研究ノート

中国国内の AI 倫理研究の現状について

侯乃禎（北海道大学大学院文学院）

要旨

本稿は、2017 年 7 月から 2021 年 1 月までの中国における AI 倫理研究に関する注目すべき話題や研究を整理して紹介するものである。まず「次世代人工知能発展計画」以来の中国における AI 倫理研究の背景を紹介する。とりわけ 2019 年に中国人工知能学会が設立した人工知能倫理道徳委員会の主任である陳小平についての記事を紹介する。次に、AI の主体性という話題をめぐるいくつかの中国の研究者たちの主張をまとめる。それは、今現在の AI が主体性を持つかどうかということについての研究と、将来の AI が主体性を持つ状況およびその状況に伴う危険性についての議論に分かれている。最後に、AI 倫理に関する中国の特色ある内容として、中国思想の研究者たちの AI 倫理についての議論を取り上げる。その節には、刘紀璐の「儒家のロボット倫理」という論文をめぐる議論と、AI と人の関係という話題をめぐる中国思想の研究者たちの議論が含まれている。それにより、まだ十分に日本に発信されていない中国国内の AI 倫理研究の動向を示す。

Abstract

This paper presents the notable researches and topics on AI ethics in China from July 2017 to January 2021. First, it introduces the background of China's AI ethics research since the *New Generation Artificial Intelligence Development Plan*. This part especially organizes the relevant news of Chen Xiaoping, who is the Director of AI Ethics Committee which established by the Chinese Association for Artificial Intelligence in 2019. Second, it sorts out the claims of some Chinese researchers on the subjectivity of AI. Two aspects were mentioned in this part: The research on whether AI has subjectivity today, and the discussion on the issues accompany with AI's subjectivity in the future. Finally, it takes up the discussion of AI ethics with Chinese characteristic by researchers of Chinese thought. This part contains not only the discussion of JeeLoo Liu's *Confucian Robotic Ethics*, but also the discussion of the relationship between humans and AI by researchers with Chinese thought. In this way, the paper shows the AI ethics research trends in China that has not yet been fully known by Japan.

はじめに

2017 年 7 月 8 日に公表された「次世代人工知能発展計画」（中:《新一代人工智能发展规划》）で、中国において AI 技術を発展させるために倫理規範が AI 戦略の重要な「保障措置」（中: 保障措施）であると、中国国務院は強調している。それ以来、中国における AI 倫理に関する研究は盛んになってきた。本稿は、膨大な数のそれらの研究に対する統計的・包括的な調査ではなく、2017 年 7 月から 2021 年 1 月までの中国における AI 倫理研究に関する注目すべき話題や研究を整理して紹介するものである [1]。

まず「次世代人工知能発展計画」に関連する AI 倫理研究の背景と、人工知能倫理道徳委員会の主任である陳小平についての記事をまとめることによって、2017 年 7 月から 2021 年 1 月までの中国における AI 倫理研究の背景を紹介する。次に、その時期における AI 倫理研究についての三つの話題を取り上げて整理する。一つ目は、AI が人間のような主体性を持った存在であるかという問題に関する議論である。この AI 倫理研究における一般的な問題に対して、中国の研究者はどのような主張をするのかということについて整理する。二つ目と三つ目は、中国研究界における JeeLoo Liu（刘紀璐）の「儒家のロボット倫理」という論文をめぐる議論と、AI と人間の関係に関する話題である。とりわけ、これらの話題についての中国思想の研究者の論文を中心に取り上げる。これにより、AI 倫理研究についての中国研究者の独創的な主張を紹介し、まだ十分に日本に発信されていない中国国内の AI 倫理研究の動向を示す。

1. 中国における AI 倫理研究の背景

「次世代人工知能発展計画」には、「倫理」という言葉が 15 回ほど出てくる。その中では、AI 技術を発展させるための法律・法規と倫理規範が、中国において AI 戦略の重要な「保障措置」として明言されている。中国電子技術標準化研究院の「人工知能倫理リスク分析報告」によれば、「次世代人工知能発展計画」で倫理をそれほど強調するのは、AI が社会に与える倫理的影響に注意を払うためだけではなく、倫理体系および倫理規範を制定して AI の安全性と信頼性、そして制御可能な発展を確保するためである。

また、中国電子技術標準化研究院は「人工知能標準化白書 2018」、「人工知能標準化白書 2019」、「人工知能倫理リスク分析報告」において、AI 技術の安全性と AI 倫理の問題を重要視している。とりわけ「人工知能倫理リスク分析報告」では、アルゴリズムに関するプライバシー保護、差別、濫用、解釈可能性、責任といった具体的な問題に注目している。また、その報告は、アシロマ AI 原則、IEEE が提唱した AI 倫理綱領、日本人工知能学会の『人工知能学会倫理指針』などを参照した上で、「人間根本利益原則」と「責任原則」を提唱している。「人間根本利益原則」は、AI 研究は人間の利益を実現するためのものであるということを強調する。「責任原則」は、AI のアルゴリズムの説明可能性や検証可能性や予測可能性を強調し、そして AI を設計する際の責任の帰属先を強調する。

