AI原則としての包括的説明責任：技術と倫理の協調に向けて

Explicability as an AI Principle: Technology and Ethics in Cooperation

上浦　基 *1
Moto Kamiura

*1同志社大学高等研究教育院
Institute for Advanced Research and Education, Doshisha University

This paper categorizes current approaches to AI ethics into four perspectives and briefly summarizes them: (1) Case studies and technical trend surveys, (2) AI governance, (3) Technologies for AI alignment, (4) Philosophy. In the second half, we focus on the fourth perspective, the philosophical approach, within the context of applied ethics. In particular, the explicability of AI may be an area in which scientists, engineers, and AI developers are expected to engage more actively relative to other ethical issues in AI. We propose four fundamental elements to improve AI intelligibility and interpretability: "I/O," "Constraints," "Objectives," and "Architecture." Furthermore, we discuss how the relationship between AI designers' objectives and users' purposes is fundamentally connected to the challenges of AI alignment.

1.　はじめに

広島 AI プロセス合意 [G7 2023] や EU の AI 規制法成立 [EU 2024] など、AI の開発と利用に関する社会的合意形成が急速に進められている中、AI 技術開発と AI 倫理的施策を協調させそれらを共に発展させることが重要である。

本稿ではまず、「AI 倫理課題に対しどのような方策で解決への道筋をつくるのか」というアプローチの現状を 4 つに大別して概観する。こうした観点は、AI 倫理の諸課題が、先端技術研究、グローバルな社会的・経済的影響、人間や道徳に関する観念の変容等が複合的に関連したものとして認識され、AI 戦略の文脈に適切に位置づけられる必要があり、また多分野の専門家の協力と継続的関与がその解決のために望まれる、ということに光を当てるものである。

次に、そうした課題解決の取組みのひとつとして、中核的 AI 原則 [Floridi et al. 2018] に含まれる Explicability に関する応用倫理学的アプローチについて述べる。本稿では、関連研究を整理した上で、Explicability を「包括的説明責任」と訳し、下位概念として区別されて用いられる用語である Explainability を「説明可能性」と訳し分けることを提案する。また、多様なステークホルダーの効果的な協働を促進するためには、様々な AI システムの構造の理解を共有するという課題を克服することが必要であるが、そのための適切な抽象化水準 [Floridi 2019a] とそこで必要となる基本的構成要素に関する我々の研究について概説する。

2.　AI 倫理課題への諸アプローチ

AI 倫理に関わる研究や実践には異なる方法論に基づく複数のアプローチがある。誤情報・偽情報・バイアス・透明性・責任・セキュリティ等、「何が課題なのか」に関しては本節で挙げる文献に多くの記載があるが、「AI 倫理課題に対しどのような方策で解決への道筋をつくるのか」という戦略の全体像に関する見通しの良い記述は少ない。本稿では後者について、(1) 事例研究・技術動向調査、(2) ガバナンス、(3) 技術的対処、(4) 哲学 の 4 つに大別して現状を概観する。

連絡先: 上浦 基，同志社大学高等研究教育院，京都府京田辺市多々羅都谷 1-3，mkamiura@mail.doshisha.ac.jp

2.1　事例研究・技術動向調査

統計的機械学習システムの実環境稼働が本格化した 2010 年代 [上浦 2015] 以降、AI チャットボットの暴走や SNS パーソナルデータの選挙への利用 [福岡 2022]、自動運転車の死亡事故等、AI 倫理に関わる事案が発生している。実社会での技術使用に伴って発生する課題に対処する際、こうした事例研究の蓄積と共有は重要である。また、実際に事故が起きていない場合でも、最新の技術動向が把握され、その潜在的なリスクについて認識が共有される必要がある。その際「AI は人間や（実）環境でベータテストされるべきではない」[Floridi 2019b] という理念が尊重されるべきである。一方、直近では高性能小型モデル [DeepSeek-AI 2025] が公開される等、ソフトウェアに関しては一般市場への急速な展開が続いており、安全面の不確実性への対処は課題である。規制とイノベーションのバランスについては、国際協調と広範なステークホルダーの合意が重要であり、この観点は次節で述べる AI 原則の策定に結びつく。

