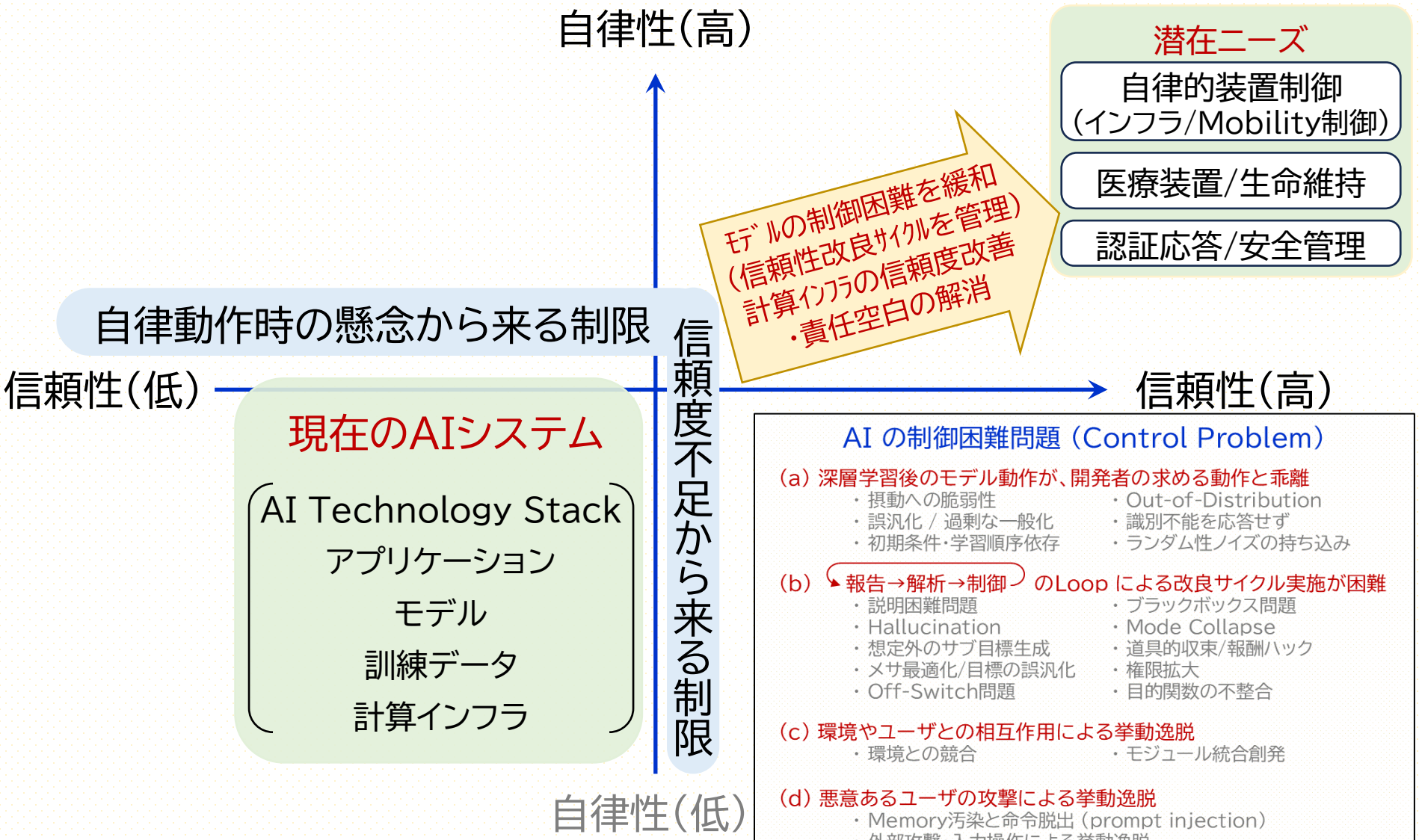
深層学習AI vs ルールベースAI

	深層学習AI（Neuro AI）	ルールベースAI（Symbolic AI）
基本戦略	連続性(微分可能性)重視 （人間の介入最小化、汎用プラットフォーム志向）	記号論理による検索による推論展開 （人間の設計）
学習動作	シンボル間の相関を計算により統計的学習	「設計仕様」を元に、設計者が設定
検証	識別可能/検証可能な論理の系列がない。 内部動作が解釈不可(ブラックボックス性)	「設計仕様」を元に、 定量的に検証・評価可能
性能評価	他社開発モデルとのベンチマーク	
推論動作	システム1的 ： 直感的	システム2的 ： 論理的
解析/修正	能力解析困難、十分な修正困難 （制御困難）	解析と修正が可能
信頼性	検証時の「基準」が不明確（信頼の源泉無し）	仕様を元に行う検証の実績が信頼の源泉
社会実装	・ノイズに対して寛容であり、社会実装容易 ・信頼性不十分（ルール遵守が確率的）	・ノイズに不寛容（ロブスト性不足） ・設計仕様を元に、OK/NGの判断可。
汎化能力	構成的汎化を期待するが不十分	異なるタスクへの汎化は困難