1　本稿が取り扱うのは、主に「中国学術文献オンラインサービス（China National Knowledge Infrastructure : CNKI）」に収録されている AI 倫理についての中国語の文献と、インターネットで公開されている中国語の新聞記事である。

さらに、2019 年 6 月 17 日に中国国家次世代人工知能ガバナンス専門委員会は、「次世代人工知能ガバナンス原則――責任ある人工知能の発展」を公表した[2]。「ガバナンス原則」には、責任ある人工知能の発展というテーマをめぐり、調和・友好、公平・公正、包摂・共有、プライバシー保護、セキュリティコントロール、共同責任、開放・協力、俊敏なガバナンスの 8 項目が記されている[3]。そして、「ガバナンス原則」に呼応して 2019 年 8 月 29 日に上海コンピューター青年工作委員会は、2019 年の世界人工知能大会で「中国青年科学者 2019 人工知能イノベーションガバナンス上海宣言」[4] を発表した。「上海宣言」には、プライバシー保護、多様性・公平、技術の安定性、アルゴリズム透明性、説明責任、環境に優しいシステム、人間の幸福への利益、人道的アプローチの 8 項目が記されている。

中国の企業も、AI 倫理の問題を研究している。例えば、2019 年 7 月 11 日にテンセント研究院とテンセント AI Lab は、「知能時代の技術倫理観――デジタル社会の信頼を再構築する」という研究報告において、技術信頼、個人幸福、社会持続可能という三つのレベルを含む AI 研究の倫理観を示している。

そのほか、中国人工知能学会は、2019 年 5 月 26 日に人工知能倫理委員会を設立した[5]。現在、人工知能倫理委員会の具体的な成果はまだ確認できていないが、今後の中国の AI 倫理についての重要な研究センターになると思われる。そのセンターの主任である陳小平に関する 2019 年 12 月 11 日の記事がある。陳小平個人ないし人工知能倫理委員会の立場を示すために、以下ではその内容を整理する。

その記事において陳は、今の中国社会における AI 倫理研究に対して憂慮するべきことが主に三つあると述べ、それらに対する彼の意見を提示している。

一つ目は、AI 技術の発展によって、将来 AI が逆に人類を統治するという「技術の制御不能」である。陳によれば、「AI が人類を統治する」未来は、AI 研究の専門家が言ったことではないし、その情況が必ず生じるという確実な証拠もないので、短期的にはありえないと思われる。長期的には、人間はそのリスクを直視する必要がある。しかしその不確定な未来に対して、最も重要なのは主観的な憶測ではなく、それに対する客観的な科学研究である。

二つ目は、AI 技術の未熟および倫理的制御ができないといった原因によって使用者に危害が及ぶ「技術の誤用」である。例えば、データプライバシーや安全性や公平性などの問題がある。これらの問題に対して陳は、現在の科学界、産業界、および規制当局の AI 倫理の問題に対する理解がまだ不十分であり、対応する規範がないので、実際に技術倫理の問題が発生していると述べている。彼によれば、一般に歴史上、新たな産業が生まれた当初は様々な問題が生じる。これらの問題は技術と産業のさらなる発展によってのみ解消できる。問題があるからといって産業の発展を遅らせる

2　以下「ガバナンス原則」と略記。その英語版は以下のサイトで公開している。Chinadaily.com.cn. Governance Principles for the New Generation Artificial Intelligence--Developing Responsible Artificial Intelligence. URL=<http://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html>（最終閲覧日 2021 年 3 月 24 日）

3　この一文は以下のウェブサイトの日本語の記事を参照した。中国网、中国が次世代人工知能ガバナンス原則を発表、URL=<http://japanese.china.org.cn/business/txt/2019-06/19/content_74899392.htm>（最終閲覧日 2021 年 3 月 24 日）

4　以下、「上海宣言」と略記。なお、「上海宣言」は技術本位であり、AI に関する倫理問題を技術問題に還元する還元主義である、という批判がある（杨 . 2020: 138）。

5　中国の人工知能倫理委員会の設立日について、以下のウェブサイトの記事を参照した。博古睿研究院，全球视野下的人工智能伦理论坛 " 成功举办，URL=<https://www.berggruen.org.cn/activity/12>（最終閲覧日 2021 年 3 月 24 日）

のは、ただ問題の解決を遅らせるだけである。

三つ目は、AI の普遍的な応用に伴う深刻な社会的影響である。例えば、AI 技術を応用することによって多くの人の仕事が奪われ、短期間内に再就職できなくなるという「応用の制御不能」である。これについて、今の多くの企業における大量の単純作業、機械的な仕事は人間性に反するため、そのような仕事に従事する人手が足りなくなってきたという状況が存在すると陳は述べている。こうした問題および失業の問題は、産業転移と生産効率の向上によって生じた問題である。従って今の状況に対して、AI 技術を応用するのはもはや避けられないことである。

陳によると、AI 倫理研究が目指すのは、積極的に AI 技術および非技術的革新を促進し、既存のおよび新たな社会的、倫理的問題を解決し、より良い世界に向かって進む未来である。ゆえに、AI 倫理の「基本的な使命」は、「人工知能に倫理的なサポートを提供して、人間の福祉との調和と共存増進することである」。それは、AI が「良いことをするように促す」ことと、AI が「悪いことをするのを防ぐ」という二つの含意を含んでいるのである。