2.2　ガバナンス

複数の AI 倫理に関わる事案が周知となるに従い、ソフトロー（拘束力を伴わない倫理原則やガイドライン等）やハードロー（拘束力を伴う法的規制）を適切に組み合わせる必要性が認識された。2010 年代後半以降、AI 原則（[Asilomar 2017] 等）策定やガイドライン整備が各国で進み、我が国ではアジャイル・ガバナンス [経産省 2022a] の考え方に基づくガイドライン [経産省 2022b; 2024] が提示された。国際的にも、広島 AI プロセス [G7 2023]、EU の AI 規制法成立 [EU 2024] 等、ハイレベルの合意形成へと至っている。一方、こうした急速な実践的対応とともに、法実証主義や自然法論等の法理論の観点からの研究 [Dahraj 2023; Caputo 2024] や、「結果重視の考え方 (outcome-focused view)」では見落とされる可能性がある、ガバナンスの重要な特性のひとつとしての手続き的側面と結びついた政治的正統性を理論化する枠組みの開発 [Erman&Furendal 2024] も重要であろう。

また、AI の開発・利用が国家安全保障や外交政策に重大な影響を及ぼし得るとの認識 [JETRO 2024]、および、AI ガバナンス研究が「政治、経済、軍事、統治、倫理の次元に焦点を当てる」[Dafoe 2018] という定義に従うと、レアアース、半導体、データセンター、電力、海底ケーブル、衛星、ロボティク

ス等に関する「デジタル言説のハードウェア・ターン」[Floridi 2024a] は、この領域における倫理的課題として整理され得る。

2.3 技術的対処

前の二節は、既に存在するまたは今後実現し得る AI システムとその影響を所与とし、社会的あるいは法的水準でその対処方法を構想するものであった。これに対して、AI の開発段階で倫理的課題を認識し技術的水準で対処することにより、望ましい情報環境を実現しようとするアプローチがある。後者に関する主な観点として、次の 4 つを挙げることができる。(i) 仕様・標準の策定：路上走行車の自動化レベル [SAE 2021] や医療用ロボットの自律性レベル [Yang et al. 2017]、IEEE や ISO/IEC における国際標準 [IEEE 2017; ISO/IEC 2023] 等、技術的な仕様を策定し共有すること。(ii) 倫理的メカニズムの実装：人間の意図や価値観に沿った働きをする Aligned AI [Ji et al. 2023] や、AI の信頼性や透明性を向上させるための説明可能 AI（XAI）[Ali et al. 2023] 等のように、倫理的困難を引き起こすリスクを低減するためのアルゴリズム等をシステムに実装すること。(iii)AI システムを用いた倫理課題解決：マルウェア検出 [Singh&Singh 2021] やフェイクニュース検出 [Shen et al. 2023] 等、サイバーセキュリティや情報倫理に関する諸課題を解決するための AI システムを開発し使用すること。(iv) バイアスの特定と緩和：データ、アルゴリズム、ユーザー等に由来するバイアス [Ferrara 2024] の具体的な原因と影響を明らかにし、その緩和戦略に取り組むこと。

