特集「道徳判断の自動化をめぐる問題：規範の選択と協力の進化」にあたって

岡田 勇
（創価大学）

1. はじめに

汎用人工知能の進展に伴い，多くの活動が人工知能によって代替されていく．そのとき，善悪を含む道徳判断やジレンマ状況での選択はどうあるべきであろうか．例えば，完全自動運転車には，歩行者と乗客の安全が互いにトレードオフになっている状況において，どちらの安全を優先させるかという道徳のジレンマに答えを出すことが求められる．完全自動運転車が乗客を犠牲にする判断をしたときに，歩行者を犠牲にすることを止むなしとする価値観を支持する立場からは人工知能の価値基準に異議がある．また，けが人を助けるという規範は基本的に人にもAIにも受け入れられるだろうが，今後の社会で行動プロファイルが可視化される中で，明らかな悪人に対して援助すべきか，また援助した人（しない人）をどう判断すべきか，さらには悪の度合いと困窮に差がある複数人をどのような判断基準をもって援助すべきかという規範は人間社会においてすら共有されていない．

多様な価値観が混在する状況は，人工知能と人間の混在する系のみならず，人間社会そのものにおいてもしばしば観察される．インターネットが社会の隅々にまで浸透し，人々の移動がかつてない規模で行われる現代社会にあっては，一方の価値観からは社会的に望ましい行動であっても，他方の価値観からは反社会的な行動と映る場合もあるだろう．価値観の対立は時に深刻な社会問題を引き起こす．価値基準を含む倫理的検討への科学的アプローチの開拓は，社会的に喫緊の課題の一つであるが，どのような価値基準が望ましいかを合意形成することは難しい．むしろ，道徳的判断の検討は，これまで倫理学・哲学などの限定的な分野で議論されるに留まり，価値中立を標榜する科学がこれまで検討を避けてきたテーマにほかならず，昨今に至るまで学界からの積極的な議論は回避されてきた現状があるように感じる．

本特集では，近年ますますその傾向が強まっている，複数の価値基準が混在した系において道徳判断の自動化が求められている状況に対し，どのような規範を選択し，どのような協力的な社会を構築していくのかという社会的課題に対して，さまざまな学問分野がどのような挑戦をしているのかに焦点を当てるべく企画された．幸いにも工学のみならずネットワーク分析，経済学，社会心理学，倫理学，法学など広範な分野で最先端の研究をしている方々に執筆をいただくことができた．この点だけでも，この問題が人工知能分野のみならず多くの研究者の関心を引く重要な社会的問題であることがわかる．

本特集号では，執筆いただいた内容を大きく三つのグループに分けて掲載することにした．はじめに「問題意識」のグループAである．本特集号が挑戦しようとしている社会的課題を俯瞰した解説を2本と，価値観の対立に対して，ネットワーク分析を用いて規範や道徳を計量した意欲的な解説を2本執筆いただいた．次にグループBでは，「規範の進化」に焦点を絞り，ゲーム理論や社会心理学の専門家に緻密な論理展開によって，あり得るべき規範の姿を探求していただいた．また，人工知能研究者からも執筆いただき「公平性」をどう計量するのかといった意欲的な研究も紹介いただいている．最後のグループCでは，「異分野の視点」として，倫理学や法学などがこの問題を，またこの問題に挑戦している人工知能研究者をどのように見ているのかを執筆いただいた．

2. 各解説の概要

本章では，それぞれ寄稿していただいた解説について，その概要を紹介する．特集号の全体像を把握するうえで参考になれば幸いである．

まず，「グループA」では，現代社会が「判断の自動化」を要請していることを論じていただく．山川氏の解説では，社会的複雑性と意思決定の迅速化が強まっていることを指摘し，どのような技術的・制度的課題が生じているのかを紹介している．福島氏の解説では，意思決定や合理形成の分野で実務的な課題となっている点を紹介いただき，その課題を解決するためにどのような技術開発が考えられるのかについて，現状の網羅的なレビューも踏まえて考察していただいた．鈴木氏の解説では，Twitterを用いた世論分析で，現実の価値観の対立がどのように観察されるのかについて紹介している．これらの論考により，現代社会が価値観の対立を引き起こしやすい状況になっているにもかかわらず，これらの混在した状況をどのように合意形成していくのかが難しい状況であることを確認する．笹原氏の解説では，ソーシャルデータを用いた道徳基盤の測定に関する意欲的な研究を紹介いただいている．

山川氏には，本特集号の最初の論考にふさわしく，な

ぜ「道徳の自動化」という問題が生じているのかについて解説していただいた．意思決定が加速化し，対処すべき問題が複雑に絡み合っている現代社会の特徴を紹介しつつ，社会的決定を AI に委ねざるを得ない状況が現出していることを指摘している．人工知能研究を中心的に牽引している一人の論陣であるだけに，不気味な説得力がある．本特集号の意義について高い位置付けをいただいた．