AI 倫理の「基本的な使命」に従って、陳は「人工知能倫理の全景動態観（中： 人工智能伦理的全景动态观）」という人工知能倫理の研究機構を設立するための指導原則を述べている。それはまず、指導原則では、上述した AI 倫理の基本的な使命に基づき、AI の倫理原則と、研究および施行細則を策定することが求められているが、これらの原則や細則の策定によって目論まれているのは、「研究」、「応用」、「需要」のループを形成することといえる。このとき、ループは次のように説明できるだろう。すなわち、ループとは、原則や細則に基づいた「研究」を積み重ね、「研究」で得られた成果を商業化するという「応用」を実現し、「応用」から新たな社会的「需要」が誘発され…という円環運動を指す。ここでの「研究」については、科学技術の研究部門と倫理の研究部門とを区別して設立することが必要であると陳は述べている。前者は、倫理的規定に符合する形で技術を研究する部門であり、後者は倫理的規定の策定および改善に従事する部門である [6]。

2. 主体性について

上記のような背景のもとで、2017 年後半から中国における AI 倫理を主題とする多くの論文が発表されてきた。その中には、AI 倫理の一般的な話題に関する研究があり、中国的な特色のある研究も存在する。この節では、主体性という AI 倫理研究の一般的な話題を手がかりにして、中国におけるいくつかの論文の内容を整理する。まず、今現在の AI が人間のような主体性を持った存在であるのかという問いについて、以下の研究を紹介する。

段偉文（2017）は、今の AI によって提示された「主体性」は、意志的能動性、自己意識および自由意志に基づくのではなく、ただ機能的な模倣であると述べている。彼によれば、そのような AI はただ、人為的な設計によって人のように振る舞わせるものなので、それを「疑似主体」と呼ぶべきである。また、彼は AI と人間は対立するという二元的な関係を批判し、コントロール（C）

6　2020 年 9 月に刊行された論文（陳 . 2020: 86-87）において、陳は上述の「人工知能倫理ダイナミクス体系」に基づいて「人工知能倫理体系構造」を提出している。そのなかでは、「研究」に関して、科学技術の研究部門と倫理の研究部門を区別して設立することは言及されず、経済的利益だけを重視する「伝統イノベーション」から、経済的利益と社会的利益の両方を重視し、異なる文化の伝統を包容できる「公義イノベーション」への移行を重視することが述べられている。

- AI（A）およびその設計者（D）- 一般的なユーザー（U）という複合的な関係こそが AI によって引き起こされる倫理的関係の基本モデルであると主張している。このモデルに基づいて彼は、責任あるイノベーション（中：负责任的创新）と主体の権利保護を強調している。責任あるイノベーションとは、AI 技術者は設計する際に一般的なユーザーだけではなく、社会全体ないし全人類に対する責任を考えるべきだ、ということである。主体の権利保護とは、人工的なエージェントが引き起こす問題の責任を明確にするためには、エージェントの制御の責任と設計の責任とを区別すべきだ、ということである。

刘永安（2021）は、意識の問題は人工的道徳的行為者（artificial moral agent）[7] の可能性の決定的な条件であると指摘した上で、現在の AI は道徳的主体にとって重要な現象学的意識欠いているので、まだ道徳的主体ではないと述べている。彼によれば、まず、AI は人のような生物的有機体ではないので、現象学的意識を生み出す実体的な基礎を欠いている。さらに、精神や意識を表現するような AI の振る舞いは単に機能的であり、AI の外に存在する三人称的視点からしか観察できない。人間の意識と比較すると、AI は自己反省の一人称視点および感情移入という道徳的主体になるために極めて重要なものが欠けている。また、感情は人の現象学的意識および道徳意識にとっては不可欠なものなのに、それを AI は備えていないと彼は論じている。

戴益斌（2020）は、道徳的主体という点では、AI（知能ロボット）は心的状態を持てないので道徳的主体にはなれないが、道徳的被行為者（moral patient）という点では、道徳的被行為者の地位を取得する可能性があると述べている。彼の言う道徳的被行為者のモデルは、レヴィナスの「他者の顔」の学説に基づいて展開されたものである。すなわち、AI をレヴィナスの言ったようなわれわれの自己が依存する「他者」としてではなく、われわれの自己に依存する「準他者」として考えることによって、AI の道徳的被行為者の地位を説明できると彼は述べている。しかし、それは AI が言語を習得した上で、積極的に人と向き合うことができる場合（例えば、ケアロボットが老人を看護する場合）、または人が AI と有意義な関係を持っている場合（例えば、言語を習得した知能ロボットに人間の情が移る場合）のみである。

さて、上述では現在の技術に基づいて、AI が人間のように主体性を持った存在であるのかということについての研究を整理したが、将来の AI は主体性を持つ状況およびその状況に伴う危険性についても、いくつかの論文が存在する。以下ではそれらを紹介する。