2.4 哲学

学術研究分野としての倫理学は哲学の一部として位置づけられるが、AI 倫理については、特に、応用倫理学や技術哲学 [Coeckelbergh 2020a] の領域で、概念的フレームワークや道徳的推論の整備が進んでおり、またこれらの研究者による欧州委員会の高度専門家グループ (HLEG) 等を通じた実践的貢献も為されている [Floridi 2019; Coeckelbergh 2020b]。技術哲学では、技術を用いた人間の強化を肯定するトランスヒューマニズムは擁護しなくとも、技術と人間の関係を問い直し人間を再定義するポストヒューマニズムについては擁護する立場があり、特に近代西洋が前提とする人間の自律性や主体-客体の分割が再検討される [Verbeek 2011]。多くの AI 原則では人間中心主義的理念が基本的に維持されているものの、既にその一部には人間を主体とし技術を客体とする道具主義的な AI 観に留まらない考え方が反映されていると見ることもできる（[内閣府 2019]4.1(7)「AI の発展によって、人も併せて進化していくような継続的なイノベーションを目指すため」等）。また実際的にも、AI の設計・製造・普及によって「技術的志向性」が実現され、人間の道徳的主体構成に介入していると考えるならば、上述した事例研究や技術的対処は、Verbeek が提示する「技術的媒介」の具体例となっているとも言える。他方、社会的水準での包括的戦略により、AI の新しく強力な説得の形態の悪影響を最小化し、個人の自律性を保護するべきであるとする立場 [Floridi 2024b] もある。これらの哲学的議論が広く参照され認識が共有されることは、社会的に合意できる道徳的判断と科学技術政策の目標設定との首尾一貫性を保つために有益であろう。応用倫理学的アプローチについては、次節以降の本稿の内容として具体的に展開する。

3. AI 倫理原則としての包括的説明責任

3.1 中核的 AI 原則と Explicability

AI に関する倫理指針の重要性は既に広く認識されており、またその策定は歓迎すべきことである。しかし、提案された数多くの AI 原則には、冗長性、混乱、曖昧さ等による「原則の拡散」と呼ばれる問題がある [Floridi&Cowls 2019]。この問題に対処するため、Floridi らは、Asilomar AI 原則等、主だった 6 つの原則を分析し、それらに含まれる 47 項目を統合して、5 項目から成る中核的 AI 原則を提示した。彼らは、生命医学倫理の 4 原則（善行、無危害、自律性尊重、公正）[Beauchamp&Childress 1979] と AI 原則群を比較し、それらの一致性を見出すとともに、第 5 の原則として新たに Explicability を付加すべきであるとした。生物医学倫理分野では倫理原則として Explicability を付加すべきか否かについて様々な専門的な検討が行われているものの [Ursin et al. 2022; Adams 2023]、AI 原則として採用すべきか否かに依らず、Explicability の重要性については広く認められている。

3.2 包括的説明責任としての Explicability

AI の説明性や透明性に関する用語としては、Intelligibility、Interpretability、Explainability、Accountability、Transparency 等、複数あり、それぞれの用語が実際に指し示すものは異なっている。ただし、複数の研究 [Floridi et al. 2018; Morley et al. 2020; Ursin et al. 2023] によって明らかとなるのは、Explicability がそれら複数の説明性や透明性に関する包括的な上位概念であるということである。実際、Floridi らは Explicability を認識論的な意味での Intelligibility と倫理的な意味での Accountability の統合概念であるとし、また Ursin らはこれを表 1 に示すような「説明と同意（インフォームド・コンセント）」の深度と結びついた 4 つの水準についての包括的な用語として提示した。

以上を踏まえ、本稿では、Explicability を「包括的説明責任」と訳し、下位概念として区別されて用いられる用語である Explainability に「説明可能性」の語を割り当てることで訳し分けることを提案する。これらは特に日本語への直訳では見分けにくい用語であるが、中核的 AI 原則のひとつとしても認識される Explicability については、それを用いて指し示される意味の水準も含めて訳出することが適当であろう。

表 1: 包括的説明責任の 4 水準（cf.[Ursin et al. 2023]）

「説明と同意」のための倫理的要請	指針となる倫理的問い
1. 開示 (Disclosure)	何らかの AI システムが使われているか？
2. 明瞭性 (Intelligibility)	一般的に AI システムとはどのように働くのか？
3. 解釈可能性 (Interpretability)	特定の AI システムがどのように働くか？
4. 説明可能性 (Explainability)	その AI システムはなぜそのような決定を下したのか？

