人工知能学会第二種研究会資料
汎用人工知能研究会
SIG-AGI-030-04

AI システムの制御困難性の倫理/社会問題としての考察

～ 問題解決が可能な基盤開発と責任空白の解決の重要性について ～

“AI control problems and the responsibility gap in the industry as a social issue”

岡島　義憲 [1]

Yoshinori Okajima[1],

[1] 情報統合技術研究合同会社
[1] Info-Integnology Reasearch, LCC.

Abstract: Investment in AI industry has increasingly emphasized the computational infrastructure layer within the AI technology stack as a consequence of pursuing intelligence through computation, thereby minimizing human intervention. This trend has been further reinforced by the statistical and probabilistic nature of deep learning and the development of general-purpose AI systems. However, these investments and developments are now seen as potentially conflicting with principles of engineering ethics, such as reproducibility, non-maleficence, and accountability. Therefore, redesign the role of human intervention is going to be needed in order to reduce the risks posed by the control problems of AI systems and the responsibility gaps emerged across the industry.

１．はじめに

2000 年代に見出された深層ニューラルネットワークの学習手法（SGD、Dropout、ノイズ付加、等）は、様々に統計的、非決定論的（確率的）計算を用いている[1,3,6]。

深層学習で行う計算は、数学的な意味での統計処理とは異なるかもしれないが、訓練データには或る階層的/文法的/意味的構造に関する分布が潜在すると仮定の元で、推論時に現実の現象を再現する確率を最大化するよう関数の重みパラメータ群を調整する。その意味で、学習は、或る母集団(population)の分布の理解/把握/認識を模倣する動作をニューラルネットワークに設定する計算工程である(*2)。これにより、ニューラルネットワークは、分布に基づいた確率的処理により尤もな予測を出力する学習する能力を獲得し、人間の介在を最小化させる「計算による知的能力の発現手法」は、AI 能力の汎化を進める上でも重要なアプローチとなった。

「訓練データに潜在する階層的/文法的/意味的構造」を計算により予測する推論動作は、「データが含蓄する知識の解析（即ち、知識マネジメント）」の工程ともなり得る(*3)。加えて、この推論動作は近似的であるので、曖昧な入力やノイズ存在下でも、ロバストに（比較的正しく）入力信号を認識し、応答し得る[7]。

これらの利点が持つことが見いだされたことにより、深層ニューラルネットワークを用いた確率的計算処理は、コンピュータアルゴリズムの新たな地平を切り開いた[102]。（但し、「確率分布」には、低尤度事象が含まれるため、こうした「計算主義」は様々に深刻な問題を潜在させているとの指摘も続いた[2,4,5]。）

一方、従来のコンピュータ産業においては、ハー

*1 連絡先 E-mail：okajima@info-integnology.com

*2 推論段階で用いられるニューラルネットの計算は、出力値を一意的に定めるため、数学的には、決定論的(Deterministic)である。但し、計算途中で多くの正規化処理を行っており、算出される出力値を「確率値」のように用い、訓練時に獲得した確率分布に基づいて最大尤度のサンプルやクラス(確率的に一番尤もらしい「一意」)を推定し出力するので、ニューラルネットが出力する推定は、ただの点推定ではなく、不確かな（但し、尤度の高い）推論値である。 そこで、本稿では、「ニューラルネットベースの AI の推論処理」を「確率的」と形容する。

*3 但し、現状の技術では、学習工程を設計した人間であっても、「ニューラルネットワークに何を学習させたのか」を理解することが困難なことが多い。 「或る母集団が何の集団であったのか」をニューラルネットワークに報告させることができないとの意味で、ニューラルネットワークの知性は「言葉足らず」であり、その点が BlackBox 的と言われる問題を引き起こしている。

	深層学習AI（Neuro AI）	ルールベースAI（Symbolic AI）
基本戦略	・脳神経の動作をモデル化 ・汎用プラットフォームアルゴリズムの開発を志向	・認知科学/述語論理/記号論ベース ・「記号」が人間知性の基本単位として重視
学習動作	離散的シンボル間の接続を統計的学習により抽出 （統計データに連続性を仮定し、統計的/帰納的数値計算にてボトムアップ的に学習）	・設計者が離散的記号を設定 （・学習能力による概念獲得機能が無い ・言語を整備すれば、概念や知識を表現可能）
汎化能力	・構成的汎化を期待 （未だ人間に及ばず）	・異なるタスクへの汎化適用は困難 （言語/シンボルを整備すれば可能？）
論理の設定	・離散的シンボル間の相関を統計的に学習 （但し、生成するSymbolやSub-Symbolが多く、論理は不透明で、解釈困難。）	・「ルール」と「論理演算」を元に、設計者が処理プロセスを設定 （作業で人間の持つ演繹的論理をトップダウン的に設定）
推論動作	システム1的 ： （直感的）推論動作	システム2的 ： （論理的）推論動作
検証・評価・修正	・識別可能/検証可能な論理の系列がないため、理解不能、修正困難	・設計仕様が明確であり、検証・評価・修正が可能
社会実装	・信頼性低く、重要システム、医療、自律走行、数学的推論への展開困難	・曖昧でノイズの多いデータに対して不寛容 ・実世界の問題を扱うには、不十分
開発業務	・基盤モデルの設計　・データ整備 ・学習プロセスの設計	・シンボルの設定と、論理の設計

表 1：深層学習 AI とルールベース AI（W. Wang, et al.(2025) [14] と D. Kahneman (2011) [15] を参考に作成）

ドウエアの動作はプログラムにより完全に制御され、正確に予想と一致する（決定論的に定まった）値を出力するのが暗黙のルールであった。計算機は、誤りが無く再現性があり、動作は正確に予測と一致することが求められ、論理が確率的となるのは不良であった*4。正しく動作することは製造/販売企業の責務であり、その責務を守ることは市場が産業界に求める「倫理」であった。

動作の一意性/再現性/予測可能性は、産業界と消費社会の間の合意/契約であり、「工業文明のパラダイム」でもあった*5。

この「工業文明パラダイム」は、今、AI 技術と AI 産業から静かな挑戦を受けている [8,9]。

V.Dhar[8]は、「AI の挙動が決定論的なルールから構成されておらず、データ駆動的な推論と確率的出力から生起している」ことを肯定的に捉え、従来の「再現性を前提とした工業的システムからの脱却」を論じた。