福島氏は，はじめに意思決定の困難さについて概観している．個人レベルでは，限定合理性に代表される要因の見落としやソーシャルメディアによる思考誘導，価値観の対立と分極化といった顕在している問題に加え，リアルなフェイクニュースを簡単に生成できる現状がもたらす問題にまで踏み込んでいる．また，これらを解決するための技術的観点として，人間に寄り添うエージェントの開発，健全な意見集約が可能なプラットフォームの形成，そして，人間自身の判断能力に関するリテラシーの重要性に言及し，それらを実現するための要素技術の開発現状について丁寧に俯瞰している．人工知能技術の開発には，このような社会的な文脈に対する鋭敏な感受性が重要であるとの感を深くした．

鈴木氏は，インターネットコミュニケーションの進展は一般的に社会の分断を促進するという研究に対して非常に丁寧なレビューを試み，それが米国の政治事情を反映したもので，日本では必ずしも分断化は生じていないことを説明している．また，分断が生じないための「副産物的学習」と呼ばれる，選択的接触が容易な環境であったとしても，多様な情報を目にする機会を提供することの重要性について高い説得力をもって指摘している．特にインターネットの情報と社会調査などの情報を統合させることで個人のイデオロギー特性を同定化する手法など，一連の丁寧な研究は反論の余地を少なくするため，この分野の研究の進展を確かなものにすると好感した．

笹原氏には，Twitter の数百万にわたる投稿データを定量的に分析することで，人々の道徳基盤の推定が可能であるかどうかについて検討していただいた．その結果，先行研究が提示した五つの道徳基盤について，擁護基盤が根本的な道徳基盤であるといった知見を抽出することに成功し，tweet する人々の道徳性の計測に道を開いている．人々の道徳性の計測化は，本特集号が提示する道徳判断の自動化にとって避けることのできない現実である一方，tweet されたものがその個人の全人格の投影であるかどうかといった繊細な議論も要請されるなど，さらなる研究の深化が求められるであろう．

次に，「グループ B」ではこの状況に対し学問はどのようにアプローチしているのかをサーベイしている．複数の規範が混在する系がどのような社会的帰結をもたらすかについては，協力の進化というテーマで多くの研究の蓄積がある．Han 氏らの解説では，進化ゲーム理論研究の専門家としてこれまでの研究を概観し，特に機械倫理に関して詳細なサーベイを紹介している．また山本氏は，複数の規範が混在する系で各規範がどのような役割を果たしているのかについて，進化の観点から数理的に分析する新たな手法を紹介している．竹澤氏は，社会心理学の立場から，道徳規範の存在理由について探っており，規範意識に関する実験成果と未解決の問題について紹介している．神嶌氏らの解説では，公平性配慮型データマイニングの研究を踏まえ，人工知能関連の技術者が捉える「公平」概念とその実装や課題について紹介している．

Han 氏らには，この特集号の趣旨に賛同いただき，特別に寄稿いただいたものである．筆者自身の訳で恐縮であるが，進化ゲーム理論の専門家として，誠実さ，謝罪，懲罰，寛容，コミットメントといった行動がどのような社会的帰結をもたらすかについて，豊富な研究の蓄積を紹介いただいた．道徳判断の自動化にとって，こういった研究蓄積の豊富な「進んだ」分野からの知見を活用できると有利であろうと思われる．

山本氏には，複数規範の混在系において，いかにしたら協力的な社会が構築できるのかという協力の進化研究の最新成果をご紹介いただいた．読者は，道徳規範が形式化され，理論や実験においてさまざまなレベルの検証が行われていることを実感するに違いない．特集号との関連を議論している部分において「新たな善きサマリア人の法」，すなわち，高い情報処理環境にある現代において，善意により誠実に行動した結果の失敗・損失には責任が問われないとする規範が求められると挑戦的な提言をしており，興味深い．

竹澤氏の解説では，社会心理学の立場から，道徳規範がなぜ存在しているのかを説明しようと試みている．理論的見解や実証データ・実験データの広範なサーベイを踏まえ，社会化と罰が規範の維持に重要な役割を果たしていることを緻密な論理展開で示している．そのうえで，本特集の狙いに対し「社会規範にまつわる最後の大きな謎の解明に貢献」できる可能性があると期待を寄せてくださっている．読者は道徳研究に対する別の観点からのアプローチについて理解することができるだろう．

神嶌氏らの解説では，はじめに人工知能が参照するデータ自体が人間社会に内在する偏見によってゆがめられている現状を，ネット広告配信と再犯リスクスコアの事例を通して紹介されている．機械的に単純なデータマイニングだけでは不公平な結果をもたらすこと自体が，筆者には一種の衝撃であった．この点を解決する手法として，公平性を配慮したデータマイニング研究の広範なレビューをしていただいている．かなり精緻な議論が積み上がってきており，一つの問題解決方法として，期待できる．特に結論部分において「分析技術で生じた不公平な状況は，本稿のようなアルゴリズムの改良によって対処できるものである」と記述しており，研究はかくあるべしとの見本を提示していただいた感がある．