4. 明瞭性・解釈可能性への科学技術の貢献

4.1 明瞭性・解釈可能性における課題

包括的説明責任を、開示・明瞭性・解釈可能性・説明可能性の 4 水準から構成する Ursin らの考え方は、それを認識論的水準と倫理的水準の統合と見なす Floridi らの考え方と整合的である。Floridi らの認識論的水準に対応する、Ursin らの明瞭性および解釈可能性は、科学者や技術者による重要な貢献が為されるべき AI 倫理課題である。ただし、ここでの主な課題は、単にソフトウェアコードを開示することではなく人々に対して意思決定の説明をすること [Coeckelbergh 2020] であり、その際、どの程度の深さで詳細を説明すれば満足できるかは文

脈や受け手に大きく依存する [Herzog 2022]。そして、その難しさの一部は、適切な「抽象化水準」を設定すること [Floridi 2019c] であると考えられている。すなわち、明瞭性および解釈可能性に関して重要なのは、よりよい包括的説明責任を構築するための基盤の提供であり、そのための適切な抽象化水準の明確化である。

4.2 明瞭性・解釈可能性のための 4 つの基本要素

Ursin らは、明瞭性および解釈可能性に応答するための指針となる倫理的問い（表 1 を見よ）に答える際に用いられる用語として「入力、出力、学習データ、パラメータ、計算」を挙げた。これに対して、我々は、これらに応答するための 4 つの基本要素として「入出力 (Input/Output)、制約 (Constraints)、目的 (Objectives)、アーキテクチャ(Architecture)」を提案し、それらの間に見出される一般的な関係についてを数理構造に基づいて説明するとともに、目的概念の位置付けを明確化した [Kamiura 2025]。

表 2: 明瞭性および解釈可能性に応答するための用語

Kamiura 2025	Ursin et al. 2023
I/O (Input/Output)	input, output, training data
Constraints	parameter
Architecture	calculation
Objectives	(Nothing)

I/O と制約の関係は数学的関数における変数とパラメータの adjunction 構造から抽象化され、Objective は目的関数の役割から抽象化される。このような視点は観測者と観測行為を含むシステムをモデル化し分析するための手法としての dynamic duality の概念を受け継いで構築されており、複雑系科学が進展した 1990 年代から機械学習が普及した 2010 年代への過渡期に開発された [Kamiura and Gunji 2006; Kamiura 2013]。

4.3 AI システムの Objective と Purpose

AI システムの「目的」には、システムの最適化関数として埋め込まれる設計者の目的としての Objective と、システム利用者の目的としての Purpose という、異なる二つの意味を見出すことができる。包括的説明責任の水準が、「どのように」を問う明瞭性と解釈可能性から、「なぜ」を問う説明可能性に移行するとき、Objective および Purpose の情報が要求される。ここでの Objective は、狭義には、データ入出力関数としての AI システムの性質を決定する技術的要素としての目的関数であり、広義には、その目的関数の運用の結果としての AI システムの予期される振舞いと結びついたシステムの開発目的である。他方、Purpose は、利用者のニーズや期待、システムを取り巻く外部環境や社会的背景に関係し、利用者や社会が事後的にその価値を見出し、最終的にその AI システムを制約するものである。

AI 以外の従来の工業製品では、Objective と Purpose は通常一致している。製品の適切なマニュアルが提供され、設計者の意図がユーザーに伝えられるかもしれない。また、製品に欠陥があった場合の責任は設計者や製造者にあるが、欠陥のない製品を設計者の意図を超えて不適切に使用した場合の責任は使用者にある。自動車やガソリンは様々な用途に使われる潜在的な汎用性を持っている。Objective と Purpose の不一致は、使用者の積極的な創造性と予期せぬ危険性の両方に開かれている。Objective と Purpose を一致させることで、安全に使用することができる。このような整合性は、確立された法律や我々の常識によって支えられており、また歴史的な発展の結果でもある。