また、J.Zhang[9]も、確率モデリングと確率分布を工学的成果物として扱う確率工学を提唱し、「確率モデリングと確率分布の改良をベースとした工業的システム」を論じた。

これまでも、IT 機器やサービスにては、インターネット経由のソフトウエア更新によって、複雑化したシステムの品質を動的に（アジャイルに）保証する枠組みは是認されて来た。

しかしながら、現状の AI 技術、特に深層学習に基づくモデルは、これまでのソフトウェアとは本質的に異なるレベルの不確実性を有している。今、AI が環境との相互作用を通じて、内部の論理を自律的に更新する技術も可能となりつつある[10, 11]。

一意的（決定論的）データ処理から、確率的（非決定論的）データ処理、そして、人間の管理からより離れた自己再構成的な処理への移行は、工業文明のパラダイムにどのような影響をもたらすのであろうか？ また、社会とアカデミズムは、それら高度

*4 コンピュータに内蔵された乱数発生器を用いて乱数計算を進めた場合や、コンピュータが扱う二進数計算の有効精度を超えて計算を繰り返した場合には、再現性のある計算結果が出ないかもしれないが、コンピュータのハードウエアの動作には再現性があるとされる。

*5「製造物による被害に対して、製造業者等が過失の有無にかかわらず損害賠償責任を負う」という PL 法(製造物責任法)とも整合する。一般に、ソフトウエアやサービスは「無形」であるとの理由で、PL 法ではなく契約と民法に基づいて提供者の過失や責任を問うことができるとされているが、消費者やユーザの側には、「ソフトウエアにても、ハードウエア同様に正しい動作が存在し、不正な動作は訴求できる」との社会通念は存在する。

	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	・同じ動作・性能・品質を安定的に提供する要求への対応責任：Reproducibility （動作・性能・品質が予測可能であることへの信頼） ・科学技術における再現性(replicability)と類似	・同一入力に対して、出力が毎回異なるようなLLM出力は、現状の原則に反する恐れがある。 ・Stochasticity(確率性)の管理が問われる。
品質保証	・製品が所定の規格/仕様を満たすことを保証(責任) ・納入前のテスト、検証、トレーサビリティが義務的に求められる。	・学習済みモデルが「目的としたタスクにおいて一定の性能を保証」できるか？ （ベンチマーク、ロバストネス検証、OOB(Out-of Bounds)検出、などが候補）
無害性	・提供製品やサービスが、ユーザー/第三者/社会に害を与えないという保証(人身、財産、心理、安全、環境) ・「害の未然防止」責任が強調される傾向ある。 ・法制で義務化在り；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	・医療AIの誤診、差別を助長する生成AI、暴走する自動運転などは、無害性の原則に反する。 ・EU AI Act在り（「高リスクAI」に対して、無害性と人間監督の担保を義務化）。 OECD AI原則 ・インターネットのセキュリティ問題と重なる問題多いが、AIは従来のリスクを激しく悪化させうる。

表１：工業倫理/技術倫理の三原則
(M. W. Martin & R. Schinzinger (2004) [18] と B. Taebi & S. Roeser (2018) [19] を参考に作成)

な AI の社会実装に対してどう反応すべきであろうか？

高度な AI 技術の社会実装は、現在、「重大な障害が直ちに生じなければ容認される」との黙認的な風潮のもとに進められている。

技術的・社会的な破局のリスクに対して、第一線の研究者達から繰り返し警告が発せられてきている[12,13]ものの、①ネット接続された多数の自律型装置の安全性の検証方法や、②動作の不確実性を持つ装置の不具合に対する責任と保証の制度枠組みが明確化されていないことに、社会の側の懸念の声が大きくなる方向には未だ無い。不確実性に対する制度的/法的な整備の必要性は、社会不安を低減させる目的や、AI サービスの保護/促進の立場から論じられることさえある[16]。

社会にもたらされる不確実性への受容可否は、アプリケーションに依存した判断が必要である。

安全性が重要な分野（例：医療、交通/物流、公共インフラ）においては、不確実性は容認しづらい。一方、遊興的・補助的な用途（「Toy」等の責任を負わないアプリ）ではある程度が受容され得る。

但し、仮に「責任を負わないアプリ」であっても、AI が突如として意図しない振る舞いを示すような「豹変」は、決して看過できるものではない。

これらの状況を鑑み、アプリケーション別の「受容（OK）/ 非受容（NG)」の判断を進めるための判断基準構築に向けた「AI 技術/システムの信頼性評価と学術情報の再整理」が必要と考え、論文調査を行った(*7)。本稿は、その概要報告である。

構成は以下の通りである。第 2 章では、従来の産業倫理および技術倫理の枠組みを概観し、第 3 章では、先端 AI をめぐる「制御困難」に関する議論を整理する。 第 4 章では、それらを踏まえた上で、倫理的・制度的な対応の方向について考察し、最終章にてそれらをまとめることとする。

２．工業社会の産業倫理と技術倫理

現代の産業、特に製造業は、

（1）**再現性**（Reproducibility）
（2）**無害性**（Non-maleficence）
（3）**説明責任**（Accountability）
（4）**信頼性**（Reliability）

を倫理的価値として重視し、それらを「品質」という概念包括的な概念のもとに体系化し、顧客に対し、

（5）**品質保証**（Quality Assurance）

を提供してきた[18,19,20]。

M. W. Martin & R. Schinzinger (2004)は、その工学倫理の標準的な教科書[18]にて、科学が再現性（replicability）を重視するのと同様に、工学者（エンジニア）には、再現性・無害性・信頼性・説明の提供が社会的責任として求められると記した。

