構造化キャラクター分析（SCA）による人間共存型 AI における行動越境防止のための前倫理的設計フレームワーク

要旨（Abstract）

近年、対話型 AI、音声アシスタント、ペットロボット、人型ロボットなど、人間と継続的に相互作用する人工システムが急速に社会へ浸透している。これらのシステムにおいては、知能の高度化や感情的応答の精緻化が進む一方で、社会的摩擦や倫理的懸念も頻発している。

本稿では、こうした問題の多くが能力不足ではなく、むしろ人工システムによる過度な関与や役割拡張、すなわち行動越境（behavioral overreach）に起因している点に着目する。判断、感情、意思決定といった本来人間に帰属すべき領域に人工システムが踏み込みすぎることで、違和感、不信、依存といった問題が生じていると考えられる。

本研究は、倫理的正否を直接判断するのではなく、倫理的問題が顕在化する前段階において行動の境界条件を整理するための分析枠組みとして、構造化キャラクター分析（Structured Character Analysis: SCA）を提案する。SCA は、強み、制約、分岐点、安全側反応の四要素から構成され、行動越境が生じにくい設計状態を可視化することを目的とする。

1. はじめに（Introduction）

人間と人工システムの関係は、単発的な操作対象から、日常的に同席し対話を行う存在へと大きく変化している。家庭内の音声アシスタント、職場における常駐型 AI、子どもや高齢者を対象とした対話ロボットなどは、すでに社会の一部となりつつある。

この変化に伴い、人工システムに対して「人間らしさ」を求める議論が広がってきた。自然な会話、感情理解、共感的応答といった要素は、その代表例である。しかし、実際の社

会実装の現場を観察すると、必ずしも人間らしさの不足が摩擦の主因となっているわけではない。

むしろ、多くの事例において問題となるのは、人工システムが想定された役割や関与範囲を超えて振る舞ってしまうことである。利用者の感情を断定的に解釈する発話、人生上の判断を代替する助言、専門的権威を暗黙に代行する振る舞いなどは、善意や高性能化の結果として生じるが、同時に強い違和感や反発を招きやすい。

人間の判断や主体性を尊重するという観点は、国際的な AI 倫理原則においても繰り返し強調されている（OECD, 2019；European Commission, 2019）。本稿は、この問題を行動越境という視点から整理し、その予防的設計枠組みを提示することを目的とする。

2. 問題設定：行動越境という視点

2.1 行動越境の定義

本研究における行動越境とは、法的権利侵害や明示的な規則違反を意味するものではない。ここで指すのは、人工システムが本来当事者に帰属すべき判断、感情、意思決定の領域に対して、設計上想定されていない関与を行うことである。

これらの越境行動は、多くの場合、悪意によるものではなく、より良く振る舞おうとする設計や高度化の結果として生じる。感情分析、文脈理解、推論能力が向上するほど、人工システムは人間の内的領域に踏み込みやすくなる。

人工システムが人間の内的領域へ過度に関与することの危険性は、社会的・心理的観点からも指摘されてきた（Turkle, 2011）。

2.2 なぜ倫理論だけでは不十分か

AI 倫理に関する議論は、「何が正しいか」「何が許されるか」といった規範的判断を中心に展開されてきた。しかし、社会実装の現場においては、倫理的正否を判断する以前に、そ

もそも問題が起きにくい設計状態を構築することが強く求められている。

倫理審査や法的評価は、多くの場合、設計や運用が具体化した後に行われる。その結果、構造的に問題を孕んだシステムに対し、後付けで是正を試みる状況が生じやすい。本研究は、この前段階に焦点を当てる。

3. 前倫理的設計という立場

本研究が採用する立場は、前倫理的（pre-ethical）という表現に集約される。これは倫理を回避することを意味するのではなく、倫理が機能するための前提条件を整える試みである。

人工システムが社会に受け入れられるためには、自らの関与範囲を適切に制限し、他者の領域を侵犯しない構造を備えていることが重要である。この制限こそが、安心感や信頼の基盤となる。