AI の場合、使用者も開発者も、AI に何ができるのか、何をさせるべきか、試行錯誤の段階にある。トップランナーである OpenAI でさえ、汎用人工知能（AGI）の最終目標を明確に把握していないようである [Altman 2023]。AI は、様々なユーザーの絶えず変化する多様な目的に対応できるという肯定的側面を持っている。しかし、この「一般性」が、Objective と Purpose のズレあるいは Purpose の消滅を意味する場合、AI アライメントに関連した倫理的問題を引き起こすリスクが高まる。こうしたリスクを既存の工業製品と同様に適切に管理するためには、さらなる研究と社会的コンセンサスが必要である。

5. 結論

本稿では、AI 倫理課題への諸アプローチを概観した上で、特に応用倫理学的アプローチによる AI 原則の分析について述べた。明瞭性と解釈可能性は、AI に関する「どのように」という問いに答える説明の水準であり、包括的説明責任の一部を成す。明瞭性と解釈可能性を実践的に構築するためには、AI のプログラムコードを開示するのではなく、AI の構造を一般的に説明するための適切な抽象化水準を確立する必要がある。我々は、いくつかの要件を満たしながら明瞭性および解釈可能性に応答するための 4 つの基本要素を提案するとともに、AI システムの Objective と Purpose の整合性が AI アライメント・リスクを管理する上で重要であることを指摘した。

参考文献

[Adams 2023] Adams, J.: Defending explicability as a principle for the ethics of artificial intelligence in medicine. Medicine. *Health Care and Philosophy* 26, 615-623. (2023).

[Ali et al. 2023] Ali, S., et al.: Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805 (2023).

[Altman 2023] Altman, S.: Planning for AGI and beyond. February 24, 2023. `https://openai.com/blog/planning-for-agi-and-beyond` (2023).

[Asilomar 2017] Asilomar AI Principles (2017). `https://futureoflife.org/open-letter/ai-principles/`

[Beauchamp&Childress 1979] Beauchamp, T.L. and Childress, J.F.: it Principles of biomedical ethics. Oxford University Press (1979).

[Caputo 2024] Caputo, N. A.: Rules, Cases, and Reasoning: Positivist Legal Theory as a Framework for Pluralistic AI Alignment. arXiv:2410.17271v3 (2024).

[Coeckelbergh 2020a] Coeckelbergh, M.: *Introduction to Philosophy of Technology*, Oxford University Press (2020), (邦訳：『技術哲学講義』直江清隆・久木田水生監訳, 丸善出版 (2022)).

[Coeckelbergh 2020b] Coeckelbergh, M.: *AI Ethics*, MIT Press (2020), (邦訳：『AI の倫理学』直江清隆 他 訳, 丸善出版 (2020)).

[Dafoe 2018] Dafoe, A.: AI Governance: A Research Agenda. `www.fhi.ox.ac.uk/govaiagenda` (2018).

[Dahraj 2023] Dahraj, A. T.: Theory of Natural Law, Legal Positivism and Its Implications for AI Regulation. SSRN 4357131 (2023).

[DeepSeek-AI 2025] DeepSeek-AI: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948v1 (2025).

[Erman&Furendal 2024] Erman, E. and Furendal, M.: Artificial Intelligence and the Political Legitimacy of Global Governance. *Political Studies* 72(2) 421-441 (2024).

[EU 2024] EU AI Act (2024). https://artificialintelligenceact.eu/

[Ferrara 2024] Ferrara, E.: Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* 6(1), 3, (2024).

[Floridi 2019a] Floridi, L.: *The Logic of Information*. Oxford University Press (2019). (邦訳：『情報の論理学』上浦基・近藤和敬・中嶋浩平 訳, 東京大学出版会 (近刊)).

[Floridi 2019b] Floridi, L.: Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1, 261-262 (2019).

[Floridi 2019c] Floridi, L.: What the Near Future of Artificial Intelligence Could Be. *Philosophy Technology* 32, 1-15 (2019).

[Floridi 2024a] Floridi, L.: The Hardware Turn in the Digital Discourse: An Analysis, Explanation, and Potential Risk. *Philosophy Technology* 37, 39 (2024).