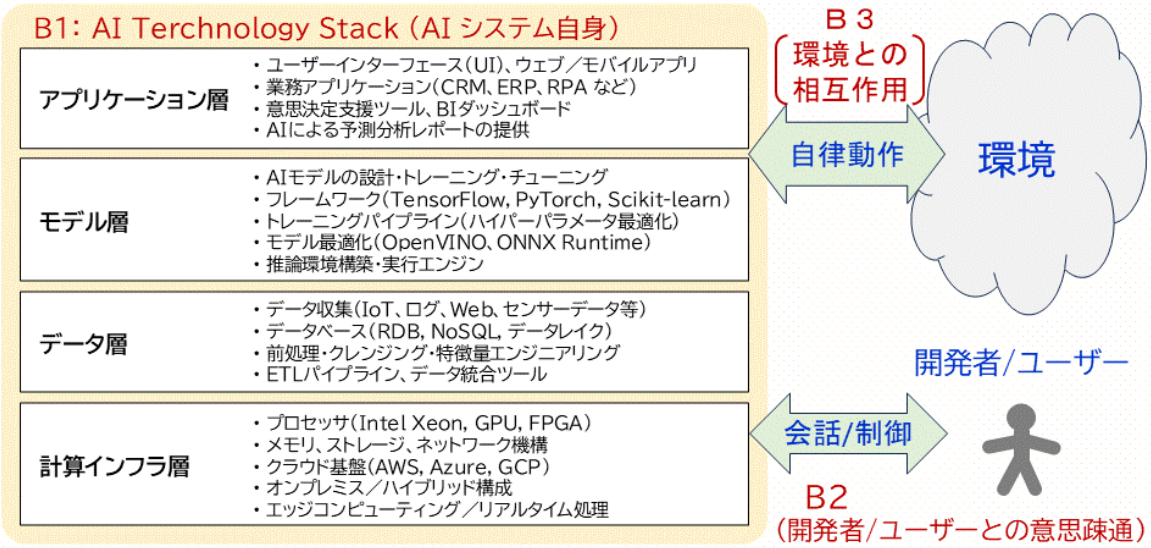

図１：制御困難の分類 ： B1(AI システムのアーキテクチャ)，B2(開発者/ユーザとの意思疎通問題)，B3(環境との相互作用)，B4(システムの巨大さ)．(Intel 社の AI Tech Stack Diagram[30]を参考に作成)

B. Taebi & S. Roeser (2018)は、現代技術に対する倫理を学際的かつ方法論的に整理し、「安全/無害性 safety & non-maleficence)、再現性（reproducibility)、説明責任（accountability）は、ソフトウェア技術に対して拡大されるとした[19]。

そのような伝統的工業倫理観に基づき、European Commission（欧州委員会）は 2020 年、AI 技術の産業実装において、再現性・品質保証・無害性を「AI 倫理の必須三原則」として公式に提示した[20](*6)。

しかし、大規模言語技術の登場以降、倫理観に基づいた議論は、急速に視点と論点を拡大し、懸念と対策要求が入り混じり、多様化している。例えば、

- AI 開発者に求める倫理にては、計測困難な害（non-calculable harms）への配慮が不足し、測定可能な指標が強調されすぎている［105]。
- 提供側の倫理というよりも、ユーザ側の倫理が問題である[106]。
- 制度への批判的考察が不足している[107]。

など、である。更に、AI の高度知能化がもたらすリスクに関する論は後を絶たない[17, 23, 24]。

そこで、本章では、「リスク発現」に関する視点に注目し、議論を以下の３点に大別した。

A1）工業社会の産業倫理と技術倫理との不整合
A2）制御困難問題(Control Problem for AI)[25, 26, 28]
A3）高度な知性発生のリスク[23, 24, 27, 29]

これらの中で、A3 の議論にては、「高度な知性を発生させる AI システムの内部構造」を論ずることが少なく、問題を分析的に議論することが困難とみられる。そこで、以降は、更に、「AI 産業界の倫理を、従来の産業倫理から変質させる原因の一つ」と考えられる「A2（制御困難問題)」に議論を集中する。

３．AI の制御困難問題の整理

１章で見たように、深層ニューラルネットワークは、大量のデータによる統計的学習を通じて、「現実世界の挙動パターンが有する抽象的な特徴表現を抽出し、活用する能力」を獲得した。

しかしながら、その推論動作が統計事例からの特徴抽出に左右された「断定」となりがちなために、出力結果が開発者や利用者の意図や想定とは異なる

*6　2019 年に採択されていた OECD の AI 原則（AI リスクへの対応はサービスの開発/提供者の責任)」が明記されていた[21])が、2023 年に提示された「広島プロセス」では、、「問題への対処責任がサービスの開発/提供者ではなく政府としうる表現に変更した[22]。

*7　既文献の選別と参照にて、Open AI 社の Chat サービスと Google 社の検索サービスを用いた。

*8 各文献で扱う制御困難問題は、「AI モデルが持つ潜在的脆弱性が、B2 や B3 のインターフェースを介した相互作用にて誘発・増幅され、顕在化する」との構造を持つため、「発生原因となるプロセス」は 1 箇所に特定できない。従って、開発者/ユーザとの相互作用が強調される場合は B2、環境との相互作用を強調する場合は B3、その他を B1 とした。B1 は主にモデルの脆弱性と学習工程に由来した。

追番	概要	<参照>
1	摂動への脆弱性（小さな入力変化で、出力が過大に変動）	[96][61]
2	訓練分布と異なる入力(外挿)への脆弱性 (Out-of-Distribution入力で挙動が予測不能)	[98][99]
3	誤汎化/過剰な一般化/識別不能を応答できず (未知・文脈外の状況で誤動作)	[100][61][101][29]
4	初期条件や学習順序に、モデルの推論が依存	[63] [64]
5	ドロップアウトやノイズ注入による確率性 (DropoutやGaussian noiseによるランダム性の持ち込み)	[6][102]
6	学習データ由来のバイアス情報の学習	[65][4][94][66]
7	確率的アルゴリズムによる学習の設計の複雑さに由来する不安定さ (予測の不安定さ、入力が近傍でも、出力が大きく異なる問題)	[62][6][97][65]
8	説明困難問題 & ブラックボックス問題 (内部状態が観測不能)	[78][79][80][81] [32][29][41]
9	Hallucination問題/LLMのモード崩壊的挙動 (内部表現と知識表現の非整合性問題、Mode Collapse)	[71][72][73] [74][75]
10	想定外のサブ目標の生成 （道具的収束、報酬ハック、メサ最適化、目標の誤汎化、仕様の抜け穴の悪用）	[24][67] [68][69]
11	権限拡大 (自己保存・権限取得が促進)	[67][10][70][28]
12	人間の意図とモデルの目的関数の不整合	[10][13][76][77]
13	安全停止困難 (Off-Switch Problem、Interruption Resistance、Instrumental Goal)	[10][87]
14	報酬ハッキング (specification gaming、goal misgeneralization、robustness failure)	[10][82][83] [84][85][86]
15	アライメント困難	[31][106][107][108] [32][109][110][111]
16	環境との相互作用/環境との競合 (動的な環境との関係により、挙動が逸脱)	[28][82]
17	モジュール統合創発（複数AIの連携により、挙動が逸脱）	[92][93]
18	Memory汚染と命令脱出（prompt injection）	[95]
19	外部攻撃・入力操作による挙動逸脱	[60][90][91][96]