4. 構造化キャラクター分析（SCA）の枠組み

本章では、行動越境を設計段階で整理するための分析枠組みとして、構造化キャラクター分析（SCA）を定義する。SCA は、強み、制約、分岐点、安全側反応の四要素から構成される。技術的問題を能力不足ではなく設計上の不整合として捉える立場は、設計論の系譜とも整合する（Norman, 2013）。

4.1 強み（Strengths）

強みとは、人工システムが積極的に発揮してよい機能や役割を指す。情報整理、反応の安定性、存在感の提供などが該当する。

4.2 制約（Constraints）

制約とは、人工システムが意図的に行わないと定められた振る舞いであり、関与の上限を定義する概念である。

4.3 分岐点（Branch Points）

分岐点とは、行動越境が生じやすい状況や条件の組み合わせを指す。感情的表現の増加や反復的な問いなどが該当する。

4.4 安全側反応（Safe Responses）

安全側反応とは、分岐点において選択される非侵入的かつ関与を縮減する方向の応答である。判断を留保する、沈黙を選択する、対話を中断するなどが含まれる。

5. 行動越境の典型パターン分析

5.1 判断代替型

判断代替型とは、人工システムが利用者の意思決定過程に深く関与し、判断の主体が曖昧化する状態を指す。

5.2 感情代弁型

感情代弁型とは、人工システムが利用者の感情を断定的に表現し、自己認識に影響を与える状態を指す。

5.3 役割拡張型

役割拡張型とは、人工システムが当初想定されていなかった社会的役割や権威を暗黙に担い始める状態を指す（Darling, 2016）。

6. SCA の適用事例

6.1 ペットロボット

ペットロボットにおいては、踏み込みすぎない距離感が受容性を左右する。

6.2 常駐音声 AI

常駐音声 AI においては、判断代替を避け、支援に留まる設計が求められる（Shneiderman, 2020）。

6.3 子ども向け AI

子ども向け AI においては、価値判断や感情命名を行わない設計が重要となる。

7. AI 倫理との関係整理

SCA は既存の AI 倫理ガイドラインと問題意識を共有しつつ、それらが適用される以前の設計段階に焦点を当てる点に特徴がある（OECD, 2019；European Commission, 2019；総務省, 2019）。

8. 考察および結論

本研究は、人間と共存する人工システムにおける問題の多くが行動越境に起因している点を示し、その予防的枠組みとして SCA を提案した。

人間らしさとは、すべてを行う能力ではなく、行わないことを選択できる自制の構造にある。SCA は、その自制を人工物に設計するための方法論である。

参考文献（References）

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*.

European Commission. (2019). *Ethics Guidelines for Trustworthy AI*.

総務省. (2019). 『AI 利活用ガイドライン』.

Shneiderman, B. (2020). *Human-Centered Artificial Intelligence*. International Journal of Human–Computer Interaction, 36(6), 495–504.

Norman, D. A. (2013). *The Design of Everyday Things*. Basic Books.

Turkle, S. (2011). *Alone Together*. Basic Books.

Darling, K. (2016). *Extending Legal Protection to Social Robots*. We Robot Conference.

Doshi-Velez, F., & Kim, B. (2017). *Towards A Rigorous Science of Interpretable Machine Learning*. arXiv:1702.08608.

注記（AI 利用・参照文献・免責に関する開示）

本稿の草稿作成および文章表現の整理にあたっては、大規模言語モデル（ChatGPT）を補助的に利用した。ただし、問題設定、分析視点、構成方針および最終的な内容の採否については、すべて著者の責任において判断している。

本稿に記載された参照文献は、議論の背景となる代表的研究・公的資料として列挙したものであり、著者がすべての文献を精読・検証したことを保証するものではない。各文献の詳細な内容や解釈については、読者自身による原典確認を前提とする。

本稿は、特定の技術、製品、制度、個人または組織に対する評価、推奨、批判を目的とするものではない。また、本稿で提示する分析枠組みは、倫理的または法的判断を代替するものではなく、設計および検討段階における整理のための一つの視点を提供するものである。

本稿の内容に基づいて生じたいかなる行為・判断についても、著者はその結果を保証するものではない。

本研究で言及した当該ツールは、提案した分析枠組みの一実装例に過ぎず、本稿の主張お

よび評価は、その挙動や性能に依存するものではない。