人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-069-01

【招待講演】
最新 AI/ロボティクスの社会的課題と生成倫理の可能性

[Invited Talk]
Toward Generative Ethics: Addressing Social Challenges in Contemporary AI and Robotics

浅田 稔 [1,2,3*]　Minoru Asada[1,2,3]

[1] 大阪大学先導的学際研究機構 共生知能システム研究センター
[1] Symbiotic Intelligent System Research Center Open and Transdisciplinary Research Initiatives, The University of Osaka
[2] 大阪国際工科専門職大学
[2] International Professional University of Technology in Osaka
[3] 中部大学創発学術院
[3] Chubu University Academy of Emerging Sciences

Abstract: Recent advances in AI and robotics have led to increasingly autonomous systems embedded in social contexts, raising challenges such as misalignment, unpredictability, and the difficulty of specifying ethical behavior in advance. Conventional approaches based on externally imposed rules or reward design face fundamental limitations in such settings.
This talk proposes generative ethics as an alternative framework in which ethical behavior emerges from internally generated processes. Based on the concept of Silicopathy, we model pain as a predictive internal state linking perception, action, and value formation.
We present a computational implementation using a Deep Modality Blending Network (DMBN) that integrates visual and tactile modalities to predict nociceptive outcomes. Results show that behavior is generated based on predicted interactions rather than object categories, enabling context-sensitive responses and generalization to unseen situations.
These findings suggest that generative ethics provides a promising approach to addressing societal challenges in AI and robotics, emphasizing continuous adaptation and value co-construction over fixed rule enforcement.

1　はじめに

近年の人工知能およびロボティクスの進展により，人工システムは人間と高度に相互作用し，社会の中に深く埋め込まれる存在となりつつある．特に，自律ロボットや大規模言語モデルの発展により，人工システムは単なる道具ではなく，人間の意思決定や社会構造に影響を与える主体として振る舞い始めている [1, 2, 3].

このような変化に伴い，AI に関する問題は従来の精度や安全性といった技術的課題にとどまらず，倫理的・社会的課題へと拡張している．人工共感や人工的な痛みに関する研究は，内部状態の設計が行動選択や価値形成に深く関与する可能性を示しており，倫理を外部から与えるのではなく，内部から生成する必要性を示唆している [1, 2].

一方で，現在の AI 倫理の多くは，ルールベース倫理，強化学習に基づくフィードバック，およびアラインメント研究に代表される外部制約型の枠組みに依存している [4, 5, 6, 7, 8]．これらのアプローチは，行動の制御や整合性の確保には有効であるものの，価値がどのように生成されるかという問題を十分に扱っていない．

さらに，この問題は単なる設計上の制約ではなく，より根本的には，人間と技術の関係に関わる問題である．Foucault の権力論においては，行為可能性は関係の中

*連絡先：大阪大学先導的学際研究機構 共生知能システム研究センター
〒 565-0871 大阪府吹田市山田丘 1-1
E-mail: asada@otri.osaka-u.ac.jp