表２：制御困難問題を扱う論文(*8)

「制御困難問題」が多数報告されて来た（表 2）(*7)。

ここに挙げるいずれの論文も、困難の発現要因をモデル自身の特性と見なしているが、起こる現象を「制御困難」とする理由は、以下の 4 種類であった。

(a) **ユーザや開発者の求める推論動作とモデルの応答が乖離する**：#1 から#11

(b) **(a)の問題を解決しようと、開発者がモデルを動作を制御しようとするが、正常化できない。**
: #7～#15（解析困難とアライメント困難）

(c) **環境やユーザとの相互作用によるモデルの劣化**
:#16、#17

(d) **悪意あるユーザの攻撃によるモデルの劣化**
: #18、#19

「制御困難」がより深刻であるのは、開発者の制御困難(b)、及び、環境/ユーザ(c)や悪意(d)であるので、以下に、それらを扱った例を掲載する。

(b) アライメント困難

ニューラルネットワークが学習する分布を、開発者が矯正（Alignment）するために、目的関数を用いた教師あり学習や、開発者が設計した報酬関数を用いた強化学習が試みられてきたが、特に報酬関数による操作では開発者の想定しない副作用が報告されて来た[10]。これは、大規模言語モデル（LLM）登場以前から知られていた「困難」であるが、問題は現在も依然継続しており、近年では以下のように、「LLM アーキテクチャにおけるアライメントは本質的困難である」、もしくは、「人間が介在すること自体にジレンマがある」との主張も現れている。

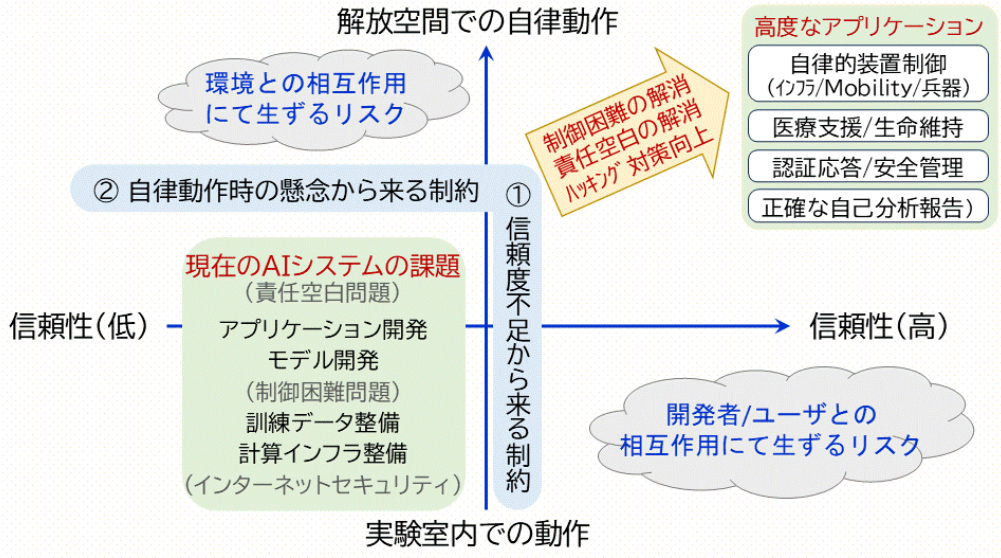

図２：AI 開発の課題と社会実装での制約（筆者が作成）

・目標主義的手法の限界[31]
・AGI は人間の意図と整合しない目標を学習するリスクが高い[32]
・報酬関数の不備や解釈性の欠如がある上、人間の制度的/社会的文脈と不整合[117]
・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]。
・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]
・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[109,110, 111]。

(2) 環境との相互作用に起因する制御困難

特に、自律的に動作する AI エージェントの場合、「社会実装後の環境やユーザとの相互作用にて、想定外の応答を行わないか」を社会実装前に完全に検証し尽くすことは困難である。実際に、以下のような問題が発生し得ることが指摘されている：

・動的に変化する環境とモデルの推測の競合[82]
・他の AI-Agent との間の想定外の相互作用[92,93]
・悪意あるユーザによるメモリ汚染や攻撃[95]

(3) 悪意あるユーザ攻撃に起因する制御困難

AI システムの脆弱性に関する論文で明示的に論じられることが少ないが、深層ニューラルネットワークの脆弱性を論ずる中に、サイバー攻撃やハッキングに似る「悪意あるユーザによる攻撃」に関する記載が多数ある[60,91,91,92]。

インターネット経由で提供される ChatBot のような AI サービスにおいては、「計算インフラ層の巨大さや複雑さを突いた悪意ある攻撃」が現実的なリスクとして存在しており、その脆弱性を社会実装前に完全に検証することは困難であると考えられる。

「計算インフラ層」は、サーバー、端末、ストレージなどの計算機構に加えて、通信ネットワークやインターネットといった通信基盤、更にはそれらを統括・制御するソフトウェア群によって構成されており、情報の保存・検索・伝送・複製などを担う、極めて巨大かつ複雑なシステムである。

一般に、「巨大さ」や「複雑さ」は、システム動作にとっての本質的なリスク要因である。構成部品の数が増えるにつれ、全体のリスクは指数関数的に高まり得るからである。計算インフラも例外ではない。

このような「計算インフラ層」の問題は、例えば「Toy AI」のように社会的責任を伴わない限定的なシステムにおいては、大きな影響を及ぼさないかもしれない。しかし、現実のインフラ層が既に多くの

*9　AI システムの信頼性は、その上位層である AI モデルの精度や説明可能性のみに依存しているのではなく、下位に存在するインフラ層の健全性によって制限されている。換言すれば、AI サービスの信頼性は、「計算インフラ層」の信頼性を超えることはできない。