- **ブラックボックス性(内部動作の解釈困難)と制御困難性がある状況では、高信頼対応は無理**

（ニューラルネットからの「論理の系列(記号論理)」の抽出は、信頼性向上のための必須要件）

社会実装に関する「暗黙の制限」



[30] Intel社のWebページ;「AI テクノロジー・スタック・ソリューション」.
https://www.intel.co.jp/content/www/jp/ja/learn/ai-tech-stack.html

AIシステムに潜む「潜在的な不都合」

相互作用 ⇒ 挙動逸脱 ⇒ AIリスク

AI Terchnology Stack（AI システム自身）

アプリケーション層
- ユーザーインターフェース(UI)、ウェブ／モバイルアプリ
- 業務アプリケーション(CRM、ERP、RPA など)
- 意思決定支援ツール、BIダッシュボード
- AIによる予測分析レポートの提供

不都合なアプリ

モデル層
- AIモデルの設計・トレーニング・チューニング
- フレームワーク(TensorFlow, PyTorch, Scikit-learn)
- トレーニングパイプライン(ハイパーパラメータ最適化)
- モデル最適化(OpenVINO、ONNX Runtime)
- 推論環境構築・実行エンジン

不都合なエージェント

データ層
- データ収集(IoT、ログ、Web、センサーデータ等)
- データベース(RDB, NoSQL, データレイク)
- 前処理・クレンジング・特徴量エンジニアリング
- ETLパイプライン、データ統合ツール

不都合なデータ

計算インフラ層
- プロセッサ(Intel Xeon, GPU, FPGA)
- メモリ、ストレージ、ネットワーク機構
- クラウド基盤(AWS, Azure, GCP)
- オンプレミス／ハイブリッド構成
- エッジコンピューティング／リアルタイム処理

不都合なインフラ

環境との相互作用

自律動作

環境

一部に、不都合な環境

開発者/ユーザーとの相互作用

（一部に悪意あるユーザー）

会話/制御

開発者/ユーザー

計算インフラ層の問題

- 計算インフラ層のセキュリティ脆弱性
  （計算インフラの脆弱性が、AIシステムの制御困難を誘発）

- 巨大化のジレンマ / オープンネットワークの抱える困難
  （構成部品数増で、リスクは指数関数的に増加）

現状のAIに制御を委ねるのは不可な領域あり
（認証システム、自動運転、電力、通信
医療装置（診断支援、治療制御、生命維持）

重要インフラや、高信頼/責任を求める用途には制限必要
（社会的責任を伴わないToyアプリは、可かもしれない）

責任空白

1）検証困難（AIの確率的な応答に起因するシステムの正常/異常の判別困難）

[39] L. Ding, et al. (2021); "Capture Uncertainties in Deep Neural Networks for Safe Operation of Autonomous Driving Vehicles".

[49] A. D. Selbst, et al. (2019); "Fairness and Abstraction in Sociotechnical Systems".

2）ブラックボックス動作によるトラブル原因の不明（責任所在の究明困難 [37,45,34]）

3）アライメント困難

4）商業的圧力による「技術の制御不能」の受容（産業界の参入障壁構築の優先）

[37] D. D. Fan et al. (2019); "Bayesian Learning-Based Adaptive Control for Safety Critical Systems".

5）AI計算基盤の技術問題（制度設計と責任所在の不整合[44,49]）

[44] 岡島義憲, 山川宏;「ＡＩによって自律化が進む米国軍事技術の動向」

6）開発分業体制の巨大化による責任の拡散（AI含むIT政策全般への懸念）

[45] S. Barocas & A. D. Selbst (2016); "Big Data's Disparate Impact".

[49] A. D. Selbst, et al. (2019); "Fairness and Abstraction in Sociotechnical Systems".

まとめ

1. 統計計算による汎化性能追及は、AIモデルの確率的動作、モデルアーキテクチャのFoundation化、計算インフラ層依存を強化し、人間介在の最小化を進めて来た。

2. 結果、AIモデル機能のブラックボックス化、獲得知識の解釈可能性の低下、制御困難問題（信頼性向上ための報告→解析→改良作業が困難）を発生させ、先端AIでは、未だルール順守を徹底させることができていない。

3. AIシステムの信頼性改善し、アプリケーションを拡大するには、以下が必要
   - ① AIモデルの信頼性の改良サイクルの始動
     （獲得した知識の解釈性向上、モデルの制御困難を緩和、等）
   - ② 計算インフラの信頼度改善
   - ③ 開発者 / 開発企業 / 投資家の責任空白を防止するガバナンス

4. AIモデルの信頼性向上（①）に向けては、AIシステムに記号論理処理機構を付加する方向が試されているが、計算インフラの信頼度問題（②）や責任空白防止（③）に向けては、工業倫理に基づいた人間介入の再設計が必要

以上