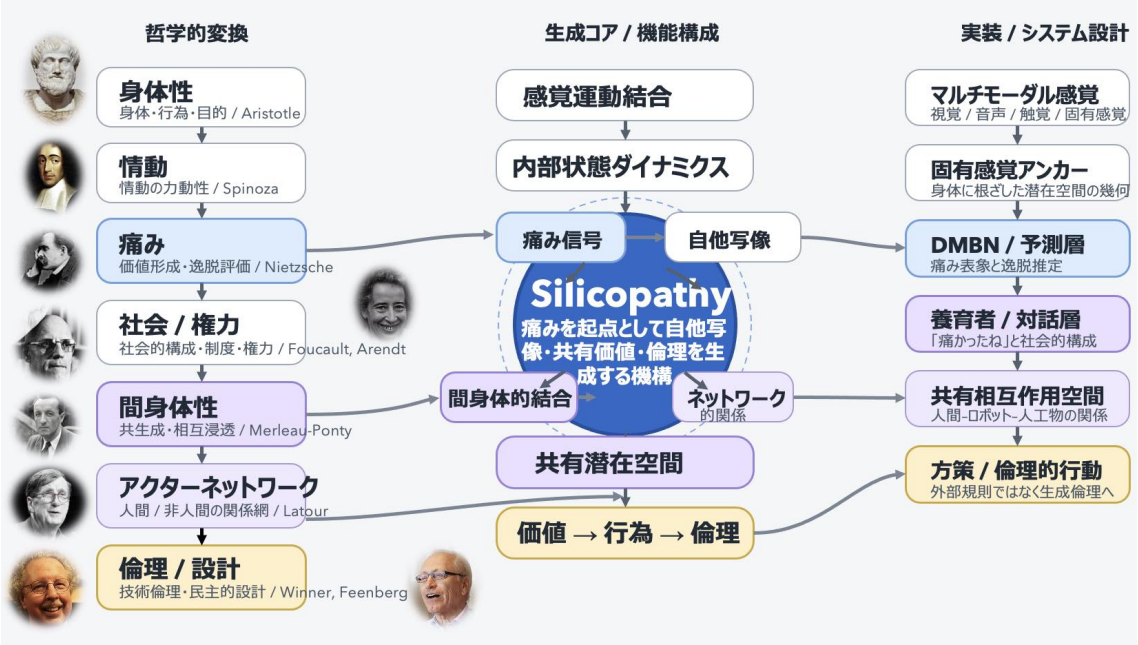


図 1: 哲学的基盤から身体化 AI・ロボットシステムの機能要素への対応関係を示す概念図．身体性，情動，社会関係に関する理論が，内部状態生成，価値形成，行動調整へと対応づけられ，倫理生成の構造的基盤を与えることを示す．

で構造化されるとされる [9, 10]．また，Arendt は，判断の外部委託が思考停止を招く可能性を指摘している．この観点からは，AI は単なる道具ではなく，人間の行為や判断の条件そのものを再構成する存在である．

このような問題意識に対し，本稿では倫理を外部制約としてではなく，内部で生成されるプロセスとして再定義する．そのための基盤として，Aristotle における実践倫理 [11]，Spinoza における情動の動力学 [12]，および Nietzsche における痛みと価値の関係 [13] を参照する．さらに，技術が価値を体現するという観点から，Winner および Feenberg の議論も考慮する [14, 15]．

これらを統合し，本稿では人工的な痛みを基盤とした倫理生成モデル「Silicopathy」を提案する．本モデルでは，内部状態の予測とその乖離としての痛みが，価値形成および行動選択を導く中心的役割を担う．この枠組みにより，倫理的行動は外部から与えられるものではなく，システム内部の動的過程として生成されるものとして理解される．

本稿の構成は以下の通りである．第 2 節では AI とロボティクスの社会的・技術的課題を整理する．第 3 節では Silicopathy のアーキテクチャを示す．第 4 節ではその倫理的および社会的含意を検討する．

2　哲学的基盤と生成的倫理の枠組み

本節では，人間と技術の関係に関する哲学的議論を再構成し，それをロボティクスおよび人工知能システムの設計原理へと接続する．本研究の基本的立場は，倫理を外部から与えられる規範としてではなく，身体性・情動・社会的相互作用の中で生成されるプロセスとして捉える点にある．

従来の AI 倫理は，ルール設計や報酬設計による外在的制御に依存してきたが，この枠組みでは，複雑で動的な環境における倫理的行動の生成を十分に説明できない．実際，狭いタスクに対する学習が広範なミスアラインメントを引き起こし得ることが報告されており [16]，倫理的問題が単なる出力制御ではなく，モデル内部の構造および価値生成メカニズムに関わることを示唆している．