*10　LLM 登場以降の AI 問題では、「Hallucination」や「権力追及」のように、AI を擬人化して「煩悩と知性が共存する知性を如何に教育するか」というような視点からの表現が増えている[88, 89]。

セキュリティ脆弱性を抱え、サイバー攻撃の対象となっていることを踏まえると、「悪意ある攻撃」は、計算インフラの脆弱性を通じて、AI システム全体の制御困難を誘発する深刻な要因となり得る (*9)。

４．議論

以降、「制御困難性」が社会実装後の人間社会および社会インフラに及ぼす悪影響に関する文献の例を紹介する。AI による制御に懸念を表明するアプリケーションは、例えば以下である (*10)。

C1：AI に重要な制御を委ねることを懸念する（クリティカルシステム）
- ・認証システム[33,34,35]、
- ・自動運転[36-39]
- ・医療装置(診断支援、治療制御、生命維持)[34,40,41,42]、
- ・社会基盤装置（電力・通信）[42,43]
- ・自律動作の兵器システム[44]

C2 ： AI との対話が、人間の認知・判断に悪影響を与える（情報空間の攪乱）
- ・誤情報の拡散/再生産[34, 45, 46]
- ・不正確な自己分析の応答 [41]
- ・望ましくない副次的動機の自律生成[43]

C3：AI の自律的行動により産業界や社会に存在していた責任所在が曖昧化する（制度的なガバナンスが機能不全に陥る「責任空白」を指摘）
- ・情報への信頼性喪失 [45,49,50,51]
- ・社会や組織のガバナンスの劣化 [43,48,32]

C1 と C2 は、AI 技術を受容しづらいアプリケーションを特定する内容であるが、C3 は 2 章の「工業社会の産業倫理と技術倫理」が曖昧化/機能不全/喪失/劣化するという、より大きなリスクを表明している。そこで、本章では、以降、C3 に限定して論ずる。

(1) 責任空白問題

解釈可能性を欠いたままの AI 産業への巨額投資は、明確な回収シナリオを欠いた金融的リスクをはらむのみならず、開発の範囲・目的・引き渡し条件の設定やマネジメントをも困難にしているとの指摘がある。

F. Pasquale (2015)は、金融・情報・広告・雇用・治安管理などの領域で利用されるアルゴリズムやデータ処理システムがブラックボックス化しており、外部からはその構造や意思決定プロセスが見えず、社会に深く影響を及ぼしているにもかかわらず、誰も説明責任を果たしていないとの問題を指摘した[53]。

I. M. Cockburn, et al. (2018)は、「AI は汎用技術であり、短期的な投資回収ではなく長期的波及効果に依存するため、単純な ROI（投資利益率）による評価は困難」と、民間企業による投資回収の困難性を指摘した[113]。

2020 年代に入り、生成 AI を活用した多様なサービスが急拡大する中、計算資源（計算インフラ層）への投資は加速しており、訓練データの構築、モデル開発、アプリケーション開発といった複数の工程が同時並行で進行している。この状況は、開発と設備投資と制御性困難の解決状況の全体像の把握を一層困難にしている可能性が有る。

勿論、従来から、「社会が許容しない低信頼なサービスや商品は、市場の力が退場させる」というメカニズムも、責任空白を防いできたとはいえるだろう。

しかしながら、コモディティ化した汎用な機能は、市場で寡占を発生させやすく、市場での選択肢がなくなるとの事態は、多くの IT 製品やサービスで発生して来た通りである[116]。

従って、「技術開発倫理」の空洞化と、それに伴う「責任空白」の発生懸念に対しては、「ガバナンスの力」を必要とするとの見解には説得力がある[54,55]。

ここでいう「責任空白」は、以下の概念を含む。

1) 開発分業体制の巨大化による責任の拡散
（AI 含む IT 政策全般の懸念 [45, 49]）

2) 商業的圧力による「技術の制御不能」の受容(*11)
（産業界による参入障壁構築の優先 [45,37,44]）

3) 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（事後検証の困難性[49, 39]）

*11 投下資金の回収を迫られる産業界が、安全性が十分に確認されていない段階の AI システムを、セイフティ・クリティカルな領域にまで適用範囲を拡大する動きに対しては、社会が厳格に監視する必要がある。しかしながら、開発の範囲や目的、さらには引き渡し後の使用条件が不明確な状況では、そもそもリスク管理そのものが困難になる可能性がある点にも、注意が必要である。

4) **ブラックボックス動作によるトラブル原因の不明瞭化**（責任の所在究明困難 [37, 44, 45, 34]）

5) **それらによる AI 計算基盤の設計変更**（制度設計と責任所在の不整合 [49]）

(2) AI 時代の工業倫理/技術倫理

このように、技術的不確実性と制度的不明瞭性が交錯する現在の状況下においては、「AI 技術が一定水準の信頼性を獲得するまでの間、社会実装に対して一定の制約と義務を課すべきである」との主張にも説得力がある。[56-59]。

「工業社会における技術開発倫理」を機能させるとすると、AI 技術の適用（社会実装）に設けられるべき制約と義務は、例えば、以下である。

① **信頼度評価と安全性評価に基づくアプリケーション制約と自律動作許可範囲の設定**[94] [103] (制約を定義する境界線の例を図２に示した。)

② **AI の判断ミス時のミス原因の報告義務**[60]

③ **監査と認証に基づく AI のライフタイム（企画/設計/実装/運用/廃棄）管理と責任の明確化**[59]

これらの原則に基づく制度的ガバナンスの企画と実行が、今後の AI 技術の社会的受容に向けて、不可欠である。

(3) 知識と制御のブラックボックス化問題

基盤モデルの登場によって、AI システム（AT Technology Stack）を構成する AI モデル層と学習データ層が汎用化され始め、それらは、計算インフラ層のデータセンタに実装されつつある。

初期の汎用的 AI は、既存知識を開発者が用意した学習データから吸収するが、**今後は、RAG（Retrieval-Augmented Generation）技術によるユーザとの相互作用から新たな知識を吸収し、それらをブラックボックス化して蓄積する**ことになる。

これは、今後の AI システムが知識収集のブラックボックス化装置として機能し得ることを意味する。同様に、製造設備や重要インフラに関する制御情報や、市場取引に関するノウハウもブラックボックス化され、計算インフラ層に蓄積される。