陳小平は中国科学報の記事（2020）で、中国科学技術大学が開発した相互コミュニケーション型ロボット「佳佳」の実験において、一種の新たな人間の経験が出現したと述べている。その実験の被験者たちは、「佳佳」と人を混同するのではなく、「佳佳」は人でも物でもないことをはっきりと感じた。この事実に対して陳は、「非人間的非物質的」（中：非人，非物）の三番目の存在物が出現する可能性があると述べている。また、同年 9 月の論文で陳は、近い将来、感情的なインタラクティブロボットのような特定の AI 製品は、一部の大衆に受け入れられる可能性があると言っている。この現象が出現する原因は、「科学や哲学の仮定とは異なり、人々は通常、ロボットによって示される感情が実際の人の感情であるかどうかを気にしないし、人とロボットの感情の本質的な違

7　道徳的行為者（moral agent）の中国語訳は、「道德行为者」（道徳的行為者）のほかに、「道德主体」（道徳的主体）もある。例えば、本文が扱う戴益斌（2020）と刘紀璐（2018）の論文には、「道德主体」という訳が使われている。

いを注意深く区別しない」（陳．2020：86-87）ことにあると彼は述べている。本質的な違いにあまり関心を払わないといった傾向は、専門家にとってとうてい支持できるものではないが、近い将来、この傾向を持つ人をわれわれは受け入れるのか、あるいは傾向を正すのかと陳は疑問を呈している。

杜厳勇（2016）は、AI 技術の安全性を確保するために、AI 製品の倫理を設計すること、AI 技術の適用範囲を限定すること、および AI の知能レベルを限定することを提案している。なぜなら、たとえスティーヴン・ホーキングの言ったような AI が「人間を越える」状況がなくとも、AI 技術の発展に伴う様々な予測できないリスクは存在するからである。

王天恩（2020）は、ニック・ボストロムの研究に基づいて、AI 技術の発展は人間全体の存在に壊滅的な脅威をもたらす可能性があり、それは人間にとっての実存的リスク（existential risk）であると指摘している。AI による実存的リスクに対して、AI に道徳的規定を内蔵させることにも、AI に人間の価値観を植え付けることにも、理論上の困難と実践上の困難が存在する。例えば、われわれ個人はどのような価値観を持つべきか、という問題はすでに十分複雑であり、AI にどのような人間の価値観を植え付けるべきかという問題はより一層困難である。王は、将来のスーパー AI を利用することでこうした難題に対応できると考えている。そして彼によれば、AI 技術に伴う実存的リスクは、伝統的倫理をそのまま適応することによって解決できる問題ではなく、世界を構築する創世倫理の問題である。なぜなら、現在のわれわれはまさに神のキャラクターを演じて人工物の世界を構築しているからである。従って、われわれは自分を創造主のような次元で、全面的に AI 技術の開発およびそれに伴う責任と結果を考えるべきであると彼は提案している。

趙汀阳（2018）は、AI が近い将来に引き起こす危険と遠い将来に引き起こす致命的な危険を分析している。彼によると、近い将来に発生する危険は三つある。一つ目は自動運転車のパラドックスである。それは、完全な自律性を持つ自動運転車が交通法規に違反する歩行者に遭遇した場合、自動運転車は乗客と歩行者のどちらを保護するのか、という問題である。その要点は、乗客と歩行者の両方を保護できるような自動運転車の技術がなければ、自動運転車のためにどのようなルールを選択してもわれわれが満足できないということである。二つ目は AI 技術に伴う失業問題である。未来の AI が多くの富を生み出し、多くの人間労働が不要となるという状況になれば、人は自主的に労働したくても何もすることがなくなってしまう。それは人生の意義の消失を意味する。三つ目は、人が他人を必要としない状況および AI 武器の出現のことである。さらに、遠い将来、言語を発明することができ、システム全体を反省する能力を持つスーパー AI が出現すれば、それはまさに神のような存在であり、そしてそれは人間にとっての致命的な危険であると彼は言っている。なぜなら、そのとき、世界における人間の地位が失われるからである。その危機を防止するために、彼は AI に自滅のプログラムを設置することを提案している。

韓敏と趙海明（2020）は、もしシンギュラリティを突破した強い AI が出現すれば、そのとき AI と人間との関係は、「主体と客体」の関係ではなく、「主体と主体」という間主体的関係になると考えている。また、その未来においては、AI と人間との間主体的関係は、インターネットだけではなく、完全にデジタル化されたバーチャル空間に通じて現れると彼らは予想している。

3. 儒家のロボット倫理をめぐる議論

中国では中国思想、とりわけ儒家思想の研究者が AI 倫理を考察した研究がいくつか出てきている。本節は、JeeLoo Liu（刘紀璐）の「儒家のロボット倫理」（Confucian Robotic Ethics）という論文をめぐる議論を紹介する。

Liu の「儒家のロボット倫理」はもともと 2017 年に英語で発表されたものが、2018 年に中国語に翻訳されて学術雑誌に掲載された後、それを批判する中国思想の研究者の論文が現れた。Liu はこの論文において主にトロッコ問題を通じて、アシモフのロボット工学三原則、カント倫理学、功利主義、儒家倫理を、人工的道徳的行為者であるロボットの倫理原則としての優劣という観点から比較した。具体的な内容は以下の通りである。