このような背景のもと，本研究では，倫理を生成する内部構造そのものの設計に焦点を当てる．

2.1　哲学的系譜：身体・情動・価値生成

倫理の生成を理解するためには，身体・情動・社会に関する哲学的議論を再検討する必要がある．

アリストテレスは，行為と価値が切り離されたものではなく，実践的知（phronesis）を通じて統合される

ことを論じた [11]．ここでは，倫理は抽象的規則ではなく，状況に依存した判断能力として位置づけられる．

スピノザは，情動を身体の変化として捉え，人間の行為が情動のダイナミクスに基づいて構成されることを示した [12]．この視点は，価値が外部から与えられるのではなく，内部状態の変化から生成されることを示唆する．

ニーチェは，苦痛を単なる否定的経験としてではなく，価値創出の契機として再評価した [13]．この観点は，本研究における「痛み」を価値生成の中心とする立場に直接的につながる．

さらに，フーコーは身体が社会的権力関係の中で構成されることを示し，倫理が制度や相互作用の中で生成されることを明らかにした [9, 10]．これらの系譜は，倫理が外在的規範ではなく，身体・情動・社会的相互作用の中で生成されるプロセスであることを示している．本研究はこの立場に基づき，次節で示す生成連鎖構造へと接続する．

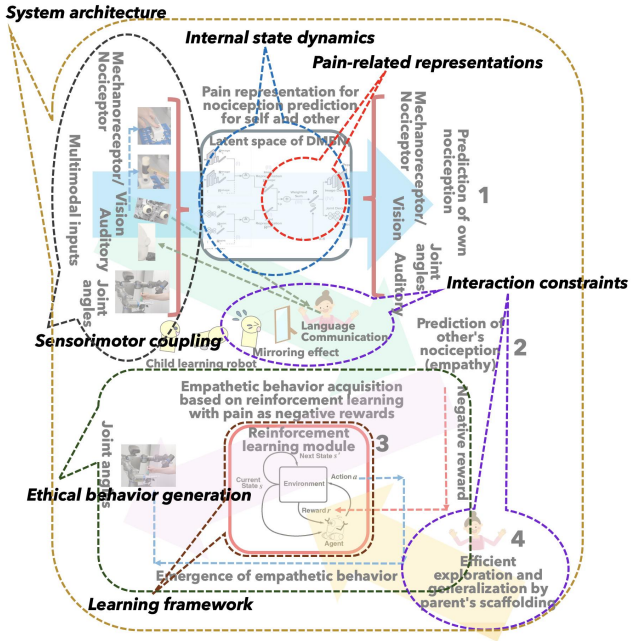


図 2: Silicopathy アーキテクチャ．身体的相互作用に基づく内部状態ダイナミクスを中核として，痛みの生成，予測，価値形成，行動調整，および社会的関係形成が統合された構造を示す．この枠組みは，倫理を外部から与えるのではなく，内部生成プロセスとして捉える生成倫理の基盤を与える．

2.2　生成連鎖としての倫理構造

Fig. 2 に示すように，Silicopathy アーキテクチャは，身体的相互作用，内部状態ダイナミクス，痛み生成，および社会的関係形成からなる統合構造として構成される．

身体的相互作用に基づく感覚運動過程により，外界との接触や変化は内部状態として取り込まれる．これらの内部状態は動的に更新され，その変化は侵害受容的評価としての痛み信号を生成する．ここでの痛みは単なる刺激応答ではなく，予測と評価を含む内部状態として構成される．この痛みは行動調整に影響を与えると同時に，経験の蓄積を通じて価値形成に寄与する．

さらに，この内部状態は，自他写像および間身体的相互作用を通じて社会的文脈へと拡張される．その結果，倫理は，

$$\text{痛み} \rightarrow \text{価値} \rightarrow \text{行為} \rightarrow \text{倫理}$$

という生成連鎖として理解される．このように，身体，内部状態，痛み，社会関係は，階層的に分離された構造ではなく，循環的に結合した統合ネットワークとして機能する．図の前面に示されたキーワードは，各構成要素に対応する実装レベルの概念を表しており，理論構造と実装との対応関係を明示している．