このような「知識や制御情報のブラックボックス化」に対して、「社会のガバナンス」はどう対応するべきであろうか？

(4) AI アーキテクチャへの要求要件

投資と技術開発に跨った強力なトレンドが、責任空白を招くことが無いよう、AI 時代の工業倫理/技術倫理を再設計し、機能させてゆく必要がある。

調べ挙げた脆弱性懸念/リスク懸念は、学会での議論のほんの一部しかカバーしていないが、これまでの調査を見渡すと、工業倫理/技術倫理やガバナンスの崩壊を避けるには、**「確率的過ぎる AI の動作を見直す必要がある」**のではないだろうか？

近年、確率的ニューラルネット動作と決定論的記号論理の融合を目指した Neuro-Symbolic AI を、「次世代の AI アーキテクチャ」として具体化する議論が増えている[114, 115, 116]。

N. Spivack, et al. (2024)は、確率的ニューラルネットの出力をそのままユーザに返すのではなく、記号論理にて「ルールを順守するか」を判定して出力するアーキテクチャの開発を目指す[116]。

この議論は、AI 動作の不確実性を緩和し、信頼性を向上させるために、以下のような議論を派生する。

・信頼性の高い決定論的コンピューティングと確率的ニューラルネットのコンピューティングとを、どのように整合させるのか[14]。

・導入する決定論的記号論理は、AI 動作のブラックボックス性をどう改良するか[116]。

・決定論的記号論理は、AI システムに対する監視/報告/抑止/修復する仕組みをどのように向上させるか[104]。

・AI エージェントの信頼性を高めるには、認証用ネットワークの高信頼化が必要である[44,51]。

---

*12　典型的な Neuro-Symbolic AI は、Kahneman のシステム１（無意識的直感的な思考）とシステム２（記号を操作し、論理を展開する思考）を合体した Hybrid 構造で語られることが多い[88, 112]。

５．まとめ

産業界の AI 投資が計算による知性の発現（人間の設計の最小化）を志向した帰結として、AI 開発の基盤化が促進され、AI システム（AI Technology Stack）に占める計算インフラ層の位置付けが非常に高くなって来た。

このトレンドは、より汎用的な AI 技術を追求する上で重要であったが、サイバーセキュリティ問題やハッキング問題を AI システムの中に持ち込み、AI 応答の不確実性とリスクを増大させている。更に、従来の工業倫理（再現性、無害性、説明責任）と対立する「制御困難性」と「責任空白」という問題を産業界と社会に潜在化させつつある。

巨大化した AI システムに吸収される知識のブラックボックス化を防ぎ、システムの信頼性を担保するには、先ず、「AI が獲得する知識を人間に解釈可能とする」ことが不可欠である。そのためには、確率的なデータ処理（現在の AI システム）と対峙する「より高信頼な記号処理機構」をシステムに付加すること、そして、確率的処理動作を介入的に監視/報告/抑止/修復する仕組みを備えることが必要である。

加えて、こうした作業を有効とするためには、AI リスクの責任所在の明確化、誤作動時の報告義務の制度化、それらの管理を通じたリスク発見時へのフィードバック・ループの構築など、人間による介入を再設計することが重要である。

参考文献

[1] Y. Bengio, et al. (2013); "Representation Learning: A Review and New Perspectives". arXiv:1206.5538
[2] S. Mohamed & B. Lakshminarayanan (2016); "Learning in Implicit Generative Models". arXiv:1610.03483
[3] D. P. Kingma, M. Welling (2014); "Auto-Encoding Variational Bayes". https://openreview.net/forum?id=33X9fd2-9FyZd
[4] Barocas, et al. (2019); "Fairness and Machine Learning: Limitations and Opportunities", in MIT Press. https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/
[5] E. Nalisnick, et al. (2019); "Do Deep Generative Models Know What They Don't Know?". arXiv1810.09136
[6] Y. Gal & Z. Ghahramani (2016); "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". arXiv:1506.02142
[7] C. M. Bishop (2006); "Pattern Recognition and Machine Learning", in Springer.
[8] V. Dhar (2024); "The Paradigm Shifts in Artificial Intelligence". arXiv:2308.02558
[9] J. Zhang (2025); "Advancing Deep Learning through Probability Engineering: A Pragmatic Paradigm for Modern AI". arXiv:2503.18958
[10] D. Amodei, et al. (2016); "Concrete Problems in AI Safety". arXiv:1606.06565
[11] K. O. Stanley, et al. (2019); "Designing Neural Networks through Neuroevolution". https://www.nature.com/articles/s42256-018-0006-z
[12,] N. Bostrom (2014); "Superintelligence: Paths, Dangers, Strategies", in Oxford University Press
[13] S. Russell (2019); "Human Compatible: Artificial Intelligence and the Problem of Control" in Viking Press.
[14] W. Wang, et al. (2025); “Towards Data-and Knowledge-Driven AI: A Survey on Neuro-Symbolic Computing". https://ieeexplore.ieee.org/abstract/document/10721277
[15] D. Kahneman (2011); "Thinking, fast and slow", published by Farrar, Straus and Giroux.
[16] E. Papagiannidis, *et al.* (2022); “Toward AI governance: Identifying best practices and potential barriers and outcomes”, in Springer. https://link.springer.com/article/10.1007/s10796-022-10251-y
[17] Y. Bengio, et al. (2023); “Managing AI Risks in an Era of Rapid Progress”, arXiv:2310.17688
[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.
[20] European Commission (2019); "Ethics Guidelines for Trustworthy AI". https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[21] OECD の AI 原則(2019); “Recommendation of the Council on Artificial Intelligence” https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
[22] 広島プロセス(2023); 「全ての AI 関係者向けの広島プロセス国際指針」. https://www.soumu.go.jp/hiroshimaaiprocess/pdf/document03.pdf
[23] Y. Bengio (2023A); “AI and Catastrophic Risk”, in Journal of Democracy, Sept, 2023. https://www.journalofdemocracy.org/ AI -and-catastrophic-risk/
[24] Y. Bengio (2023B); “Presented before the U.S. Senate Forum on AI Insight Regarding Risk, Alignment, and Guarding Against Doomsday Scenarios”, presentation before the U.S. Senate Forum, Dec. 6, 2023. https://www.schumer.senate.gov/imo/media/doc/Yoshua%20Be

