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_abstract.pdf, 論文内容の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学　位　論　文　内　容　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

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）

大規模言語モデル (Large Language Model: LLM) をはじめとして、急速に人工知能 (AI) 技術が発展し、人間社会に AI システムが組み込まれつつある。このような AI 技術は私たちの生活をより豊かにする一方で、有害な結果をもたらすこともある。例えば、差別的なテキストの生成、悪質なステレオタイプの再生産、意思決定システムに組み込まれた際の予期せぬ振る舞いなど、AI 技術が私たちに直接的に有害な結果を引き起こすことがある。

本論文では、AI、特に自然言語処理 (Natural Language Processing: NLP) 研究で用いられている言語モデルに適切な道徳を実装することで、こうした倫理的問題の解決に貢献することを目指す。そこで本論文の前半では、私たちの常識道徳および規範倫理学で研究されている規範理論を AI に実装をするために、データセットとアルゴリズムの開発および実装、評価実験を行う。また他方で、こうした常識道徳への依拠による問題を解決するために、本論文後半では、動物倫理学の知見を利用し、NLP 研究における種差別および種差別的バイアスの分析を行う。

以下では各章の概要を説明する。まず 1 章「はじめに」では、研究課題について概説した後に、本論文の構成を説明する。

次に 2 章「研究背景」では、まず AI 倫理研究の背景を概観し、次にそこで提起される様々な倫理的問題の解決を目指す、人工道徳エージェント (Artificial Moral Agent: AMA) と呼ばれる研究分野を紹介する。しかし AMA 研究では哲学的研究が多く、ほとんどの研究は実装や評価実験を伴っていない。そこで、近年の深層学習の成功を受けて、AI に適切な道徳を実装するというモラル AI の研究を紹介する。また 2 章では、既存のモラル AI 研究の課題を二つ指摘する。第一に、非アメリカ圏のデータセットが不足しており、文化的な多様性に欠ける。第二に、常識道徳に対して単純に依拠することには、常識道徳の曖昧さや、常識道徳が間違いを含む可能性があるなどの問題があると指摘する。本論文ではこれらの課題を解決することを目指す。

3 章「日本語道徳理解度評価用データセットの開発」では、日本語のモラル AI 研究を促進するために、日本語の道徳理解度評価用データセット JETHICS を開発する。JETHICS は、規範倫理学の主要な理論 (功利主義、義務論、徳倫理) および政治哲学における正義の概念を参考に作成し、また常識道徳に基づいたデータセットも作成する。3 章ではまず各理論の概要について概説し、次に、そのような理論に基づいて JETHICS を開発する。次に、JETHICS を用いて、日本語 LLM と商用の LLM を対象として評価実験を行う。その結果、日本語 LLM は現状では道徳理解能力に乏しいことが示唆された。また商用の LLM である OpenAI の GPT-4o は高い道徳理解能力を示したが、日本特有の文化規範の理解には課題が見られることが示唆された。

4 章「期待選択価値最大化による道徳的意志決定アルゴリズムの実装」では、道徳的不確実性下における AI の意思決定アルゴリズムとして、期待選択価値最大化 (Maximizing Expected

Choiceworthiness: MEC) アルゴリズムを提案・実装する。道徳的不確実性とは、道徳に関する事実が正しいかどうかわからない状態のことである。このような道徳的不確実性がある状況下での望ましい意思決定手続きが MEC である。MEC は、複数の規範倫理理論に基づくモデルの出力を集約し、道徳的により適切な判断を下すことを目指す手続きである。4 章では、MEC を用いたモラル AI の実装を行い、比較として、MEC を実装したモデルと、既存の常識道徳モデルである Delphi とを比較評価する実験を行う。実験では哲学を専攻する博士課程の学生に評価をしてもらう。実験の結果、MEC を用いたモデルと Delphi は、モデルサイズが 10 倍程度異なるにも関わらず、同等の性能を示した。

5 章「常識道徳を超えて: 反種差別の観点から」では、モラル AI の開発において常識道徳に依拠することの限界を論じる。常識道徳は曖昧で誤った価値観を含む可能性があり、種差別のような倫理的問題を見過ごす可能性があるため、これに依拠したモラル AI の開発には問題がある。そこで本論文では、動物倫理学の知見を導入し、AI 倫理研究において種差別を考慮すべきだと主張する。

6 章と 7 章では、5 章の議論を受けて、NLP 研究における種差別に関する研究を行う。まず 6 章「英語のマスク言語モデルに内在する種差別的バイアスの分析」では、マスク言語モデル (Masked Language Model: MLM) に内在する種差別的バイアスを分析する。6 章ではまず、種差別と言語に関する既存研究を説明し、種差別的実践が言語使用と関連することを示す。次に、このような種差別的言語を参考に、MLM に内在する種差別的バイアスを明らかにする。実験では、テンプレートベース・アプローチとコーパスベース・アプローチによる実験を行う。テンプレートベース・アプローチではテンプレート文を用いた評価実験を行い、コーパスベース・アプローチではコーパスから抽出した文を用いて評価実験を行う。実験の結果、どちらのアプローチに基づく実験でも、複数の MLM で種差別的バイアスが確認された。

次に 7 章「自然言語処理における種差別の体系的調査」では、NLP 研究全体における種差別を明らかにするために、NLP 研究者、NLP データ、NLP モデルのそれぞれにおける種差別を調査する。6 章では NLP モデルにのみ焦点をあて、種差別的バイアスが示唆されたが、このようなバイアスは、そのモデルの開発に用いられるデータ、またそのモデルやデータセットを開発した研究者に由来する可能性がある。そこで 7 章では、研究者が出版している文献、公開されているデータセット、NLP モデルを対象として、種差別ないし種差別的バイアスの調査・分析を行う。調査の結果、NLP 研究者が種差別の問題を認識していないこと、NLP データやモデルに種差別的バイアスが存在することが示された。こうした結果は、NLP 研究において種差別を軽減することが困難であることを示唆する。そのため、NLP 研究者はより積極的に種差別に反対することが求められると主張する。

最後に、8 章「結論」では本論文全体を要約する。

また本論文の自己批判的側面について議論する。3 章と 4 章では常識道徳に依拠したデータセットおよびモラル AI の作成を行ったが、5 章で議論したように、常識道徳に依拠することには問題がある。それにもかかわらず 3 章と 4 章では常識道徳にある程度は依拠した研究を実施した。また他方で 6 章と 7 章では、常識道徳に依拠しない研究として、AI モデルに内在する種差別的バイアスの分析と NLP 研究全体における種差別の調査を行った。こうした相反するアプローチが、本論文全体としてどのように整合的であるかを最後に議論する。