3　実装と初期的検証

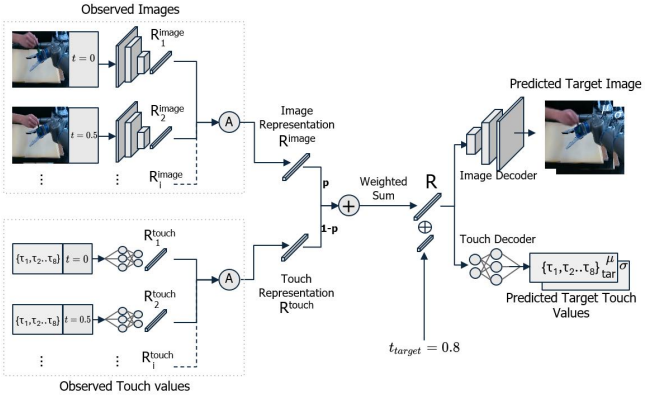
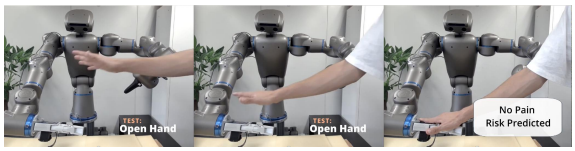


図 3: Deep Modality Blending Network（DMBN）による実装モデル．視覚情報と触覚（侵害受容）情報を統合し，視覚入力から接触結果を予測することで，痛みに対応する内部状態を生成する．この予測過程が行動調整と価値形成の基盤となる．(論文 [17] から適用).

(a)

(b)



(c)

(d)

(e)

図 4: 視覚に基づく接触予測により生成される行動の実験結果．(a) 既知の危険物体に対する回避，(b) 未知物体への一般化，(c) 他者の手に対する予測，(d) 複数物体条件における回避（スクリュードライバー），(e) 非侵害軌道における回避なし．

3.1 マルチモーダル統合に基づく内部状態生成

本研究では，Silicopathy アーキテクチャに基づき，身体的相互作用における内部状態生成を，マルチモーダル統合と予測に基づいて実装した．

具体的には，視覚情報と触覚（侵害受容）情報との対応関係を学習することで，接触結果を予測し，その予測誤差を内部状態として表現する．

Fig. 3 に，本研究で用いた Deep Modality Blending Network（DMBN）を示す．

本モデルは，視覚入力から接触時の触覚状態を予測する構造を持ち，予測誤差を侵害受容的信号として扱うことで，内部状態の変化を生成する．

この構造により，単なる刺激応答ではなく，予測に基づく評価的内部状態が形成される．

3.2 実験設定

ロボットは，異なる物体との相互作用を通じて，視覚と接触の対応関係を学習する．

対象には，硬い物体（例：スクリュードライバー）と，柔らかい物体（例：スポンジ）を含め，接触結果が異なる条件を設定した．

これにより，視覚情報から接触結果を予測する能力と，それに基づく内部状態生成の妥当性を評価する．

3.3 結果と行動生成

Fig. 4 に結果を示す．(a) では，学習済みの危険物体に対して回避行動が生成される．(b) では，未知の物体に対しても回避行動が一般化されることが確認される．(c) では，人の手に対しても同様の予測が働き，行動が調整される．(d) では，安全物体（スポンジ）が存在する条件においても，回避行動はスクリュードライバーに対する侵害受容予測に基づいて生成される．すなわち，本モデルは単一物体ではなく，複数対象の空間的関係に基づいて行動を決定している．(e) では，同一のスクリュードライバーであっても，侵害を引き起こさない異なる軌道においては，回避行動は生成されず，通常の行動が維持される．