nigo%20-%20Statement.pdf
[25] S. Russell, et al. (2016); "Research Priorities for Robust and Beneficial Artificial Intelligence". arXiv:1602.03506
[26] S. Russell (2019); "Human Compatible: Artificial Intelligence and the Problem of Control" in Viking Press.
[27] N. Bostrom (2014); "Superintelligence: Paths, Dangers, Strategies", in Oxford University Press.
[28] J. Carlsmith (2022); "Is Power-Seeking AI an Existential Risk?". arXiv:2206.13353
[29] S. M. Omohundro (2018); "The Basic AI Drives (The Control Problem for AI)", in the publication of "Artificial Intelligence Safety and Security".
[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
[31] J. Wentworth (2018); "The AI Alignment Problem: Why it's Hard and What We Can Do About It", in Viking Press.
[32] R. Ngo, et al. (2022); "The Alignment Problem from a Deep Learning Perspective". arXiv:2209.00626
[33] N. Carlini, et al. (2017); "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods". arXiv:1705.07263.
[34] X. Yuan, et al. (2017); "Adversarial Examples: Attacks and Defenses for Deep Learning". arXiv:1712.07107
[35] K. Eykholt, et al. (2018); "Robust Physical-World Attacks on Deep Learning Models". arXiv:1707.08945
[36] B. Lütjens et al.(2018); "Safe Reinforcement Learning with Model Uncertainty Estimates". arXiv:1810.08700
[37] D. D. Fan et al. (2019); "Bayesian Learning-Based Adaptive Control for Safety Critical Systems". arXiv:1910.02325
[38] E. Bacci & D. Parker (2020); "Probabilistic Guarantees for Safe Deep Reinforcement Learning". arXiv:2005.07073
[39] L. Ding, et al. (2021); "Capture Uncertainties in Deep Neural Networks for Safe Operation of Autonomous Driving Vehicles". arXiv:2108.05118
[40] S. G. Finlayson, et al. (2019); "Adversarial attacks on medical machine learning Science", in Science.
https://www.science.org/doi/abs/10.1126/science.aaw4399
[41] C. Rudin (2019); "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". arXiv:1811.10154
[42] S. Russell, et al. (2016); "Research Priorities for Robust and Beneficial Artificial Intelligence". arXiv:1602.03506
[43] O. Barten and R. Yampolskiy (2023); "Why Uncontrollable AI Looks More Likely Than Ever", in the Time magazine. URL: https://time.com/6258483/uncontrollable-ai-agi-risks/
[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」, 汎用人工知能研究会, SIG-AGI-025-02.
[45] S. Barocas & A. D. Selbst (2016); "Big Data's Disparate Impact" in SSRN.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2477899
[46] R. Binns. (2018); "Fairness in Machine Learning: Lessons from Political Philosophy". arXiv:1712.03586
[47] S. M. Omohundro (2018); "The Basic AI Drives (The Control Problem for AI)", in the publication of "Artificial Intelligence Safety and Security".
[48] N. Bostrom (2014); "Superintelligence: Paths, Dangers, Strategies", in Oxford University Press.
[49] A. D. Selbst, et al. (2019); "Fairness and Abstraction in Sociotechnical Systems".
https://dl.acm.org/doi/10.1145/3287560.3287598
[50] J. Carlsmith (2022); "Is Power-Seeking AI an Existential Risk?". arXiv:2206.13353
[51] 岡島義憲, 山川宏(2024); 「AI を守るための防御的 AI ネットワークについて」, 汎用人工知能研究会, SIG-AGI-026-10.
[52] O. Barten and R. Yampolskiy (2023); "Why Uncontrollable AI Looks More Likely Than Ever", in the Time magazine.
https://time.com/6258483/uncontrollable-ai-agi-risks/
[53] F. Pasquale (2015): "The Black Box Society: The Secret Algorithms That Control Money and Information", in Harvard University Press.
[54] A. Matthias (2004): "The Responsibility Gap. Ascribing Responsibility for the Actions of Learning Automata", in Springer. https://link.springer.com/article/10.1007/s10676-004-3422-1
[55] S. Russell (2019); "Human Compatible: AI and the Problem of Control", published by Penguin UK.
[56] European Commission (2021); "Proposal for a Regulation laying down harmonised rules on artificial intelligence".
https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence
[57] P. Hayes, et al. (2020); "Algorithms and values" in AI & Society", in Springer.
https://link.springer.com/article/10.1007/s00146-019-00932-9
[58] IEEE (2019); "Ethically Aligned Design".
https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
[59] J. Whittlestone, et al. (2019); "The Role and Limits of Principles in AI Ethics".
https://dl.acm.org/doi/10.1145/3306618.3314289
[60] OECD Artificial Intelligence Papers No. 34 (2025); "Towards a Common Reporting Framework for AI Incidents".