「グループC」では主に人工知能に判断を委ねるという問題に対して，異なる学問分野はどのように捉えているのかを紹介している．村上氏の解説では，科学技術社会論の立場から道徳判断の自動化を真正面に取り上げ議論していただいている．神崎氏の解説では，人工知能研究者が倫理をどう考えているのかについて用語の用法に注目した分析を展開していただいている．平野氏は，法律の観点から人工知能の倫理的側面を考察している．特にロボット法というユニークな法制度の考え方について紹介している．

村上氏は，哲学者の立場から真正面にこの問題について考察していただいた．さまざまな具体的問題を事例としてあげながら，ある種の価値基準の導出を試みようとしていることに注目したい．筆者は，そのなかでも「社会制度による対策」の部分で取り上げられている事例に非常に興味を引かれた．結語にある「いずれの段階にしても，研究開発自体が興味深いものであるから，というだけで進めてしまうことはもはや許されない」という警句はすべての人工知能研究者にも適用できるだろう．このような力強い論考が寄せられ，分野を超える研究者間対話が開始されていることを踏まえると，この特集号に意義を見いだすことができると自賛したい．

神崎氏には，IEEEが発行した倫理に関する報告書について，厳格な方針で分析していただき，人工知能分野で「道徳」関連用語がどのような意図や解釈で用いられているのかについてレビューをしていただいた．読者は本稿によって全体的な地図を得た感覚を味わうことができる一方，道徳問題に対してまだ人工知能研究が数多くの課題を残したままであることを痛感するに違いない．非常に説得性の高い筆致で描かれており，筆者自身が勉強になったが，なかでも，「人工技能技術およびその研究者は，すでに国際的に確立された人権や，持続可能な開発目標（SDGs）のような国際的に合意された目標などの統制を受けるだけでなく，できる限りそれらの実現に貢献すべき」という言明に強い説得力を感じた．

平野氏には，法学の分野において人工知能がもたらす社会的な革新をどのように考えているのかを説得力をもって論じていただいた．「倫理よりも法こそが，AIの開発や利活用において人々や事業者を望ましい方向に導く効果を有している」という言明は，事前に道徳をプログラムしなければならないAIの「宿命」に対する回答ではないだろうか．同時に，「新技術に対する早過ぎる〈法規制〉は失敗に終わる可能性を含んでいる」として，ソフトローとしての規範が重要であるというバランス感覚にも，なるほどと思わずにはいられなかった．

3. おわりに

道徳判断基準については，さまざまな機関がその指針をまとめる段階になっている．本学会も指針を出している[JSAI 17]が，IEEEも倫理的に適合したAIという指針[IEEE 18]を出したり，欧州委員会もEXC倫理指針を準備しているなど，その動きは活発化している．しかしこれらはあくまでもガイドラインであり，自動化という難問にとっては，スタートラインに立ったばかりである

MITの研究チームは「道徳マシンの実験」と題する論文[Awad 18]を発表しているが，そのなかで，完全自動運転車で生じる道徳のジレンマに関するオンラインの被験者実験について報告している．それは延べ233か国地域の4 000万データの分析結果であるが，選好パターンの違いによって世界を三つの文化圏にクラスタリングしているなど一連の知見を抽出することに成功している．一方，個別の道徳観を集積したデータをどのようにプログラムすべきかは，未解決の問題のままである．

道徳をプログラムしなければならないというのは，本来自己矛盾した発想なのかもしれない．個々の事象はそのときの総合的な判断，しかもその判断も動的なものであるはずだ，でなされるべきであり，事前にプログラムすることは困難が伴う．のみならず，人間の善性がそもそも記述可能なのかという難問にも突き当たる．その意味で，道徳をプログラムしなければならなくなってきた現代社会は，歴史的には不幸な時代なのかもしれない．

本特集号は，道徳判断の自動化が避けられない現代において，この問題がいかに難問であるかを示すとともに，どのような解決が可能かを探る知性の格闘について紹介するものである．幸いなことに重厚な論考が集まった．異なった学問分野の異なった問題意識が羅列されており非常に興味深い．まだまだ試行錯誤のテーマであるので，執筆分野に偏りがあるかもしれないし，取り上げていない重要な切り口も存在するだろう．しかし，本誌を読まれる方々にとって何らかの問題提起になれば幸いである．

謝 辞

本特集号は，筆者が本学会編集委員を務めていたときにお話をいただいたものである．基本的なアイディアは温めていたが，編集委員会で何度か議論の俎上に載せていただいたことで，多くの執筆者をご紹介いただくことができた．この特集号が重厚になったのは編集委員会の皆様のさまざまな助けがあってのことであると感謝している．

◇ 参 考 文 献 ◇

[Awad 18] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F. and Rahwan, I.: The moral machine experiment, *Nature,* Vol. 563, pp. 59-64（2018）

[JSAI 17] 人工知能学会倫理指針，人工知能学会，http://ai-elsi.org/wp-content/uploads/2017/02/人工知能学会倫理指針.pdf（2017）

[IEEE 18] IEEE Ethically Aligned Design, Version 2, IEEE, https://ethicsinaction.ieee.org/（2018）