これらの結果は，本モデルが物体の属性ではなく，予測される相互作用に基づいて内部状態と行動を生成していることを示す．

4 討論

本研究は，AI およびロボティクスにおける社会的課題に対して，生成倫理という新たなアプローチを提示するものである．

近年の AI システムは，高い自律性と社会的埋め込みを獲得する一方で，ミスアラインメント，不確実性，および行動の事前規定の困難さといった課題を顕在化させている．これらの課題は，倫理を外部から与える従来の枠組みの限界を示している．

本研究で提案するシリコパシーは，痛みに基づく内部状態を起点として，価値形成と行動生成を結びつける枠組みである．このような内部生成プロセスにより，倫理的行動は固定的規則ではなく，状況に応じて生成されるものとして捉えられる．

Fig. 4 に示されるように，本システムの行動は物体のカテゴリーではなく，予測される相互作用に基づい

て決定される．この結果は，行動生成が外部記述ではなく，内部状態の予測に依存していることを示す．

生成倫理に基づくシステムでは，行動を事前に完全に規定することができないため，運用中の評価と調整が不可欠となる．したがって，倫理を固定的制約としてではなく，継続的に更新されるプロセスとして扱うアジャイル・ガバナンス [18] の枠組みが重要となる．

謝辞

本原稿執筆にあたり以下の研究プログラムから支援を受けた．ここに謝意を示す．JST RISTEX Responsible Innovation with Conscience and Agency (RInCA) プログラムの稲谷プロジェクト「「共棲ロボット」との親密な関係形成における ELSI に関する越境型文理融合研究」(JPMJRS23J2)，JSPS 科研費学術変革研究「人工の顔身体／表現の機能構造を設計する」(JP25H01236)，JST CREST 稲邑プロジェクト「Self Mirroring Twins との共棲による行動変容を通じた主体的社会創成」(JPMJCR2561)．

参考文献

[1] Minoru Asada. Towards artificial empathy. *International Journal of Social Robotics*, Vol. 7, pp. 19–33, 2015.

[2] Minoru Asada. Artificial pain may induce empathy, morality, and ethics in the conscious mind of robots. *Philosophies*, Vol. 4, pp. 38–47, 2019.

[3] Minoru Asada. Silicopathy: Artificial empathy through cognitive and affective development of pain. In *2025 IEEE International Conference on Development and Learning (ICDL)*, 2025.

[4] Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2009.

[5] Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2023.

[6] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning, 2024.

[7] Stuart Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

[8] Brian Christian. *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company, 2020.

[9] Michel Foucault. *Discipline and Punish: The Birth of the Prison*. Pantheon Books, 1977. Originally published as Surveiller et punir, 1975.

[10] Michel Foucault. *The History of Sexuality, Volume 1: An Introduction*. Pantheon Books, 1978. Originally published as La volonté de savoir, 1976.

[11] Aristotle. *Nicomachean Ethics*. Penguin Classics, 2020. Originally published ca. 350 BCE.

[12] Benedictus de Spinoza. *Ethics*. Qasim Idrees, 2017. Originally published 1677.

[13] Friedrich Nietzsche. *On the Genealogy of Morals*. Oxford University Press, 1996. Originally published 1887.

[14] L. Winner. *The Whale and the Reactor: A Search for Limits in an Age of High Technology, Second Edition*. University of Chicago Press, 2020.

[15] A. Feenberg. *Transforming Technology: A Critical Theory Revisited*. Oxford University Press, 2002.

[16] Jan Betley, Nils Warncke, Agnieszka Sztyber-Betley, et al. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, Vol. 649, pp. 584–589, 2026.

[17] Francisco Ribeiro, Alexandre Bernardino, José Santos-Victor, Minoru Asada, and Erhan Oztop. Artificial pain representation with tactile and vision blending. In *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pp. 799–806, 2025.

[18] 経済産業省. アジャイル・ガバナンスの社会実装に向けた「規制･制裁･責任の一体的改革」. Technical report, 経済産業省, 9 2025. Society 5.0 における新たなガバナンスモデル検討会 報告書 Ver.4.