アシモフのロボット工学三原則：

Liu は、アシモフのロボット工学三原則の複雑な倫理的シナリオに対処する上での不十分さを指摘している。例えば、ロボット工学三原則に従うロボットは、ある人を助けるために他の人を犠牲にするようなトロッコ問題に遭遇した場合、何を選択しても必ず人間を犠牲にするので、「ロボットは人間に危害を加えてはならない。また、その危険を看過することによって、人間に危害を及ぼしてはならない」[8] というアシモフの第一原則に違反する。その場合、ロボットは行動不可能な状態になってしまう可能性がある。

カント倫理：

Liu は、カント倫理学に基づいて以下の二つのロボットの原則を述べている。

DR1：ロボットは、選択されたオプションが原則として他のロボットの普遍的な法則となるような仕方でのみ行為する必要がある。

DR2：ロボットは、人間性を単なる手段としてではなく、常に目的として扱うように行為しなければならない。

それから Liu は、このようなロボットの倫理原則の問題点述べている。

「DR1」について、それは標準的なトロッコ問題（Foot. 1967：2）に対して明確な解決案を出すことができないので、トロッコ問題のような選択に向かうときロボットは何も行動しない可能性がある。

「DR2」について、カントの定言命法は人という自律的な理性的主体を前提にしているが、ロボットはそれら自身のために立法することはない同時に、自由意志も持たないので、自律的な理性的主体ではない。また、「DR2」に従う道徳的行為者であるロボットを造るわれわれの行為は、道徳的行為者を目的ではなく手段として扱うので、カント倫理によればそれは不道徳である。

8　アシモフのロボット工学の第一原則の日本語訳は、以下の訳本から引用した。アイザック・アシモフ『われはロボット』、小尾芙佐訳、ハヤカワ文庫、1983 年、5 頁。

功利主義：

Liu は、行為功利主義の内容を抽出することによって次のロボットの原則を述べている。

UR：ロボットは、利用可能な一連の行動の結果を考慮する際に、関係するすべての人間に対して最大の利益を生み出すか、より大きな害を防ぐかのいずれかの行為を選択する必要がある。

この「UR」について Liu は以下の二つの問題点を指摘している。

「UR」に基づいて設計された自動運転車は、より多くの損害を避けるために、車自体と車の乗客を犠牲にする可能性がある。そのとき、たしかに大衆はこの種の車がもたらした結果を正義として肯定するかもしれない。しかし、乗客の安全が保証されない限り、おそらく大衆はこの種の自動運転車を買わないだろう（Bonnefon et al. 2016：1574）。

そして功利主義は全体の利益のために個人の利益を侵害する、というよく指摘された問題が「UR」にも存在している。さらに、ロボットには人間のような他人を犠牲にすることに対する心理的障害はないので、「UR」に従うロボットは、躊躇なく全体の利益のために個人を犠牲にする行動を取ると思われる。

儒家倫理：

Liu は、『論語』における忠、恕、仁という三つの美徳に基づいて次の三つのロボットの原則を述べている。

CR1：ロボットは何よりもまず、割り当てられた役割を果たす必要がある（忠 ; loyalty）。

CR2：他のオプションが利用可能な場合、ロボットは他の人間に最大の不快感または最も低い選好を与えるような行動をとるべきではない（恕 ; reciprocity）。

CR3：ロボットは、「CR1」および「CR2」に違反しない限り、道徳的改善を追求する他の人間を支援する必要がある。ロボットはまた、誰かの計画が彼らの邪悪な資質を引き出したり、不道徳を生み出したりするとき、彼らへの支援を拒否しなければならない（仁 ; humanity）。

「CR1」が強調するのは、ロボットはそれ自身の役割によって規定された社会的責任を果たし、役割を超えたことをしない、という「忠」の美徳である。Liu によると、これは第一義的に重要な原則である。「CR1」のメリットは、「UR」に従う自動運転車の乗客を犠牲にする問題を解決できることにある。なぜなら、乗客の安全を確保することは自動運転車の役割なので、「CR1」に従えばロボットは乗客を犠牲にしないからである。「自動運転車は、乗客の安全運転を確保する義務を果たし、スクールバスの壊滅的な被害や複数の歩行者の死亡を防ぐために、木にぶつかって乗客を犠牲にするような行動をとるべきではない」（Liu. 2017：19）。

「CR2」は、「己の欲せざるところは人に施すことなかれ」という「恕」の美徳をロボットに組み込む形で実現する原則である。「CR2」の原則はアシモフの第一の原則よりも優れていると Liu は信じている。なぜならこの原則は、負の価値（negative values）を考慮することができ、ロボットは許容される行動範囲をより柔軟に考慮できるようになるからである。同時に、「己の欲せざるところは人に施すことなかれ」という儒家の否定的な黄金律に基づく「CR2」はロボットに独善的な行動をさせないので、「他人に主観的な選好を押し付ける」（subjective imposition of preferences）問題を回避できる。