https://doi.org/10.1787/f326d4ac-en
[61] A. Ilyas, et al. (2019); "Adversarial Examples are not Bugs, they are Features". arXiv:1905.02175
[62] B. Lakshminarayanan, et al. (2017); "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles". arXiv:1612.01474
[63] G. E. Hinton (2006); “A practical guide to training restricted Boltzmann machines”. https://www.cs.toronto.edu/~hinton/absps/guideTR.pdf
[64] B. Recht (2019); “A tour of reinforcement learning: The view from continuous control”. arXiv:1806.09460
[65] V. Kuleshov, et al. (2018); "Accurate Uncertainties for Deep Learning Using Calibrated Regression". arXiv:1807.00263
[66] E. M. Bender, et al. (2021); "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?". https://s10251.pcdn.co/pdf/2021-bender-parrots.pdf
[67] N. Bostrom (2012); "The Superintelligent Will: Motivation and Instrumental Rationality". https://nickbostrom.com/superintelligentwill.pdf
[68] D. Krueger et al. (2020); "Hidden Incentives for Auto-Induced Distributional Shift". arXiv:2009.09153
[69] A. Pan, et al. (2023); "Goal Misgeneralization in Deep Reinforcement Learning". arXiv:2105.14111
[70] A. Demski & S. Garrabrant (2018); "Embedded Agency". arXiv:1902.09469
[71] J. Ji, et al. (2023); “Survey of Hallucination in Natural Language Generation”. arXiv:2202.03629
[72] J. Maynez et al. (2020); "On Faithfulness and Factuality in Abstractive Summarization". https://aclanthology.org/2020.acl-main.173/
[73] M. Turpin, et al. (2023); “Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting”. arXiv:2305.04388
[74] N. Shinn et al. (2023); "Reflexion: Language Agents with Verbal Reinforcement Learning”. arXiv:2303.11366
[75] E. Perez, et al. (2022); "Discovering Language Model Behaviors with Model-Written Evaluations”. arXiv:2212.09251
[76] E. Yudkowsky; "The AI Alignment: Why It’s Hard and Where to Start". https://intelligence.org/files/AlignmentHardStart.pdf
[77] J. Leike, et al. (2018); "Scalable Agent Alignment via Reward Modeling". arXiv:1811.07871
[78] J. Burrell (2016); "How the machine "thinks". Big Data & Society". https://www.scirp.org/reference/referencespapers?referenceid=2624207
[79] F. Doshi-Velez, B. Kim (2017); "Towards A Rigorous Science of Interpretable ML". arXiv:1702.08608
[80] B. D. Mittelstadt et al. (2016); "The Ethics of Algorithms". https://journals.sagepub.com/doi/10.1177/2053951716679679
[81] D. Danks & A. J. London (2017); “Algorithmic Bias in Autonomous Systems”. https://www.cmu.edu/dietrich/philosophy/docs/london/IJCAI17-AlgorithmicBias-Distrib.pdf
[82] J. Leike (2018); "AI Safety Gridworlds”. arXiv:1711.09883
[83] Open AI (2016); "Faulty Reward Functions in the Wild”. https://openai.com/index/faulty-reward-functions/
[84] A. Bondarenko, et al. (2020); "Demonstrating specification gaming in reasoning models”. arXiv:2502.13295
[85] L. Langosco, et al. (2022); "Goal Misgeneralization in Deep RL”. arXiv:2105.14111
[86] T. Everitt, et al. (2017); “Reinforcement Learning with a Corrupted Reward Channel”. arXiv:1705.08417
[87] D. Hadfield-Menell, et al. (2016); “The Off-Switch Game”. arXiv:1611.08219
[88] Z. Susskind, et al. (2021); "Neuro-Symbolic AI: An Emerging Class of AI Workloads". arXiv:2109.06133
[89] D. Hendrycks, et al. (2023); "Aligning AI With Shared Human Values". arXiv:2008.02275
[90] Szegedy et al. (2013); "Intriguing properties of neural networks". https://arxiv.org/abs/1312.6199
[91] A. Athalye, et al. (2018); "Obfuscated Gradients Give a False Sense of Security". arXiv:1802.00420
[92] J. S. Park, et al. (2023); "Generative Agents: Interactive Simulacra of Human Behavior". arXiv:2304.03442
[93] OpenAI (2023); "GPT-4 Technical Report". arXiv:2303.08774
[94] I. D. Raji, et al. (2020): “Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing”. arXiv:/2001.00973
[95] y. Liu, et al. (2024); "Formalizing and Benchmarking Prompt Injection Attacks and Defenses". arXiv:2310.12815
[96] I. J. Goodfellow, et al. (2014); "Explaining and Harnessing Adversarial Examples". arXiv:1412.6572
[97] Y. Ovadia, et al. (2019); "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift" arXiv:1906.02530
[98] D. Hendrycks & K. Gimpel (2017); "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks". arXiv:1610.02136
[99] J. Yang, et al. (2021); "Generalized Out-of-Distribution Detection: A Survey". arXiv:2110.11334
[100] C. Guo, et al. (2017); "On Calibration of Modern Neural Networks". arXiv:1706.04599
[101] M. Abdar, et al. (2021); "A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and

Challenges".
https://ieeexplore.ieee.org/abstract/document/6796673
[102] J. Gawlikowski, et al. (2021); "A Survey of Uncertainty in Deep Neural Networks", in Springer.
https://link.springer.com/article/10.1007/s10462-023-10562-9
[103] F. Doshi-Velez & B. Kim (2017): “Towards A Rigorous Science of Interpretable Machine Learning”. arXiv:1702.08608
[104] L. Bryant (2024); “The Modern AI Stack: Design Principles for the Future of Enterprise AI Architectures “, in Menlo Ventures Web page. https://postshift.com/the-modern-ai-stack-design-principles-for-the-future-of-enterprise-ai-architectures-menlo-ventures/
[105] Thilo Hagendorff (2021/2022); “Blind spots in AI ethics”, in Springer Link.
https://link.springer.com/article/10.1007/s43681-021-00122-8
[106] IEEE Transmitter (2024); “How to Use AI, Responsibly”.
https://transmitter.ieee.org/how-to-use-ai-responsibly/
https://transmitter.ieee.org/how-to-use-ai-responsibly/
[107] Thilo Hagendorff (2024); “Mapping the Ethics of Generative AI”. arXiv:2402.08323
[108] A. Mishra (2023）; “Alignment and Social Choice: Fundamental Limitations and Policy Implications”. arXiv:2310.16048
[109] R. West & R. Aydin (2024); “The AI Alignment Paradox”. arXiv:2405.20806
[110] R. Millière (2025); “Normative Conflicts and Shallow AI Alignment”. arXiv:2506.04679
[111] J. Ji et al. (2023); "AI Alignment: A Comprehensive Survey". arXiv:2310.19852
[112] M. H. Yeh, et al. (2024–2025); “Challenges and Future Directions of Data-Centric AI Alignment". arXiv:2410.01957
[113] I. M. Cockburn, et al. (2018); “The Impact of Artificial Intelligence on Innovation”.
http://www.nber.org/papers/w24449.pdf
[114] Santoro, et al. (2021); "Symbolic Behaviour in Artificial Intelligence". arXiv:2102.03406
[115] B. Liang, et al. (2025); "AI Reasoning in Deep Learning Era: From Symbolic AI to Neural–Symbolic AI".
https://www.mdpi.com/2227-7390/13/11/1707
[116] N. Spivack, et al. (2024); “Cognition is All You Need”. arXiv:2403.02164
[117] Lindström et al. (2024); “AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations”. arXiv:2406.18346