HOKKAIDO UNIVERSITY

Title	人工知能への常識道徳の実装および反種差別的観点からの批判的評価の試み［論文内容及び審査の要旨］
Author(s)	竹下, 昌志
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第16513号
Issue Date	2025-03-25
Doc URL	https://hdl.handle.net/2115/95184
Rights(URL)	https://creativecommons.org/licenses/by/4.0/
Type	doctoral thesis
File Information	Masashi_Takeshita_review.pdf, 審査の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　氏名　竹下 昌志

審査担当者　主　査　准教授　Rzepka Rafal
副　査　特任教授　坂本 雄児
副　査　教　授　長谷山 美紀
副　査　教　授　土橋 宜典
副　査　教　授　小川 貴弘
副　査　准教授　伊藤 敏彦

学位論文題名

人工知能への常識道徳の実装および反種差別的観点からの批判的評価の試み
（Implementation of Commonsense Morality in Artificial Intelligence and Its Critical Evaluation from the Anti-speciesist Perspective）

竹下昌志氏の博士論文は、人工知能 (AI) 技術の人間社会への統合に伴う倫理的課題、とりわけ自然言語処理 (NLP) の分野における新たな研究について記述している。

竹下氏は、大規模言語モデル (LLM) をはじめとする AI システムにおける倫理の実装を通じて、差別的なテキストの生成や悪意あるステレオタイプの再生といった負の影響を軽減することを目指している。論文の前半では、日本語における道徳的価値を評価する初のデータセット「JETHICS」の開発に焦点を当てる。本コーパスは、クラウドワーカーによる品質検証を経た 13 万 4 千以上のサンプルを含み、代表的な五つの倫理学派に基づくカテゴリーに分類されている。その後、日本語および多言語対応の大規模言語モデルの各種タスクにおける評価に活用される。

次に、竹下氏は、常識的道徳 (common sense morality) と規範倫理学における諸理論を統合するアルゴリズムを提案・実装する。この手法は「期待選択価値最大化 (Maximizing Expected Choiceworthiness, MEC)」と呼ばれ、道徳的不確実性のもとで AI の意思決定アルゴリズムとして利用可能である。実験結果により、MEC は倫理的選択が困難な状況において望ましい意思決定手続きであることが示された。本手法は、複数の規範倫理理論に基づくモデルの出力を統合し、異なる倫理的視点から問題を分析する人間の倫理学者の思考プロセスを模倣しながら、より適切な道徳的判断を目指すものである。

人工倫理エージェント (Artificial Moral Agents, AMA) に関する研究の多くが理論的なものである中、竹下氏は、倫理的課題の解決にニューラルネットワークを活用したアルゴリズムを提案・評価する。本手法は、英語のテストセットにおいて 0.823 の精度を達成し、一方で単独の倫理理論 (義務論、功利主義、美徳倫理) に基づく手法の精度はそれぞれ 0.715、0.765、0.771 にとどまった。

論文の後半では、動物行動学の知見を活用し、自然言語処理 (NLP) 研究における種差別およびバイアスを分析する。特に、道徳的 AI の開発において常識的道徳に依拠することの限界を批判的に検討し、曖昧で誤った価値観を再生産する危険性や、種差別といった倫理的問題が見過ごされる可能性を指摘する。さらに、こうした課題に対処するために動物倫理の知見を導入する重要性を強調し、AI 倫理研究において種差別問題を体系的に取り扱う必要があることを新たに主張する。

本論文では、NLP における種差別を扱った二つの主要な研究を詳述する。第一の研究は、マスク付き言語モデル (Masked Language Models, MLMs) に内在する種差別的バイアスの分析である。本研究では、テンプレートベースおよびコーパスベースのアプローチを用い、統計的手法によって複数の MLM における種差別的バイアスの存在を実証する。次に、本研究の範囲を拡大し、NLP モデルのみならず、データセットや研究者における種差別的バイアスを体系的に調査する。その結果、種差別的バイアスが広範に存在しながらも、研究者によってほとんど認識されていないことが明らかとなった。これらの知見は、現在のデータを「正常」とみなす一方で、既に存在するが認識されていない、あるいは将来的に問題となりうる倫理的課題を予測できないという、バイアスと常識的道徳に内在する根本的な問題を浮き彫りにする。

本論文は、現行の道徳的 AI 研究における二つの主要な問題、すなわち非アメリカ圏のデータセットの不足と、常識的道徳に依拠することの問題点を指摘し、これらの課題に対する解決策を提案する。

本研究における「日本語道徳理解データセット (JETHICS)」の構築、道徳的意思決定アルゴリズム「期待選択価値最大化 (MEC)」の実装、および大規模言語モデル (NLP 研究者を含む) における種差別的バイアスの分析は、国際的にも独創的な貢献である。

よって著者は、北海道大学博士 (情報科学) の学位を授与される資格あるものと認める。