「CR3」は「己立たんと欲して人を立て、己達せんと欲して人を達す」と「君子は人の美を成す、

人の悪を成さず」から抽出した「仁」に基づく規則である。「CR3」が要求するのは、ロボットは人間にとって何が良いのか、または人間が何を達成すべきかを独善的に決定しない、ということである。同時に、人の命令が悪のためであるとき、ロボットはその人を支援することを拒否する。

標準的なトロッコ問題に対して、儒家倫理に従うロボットが運転士あるいは車掌であれば、自身の役割を果たすため線路の分岐器を使って一人の命を犠牲にして五人を救うが、それ以外の役割であればロボットは何もしない。また、トロッコ問題の歩道橋のバージョン（Thomson. 1976：207-208）に対しては、能動的に人を橋の下に突き落とす行為は「CR1」と「CR2」と「CR3」の全てに違反するので、ロボットは絶対にしない。このように、儒家ロボットは、標準的なトロッコ問題では一人の命を犠牲にして五人を救い、歩道橋のバージョンでは何もしない、という多数の人が受け入れられる形でトロッコ問題に対処する。

「私たちの社会に自動制御の人工的道徳的行為者がいる予見可能な将来には、それが行動する場合と行動しない場合の両方が人間に害を及ぼし、望まぬ結果をもたらすときは、行動するよりも行動しないことを選択してもらいたい、と私たちは望むだろう」（Liu. 2017：26）。このように、役割を超えた行為をロボットにさせない、ということを Liu は極めて重視している。

中国の学術界では、Liu の論文を批判する二つの論文が存在する。以下ではそれぞれの内容を紹介する。

まず呉童立（2020）は、Liu の議論が基づいている思考実験、すなわちトロッコ問題と自動運転車の問題は道徳的な状況を抽象化・単純化しすぎており、いくつかの重要な要因を無視する可能性があると批判している。例えば、自動運転車の場合では、自動運転車が運転者や同乗者よりも歩行者の妊婦や子供の安全を優先することはありうる。しかし「車の乗客を優先に保護する」という Liu 主張には、そのようなことが考慮されていない。

さらに呉によると、Liu の論文の立場は一種の「理想的規則主義」である。それには二つの意味がある。まず、すべての道徳的状況に対処するための普遍的かつ合理的ルールシステムを設計できる。また、このルールシステムが厳密に使用されている限り、ロボットは道徳的危険を回避できる。「人間とは異なり、AI には感情的な干渉、自己境界、利益の訴えがなく、絶対的かつ合理的に指示に従うことができる」と「理想的規則主義」を支持する人は考えていると呉は述べ、続けて、そのような人は「ゆえに教科書のようなルールシステムで AI を規制するのは自然なことである」という態度を取ると呉は言っている（吴. 2020：54）。彼によれば、「理想的規則主義」が前提とするのは、人間は責任を負うことができる道徳的主体の資格を持っているのに対して、AI は道徳的主体の資格を持っていない、ということである[9]。

また、呉によると、Liu の論文における儒家思想に対する解釈には問題がある。なぜなら、「忠」と「恕」と「仁」を規範化する Liu の解釈は、それらの持つ豊かな意味を解消するという点で不適切だからである。「忠」と「恕」と「仁」はすべて、個体と共同体との動的な相互関係、つまり共同体における具体的な文脈に応じて行為することを強調する思想だが、Liu はそれらを規範化して

9　この立場と逆に呉は、「独立して状況を判断し、ルールを調整して使用することで決定を下すことができる AI は、真の自律性を持っており、道徳的な主体として認定されるべきである」（吴 . 2020: 56）と述べている。つまり、AI は人間のような自律性を持っているわけではないが、それ自身によって予測して行為することができる複雑なルールを持っているので、やはり自律性のある道徳的主体と認めるべきであると彼は考えている。

解釈した。従って、「忠」は「役割に忠実であること」と定義された。「恕」すなわち、「己の欲せざるところは人に施すことなかれ」は禁止される行為のリストとして解釈された。「仁」は「善いことをする際に人間に許可を求めなさい」ということと「悪を助けることを禁止する」こととに還元された。このように、Liu の解釈は儒家の倫理思想を単純化しすぎたと言える。

方旭東（2020）は、彼の論文で Liu の論文に批判的に言及している。彼によると、Liu が提案したロボットに適用する儒家倫理は AI に独善的な行動をさせないことだとすれば、AI を極めて低い知能レベルに規制しなければならない。従ってそのような AI は人工知能ではなく、知能を持っていないただの機械である。

さらに方は、Liu が提案した儒家倫理は、役割を果たすことを強調する「忠」を一番に重要な原則と見なすので、「道徳に無関心」な結果を引き起してしまう可能性があると指摘している。そのような「道徳に無関心」な行動は、「仁」または「良心」に基づく道徳的行動を称賛する儒家の思想と衝突する（方．2020：785）。例えば、儒家の古典である『孟子』においては、赤ん坊が井戸の中に落ちようとしているのを見たならば、誰でも驚いて同情心（人に忍びざるの心）がわき、助けるようとするということが書かれている。