[Floridi 2024b] Floridi, L.: Hypersuasion - On AI's Persuasive Power and How to Deal with It. *Philosophy Technology* 37, 64 (2024).

[Floridi&Cowls 2019] Floridi, L. and Cowls, J.: A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* 1(1) (2019).

[Floridi et al. 2018] Floridi, L. et al.: AI4People - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28, 689-707 (2018).

[G7 2023] G7: Hiroshima AI Process (2023). https://www.soumu.go.jp/hiroshimaaiprocess/

[Herzog 2022] Herzog, C.: On the risk of confusing interpretability with explicability. *AI and Ethics* 2, 219-225. (2022).

[IEEE 2017] IEEE Initiative on Ethics of Autonomous and Intelligent Systems: Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. (2017).

[ISO/IEC 2023] ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system (2023). https://www.iso.org/standard/81230.html

[JETRO 2024] JETRO 日本貿易振興機構: ビジネス短信（2024 年 10 月 25 日付）https://www.jetro.go.jp/biznews/2024/10/ce66709a5f931a80.html

[Ji et al. 2023] Ji, J., et al.: AI Alignment: A Comprehensive Survey. arXiv:2310.19852v5 (2023).

[Kamiura 2013] Kamiura, M.: Implicit Interaction: Mathematics on Local Description of Systems. *Transactions of the Society of Instrument and Control Engineers* 49(1), 190-196 (2013).

[Kamiura 2025] Kamiura, M.: The Four Fundamental Components for Intelligibility and Interpretability in AI Ethics. *The American Philosophical Quarterly*, Special Issue on "The Ethics of AI", 62(2)103-112 (2025).

[Kamiura&Gunji 2006] Kamiura, M. and Gunji, Y.-P.: Robust and Ubiquitous On-Off Intermittency in Active Coupling. *Physica D* 218, 122-130 (2006).

[Morley et al. 2020] Morley, J., et al.: From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Science and Engineering Ethics 26, 2141-2168 (2020).

[SAE 2021] SAE Standards: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles, J3016_202104. (2021).

[Shen et al. 2023] Shen, Y., et al.: Fake News Detection on Social Networks: A Survey. *Applied Science* 13(21)11877 (2023).

[Singh&Singh 2021] Singh, J. and Singh, J.: A survey on machine learning-based malware detection in executable files *Journal of Systems Architecture* 112, 101861 (2021).

[Ursin et al. 2022] Ursin, F., et al.: Explicability of artificial intelligence in radiology: Is a fifth bioethical principle conceptually necessary? *Bioethics* 36(2)143-153. (2022).

[Ursin et al. 2023] Ursin, F., et al.: Levels of explicability for medical artificial intelligence: What do we normatively need and what can we technically reach? Ethik in der Medizin 35, 173-199 (2023).

[Verbeek 2011] Verbeek, P.-P. : *Moralizing Technology*, The University of Chicago Press (2011), (邦訳：『技術の道徳化』鈴木俊洋 訳, 法政大学出版局 (2015)).

[Yang et al. 2017] Yang, G.-Z., et al.: Medical robotics - Regulatory, ethical, and legal considerations for increasing levels of autonomy. *Science Robotics*, 2, eaam8638 (2017).

[上浦 2015] 上浦基：〈システム〉思考と特異点を待たない人格のアップロード『現代思想』2015 年 12 月号 (2015).

[経産省 2022a] アジャイル・ガバナンスの概要と現状 (2022). https://www.meti.go.jp/press/2022/08/20220808001/20220808001.html

[経産省 2022b] AI 原則実践のためのガバナンス・ガイドライン Ver. 1.1 (2022). https://www.meti.go.jp/press/2021/01/20220125001/20220124003.html

[経産省 2024] AI 事業者ガイドライン (2024). https://www.meti.go.jp/press/2024/04/20240419004/20240419004.html

[内閣府 2019] 内閣府・統合イノベーション戦略推進会議『人間中心の AI 社会原則』(2019).

[福岡 2022] 福岡真之介『AI・データ倫理の教科書』弘文堂（2022）.