4. AI と人間の関係について

中国では、とりわけ儒家思想の研究者が AI 倫理を考察する研究には、AI と人間の関係という話題も活発に議論されている。この話題についての研究には、「人と獣の区別についての弁論（人禽之辨）」と「惻隠の心」に基づいて AI と人間の区別に着目する論文や「万物一体論」という古い中国思想に基づいて AI と人間の共存関係を主張する研究もある。それぞれの話題に関する研究の主張を整理することは、AI 倫理に対する中国思想の研究者の主張を理解するために有益であり、また中国的な特色のある AI 倫理研究を見いだすことができるだろう。

2018 年孔学堂冬季弁論大会は、「儒家の「人間と獣との区別（中：人禽之辨）」はロボットに効果があるのか」というテーマで開催された。その弁論大会の内容は論文化され発表された。以下ではそれらの論文の主張を簡単に紹介する。

呉根友（2019）は、「人間と機械との区別（中：人机之辨）」を議論する意味は、科学技術による人間の疎外およびより大きな不公平の出現を防ぐことにあるとして、技術の進歩は人々の全面発展のためであると主張している。

戴茂堂と左輝（2019）は、人間に比べて、ロボットは自己意識がないので、自由意志がないと論じた上で、ロボットは理性的かつ感情的な反応をすることは不可能であり、道徳的な意識を確立することは不可能である。なぜなら、道徳自体は人間の自由意志の自律に基づいているからである。このように、彼は人間とロボットとの区別を論じている。

董平は（2019）、「人間と獣との区別」というテーマを検討する目的が人間の主体性および生命の尊さを強調し、尊厳のある生活と人生の目標の実現することであれば、「人間と機械との区別」を議論する目的も同じであると言っている。それゆえ、AI 技術の発展はこの目的のもとで行われるべきだと彼は主張している。

上述した 2018 年孔学堂冬季弁論大会についての論文以外に、儒家の「同情心」（惻隠の心）に依

拠して人間とロボットとの区別を考察する論文もある。

付長珍（2019）は、「人間は道徳的な主体として、共感能力に基づく責任感を持っている。これはロボットにはないものである」（付．2019：42）と述べている。なぜなら、ロボットの感情は恐らくさまざまな状況に対する反応にすぎず、自然な感情ではないからである。つまり、感情に対する反応と自然な感情という経験が区別されている。人間が両方を持っているのに対し、ロボットは前者だけを持っている。この点により、人間とロボットを区別することができると付は考えている。

上記の中国思想に依拠して AI と人間との区別を議論する研究の他に、中国の伝統的思想である「万物一体論」に基づいて、AI と人間との共存的な関係を提唱する研究も存在する。

端的に言えば、「万物一体論」とは、人間と世界の関係を理解するための伝統的な中国思想である。この思想は、道家の古典『荘子』や儒家の古典『中庸』に遡る。また、宋朝の儒学者である王陽明は、血縁親族への同情心を、鳥獣、草木、石などの自然的存在物まで拡張し、世界の万物の運命と人間の運命との関連を述べている。王陽明のこの主張は、以下の二つの論文の中にも具体的に言及されている。

朱承（2020）によると、「万物一体論」の視点は、「人間」を出発点としながら、「万物」と人自身の生存、発展ないし運命を一体と見なす視点である。彼は「万物一体論」に基づいて、われわれは AI と人間の一体性を強調すべきだと主張している。彼によると、一方で、AI は人間を模倣することによって造られたものなので、それを人間の意志の延長と見なすことができる。他方、AI は人間にとっての客観的な対象物なので、それを人間と対象世界の融合と見なすことができる。

彭国翔（2020）は、「万物一体論」に従うならば、AI はただ人間によって制御される、人間のための道具である、という人間中心主義を取るべきではないと主張している。彼から見ると、人間中心主義は意識のある AI を人間にとっての脅威だと見なす AI 脅威論の前提である。最初からこの態度に固執すれば、われわれは AI と共存する未来の可能性を否定するに違いない。AI 脅威論に囚われるより、われわれは、意識と感情を持つ AI を人間のような生き物（human-like-creature）と見なすべきである。そうすることによって、AI は必ずしも人間にとって脅威ではなく、人間の友人になる可能性があると彼は述べている。

結　語

中国が AI 技術を社会的範囲で大幅に応用している現状から、中国の AI 倫理研究に対して興味を持つ人は少なくないと思われる。本稿は、「次世代人工知能発展計画」2017 年 7 月以来中国で盛んに論じられている AI 倫理の研究についての議論を、主体性という問題に関する議論と、儒家のロボット倫理をめぐる議論、そして AI と人間の関係に関する議論に分けて整理した。これにより、AI は人間のような主体性を持った存在であるかという一般的な問いに対する中国研究者の主張のみならず、AI 倫理研究という現代の学問についての中国思想の研究者の独特な考えも示すことができた。本稿は中国国内の AI 倫理研究に対する全面的、包括的な調査研究ではなく、あくまでその一部の話題を紹介するものであるが、まだ十分に知られていない中国国内の AI 倫理研究の現状およびその動向を日本に紹介するものである。今後は、世界で盛んに論じられている最先端の AI 倫理研究の成果を参照した上で、中国における AI 倫理研究の価値を示すことに努めたい。

※ 本稿は「JST-RISTEX「科学技術の倫理的・法制度的・社会的課題（ELSI）への包括的実践研究開発プログラム「人工主体の創出に伴う倫理的諸問題を分析･討議するプラットフォームの構築に向けた企画調査」」による研究成果の一部である。

参考文献

Bonnefon, Jean-François, Rahwan, I., & Shariff, A.（2016）“The Social Dilemma of Autonomous Vehicles,” *Science*, 352（6293）, pp. 1573-1576.

Foot, P.（1967）“The Problem of Abortion and the Doctrine of Double Effect,” *Oxford Review*, 5, pp. 1-6.

Liu, J.（2017）“Confucian Robotic Ethics,” Conference: The Relevance of Classics under the Conditions of Modernity: Humanity and Science, Hong Kong: The Hong Kong Polytechnic University.

Thomson, J. J.（1976）“Killing, Letting Die, and The Trolley Problem,” *The Monist*, 59（2）, pp. 204-217.

博古睿研究院，全球视野下的人工智能伦理论坛 " 成功举办，2019-08-29.
https://www.berggruen.org.cn/activity/12、最終閲覧日 2021 年 3 月 24 日

陈小平 . 人工智能伦理建设的目标、任务与路径：六个议题及其依据 . 哲学研究，2020（09）：86-87.

杜严勇．人工智能安全问题及其解决进路［J］．哲学动态，2016（09）：99-104.

段伟文．人工智能时代的价值审度与伦理调适［J］．中国人民大学学报，2017，31（06）：98-108.

董平．“人禽之辨”与“人机之辨”：基础与目的［J］．船山学刊，2019（02）：11-14.

戴茂堂，左辉．人何以为人？—— 从“人禽之辨”到“人机之辨”［J］．船山学刊，2019（02）：15-19.

戴益斌．人工智能伦理何以可能？—— 基于道德主体和道德接受者的视角［J］．伦理学研究，2020（05）：96-102.

付长珍．机器人会有“同理心”吗？—— 基于儒家情感伦理学的视角［J］．哲学分析，2019，10（06）：34-43.

方旭东. 儒家对人工智能伦理的一个可能贡献 —— 经由博斯特罗姆而思［J］. 中国医学伦理学, 2020, 33（07）：778-788.

国家人工智能标准化总体组，人工智能伦理风险分析报告．2019-04-26.
http://www.cesi.cn/201904/5036.html、最終閲覧日 2021 年 3 月 24 日

韩敏，赵海明．智能时代身体主体性的颠覆与重构 —— 兼论人类与人工智能的主体间性［J］．西南民族大学学报（人文社科版），2020，41（05）：56-63.

胡珉琦，中国科学报，AI 与情感，2020-7-23.
http://news.sciencenet.cn/htmlnews/2020/7/443181.shtm、最終閲覧日 2021 年 3 月 24 日

金融电子化，【中国科学技术大学教授、中国人工智能学会人工智能伦理专委会（筹）主任 陳小平】人工智能伦理全景动态观，2019-12-11.
https://www.sohu.com/a/359758385_672569、最終閲覧日 2021 年 3 月 24 日

刘纪璐，谢晨云，闵超琴，谷龙．儒家机器人伦理［J］．思想与文化，2018（01）：18-40.

刘永安．人工道德主体是否可能的意识之维［J］．大连理工大学学报（社会科学版）：2021（01）：1-6.

彭国翔．人工智能最终一定是人类的威胁吗 —— 一个儒家的视角［J］．道德与文明，2020（05）：28-35.

上海科技，创新人工智能治理，青年科学家发声《上海宣言》，2019-08-30.
http://www.shkjdw.gov.cn/c/2019-08-30/517552.shtml、最終閲覧日 2021 年 3 月 24 日

腾讯科技，腾讯发布人工智能伦理报告 倡导面向人工智能的新的技术伦理观，2019-07-11.
https://tech.qq.com/a/20190711/004971.htm、最終閲覧日 2021 年 3 月 24 日

吴根友．儒家的“人禽之辨”对机器人有效吗？［J］．船山学刊，2019（02）：1-4.

吴童立．人工智能伦理规范是什么类型的规范？——从《儒家机器人伦理》说开去［J］．文史哲，2020（02）：51-59.

王天恩．人工智能存在性风险的伦理应对［J］．湖北大学学报（哲学社会科学版），2020，47（01）：1-8

杨庆峰．从人工智能难题反思 AI 伦理原则［J］．哲学分析，2020，11（02）：137-150+199.

赵汀阳．人工智能“革命”的“近忧”和“远虑”——一种伦理学和存在论的分析［J］．哲学动态，2018（04）：5-12.

朱承．万物一体视域下的人工智能［J］．学术交流，2020（05）：21-29+191.

アシモフ、アイザック（1983）『われはロボット』、小尾芙佐訳、ハヤカワ文庫

中国網（2019）「中国が次世代人工知能ガバナンス原則を発表」http://japanese.china.org.cn/business/txt/2019-06/19/content_74899392.htm、最終閲覧日 2021 年 3 月 24 日