2E5-OS-12b-1

人工知能と共存する人間:哲学へのインパクト

Symbiosis of AI and Human: impacts on philosophy

村上 祐子[*1]
Yuko Murakami

[*1] 東北大学 1
Tohoku University #1

To understand artificial intelligence is a backdoor to understand human beings. Philosophical debates and technological developments coevolve. This paper will focus on individualization of responsible agents under unlimited duplication and incarnation of information. The key issue is human readability of AI process and outputs concerning deontic evaluation of real-world problems, since social agreements, in particular responsibility and commitments of human beings, cannot be omitted in decision making process. It will be both a target of technological development and elaborations of philosophical theories on concept formation and uses of language.

1. 人工知能(AI)に照らした人間そのものの理解

AI と人間が共存する社会でも，我々の日常生活の様式や考え方が哲学理論と現実とのインタフェースであるのは不変だ．一方で，哲学及び周辺領域における長年の議論は AI をめぐる考察に洞察を与える部分がある．たとえば，プライバシーやコミュニケーションをめぐる情報倫理をめぐる議論は人工知能の哲学への直接の拡張が可能である．また動物倫理には人間とは認知・思考の様式が異なる存在に対する公正を論じる点で，AI をめぐる議論に応用可能な側面がある．

科学技術の進歩を含めたその時代の状況を踏まえて哲学理論は展開されるので，人工知能が普及する社会の当然の帰結として哲学理論は修正を迫られている．だがこれはまっさらから始まるのではなく，既存理論と整合性が取れる限りで概念や語彙の追加・理論の拡大といった形で進む．

特にこれからの人工知能技術がテーマとなる場合，人間の身体能力・思考能力の拡張・補完を伴う応用が見込まれるため，「人間」そのものをめぐる諸概念の再考が必至だ．

2. 情報の増殖する具現化 (incarnation) と行為主体の個別化問題

情報が因果系列に影響することは特に驚くことではない．ガラスのコップを落としたら壊れることがあると知っているからこそ，これまで一度も落ちたことがないコップを落とさないように配慮するなど，周辺環境の知識により手段を選び取り実行する．また，行為は物理的である必要はなく，約束や結婚のプロポーズなど，言語・情報伝達によってさまざまな行為がごく普通に実現する．さらにいえば行為主体が人間である必要すらない．機械が行う電子商取引によって為替や株価は上下し，我々の日常生活はゆすぶられる．

自動運転車開発で見られるように，問題は，このような行為につながる選択肢を評価し選択するシステムの複製が可能であること，また現実の因果系列への干渉方法も自律システムとして制御できるよう開発されてきたことである．複製された「同じシステム」が相互に干渉しながらそれぞれに判断を行うときに発生したミスはそのローカルシステムに固有の支障とシステム全体の問題との可能性がある．だがシステムを改修すればそのまま使用可能であると考えてよいのか？

従来の行為主体に関する考え方をそのまま適用できればそれほど問題はない．たとえばコンピュータ囲碁 AI が指示する着手点に石を置く人間は盤上に石を置く行為を行っているが，着手点を決定するという行為を行っているわけではない．その意味で，AI に使役されているとみなすこともできる．だから意思を置く場所を間違えたとしたら責任を問われるが，見るからにひどい手を打っても非難されることはない．つまり行為ごとに責任を問われる主体が存在することを我々は暗黙に了解しており，そこに AI が介在したとしても考え方は変わらない．

だがいまや AI と機械の組み合わせにより，この主体が物理的に唯一であるという前提を再検討しなければならない．操作者を機械化・AI 制御にすれば，エージェントの能力によらない均質な操作を行える．電王戦で用いられたロボットアームでは一台だったが，複数台を同時に作動させればほぼ同じ物理的操作の繰り返しが可能だし，3D プリンタのような遠隔での複製も可能になっている．また学習についても，個々のエージェントが学習した内容を不完全ながらも伝え合うという仕方ではなく，複数のエージェントが並行して行った学習の成果を統合する分散学習システムを Google が開発している．

さらに人間が AI の支援の下で行う決定によりミスを犯したとしたら，決定した人間が責任を持つのか，AI とその人の「共同責任」になるのか，それとも AI（とその開発者）が責任を担うべき行為主体となるのか？

このような状況下で，知識と行為主体の境界はどこか？個別の主体としてどうやって切り分ければいいのか？

3. 機械の意識と社会的合意

このとき，個別の行為主体については十全な意識が必要だという主張がありうる．通常十全な行為能力があるとみなされる人間であっても何らかの理由で意識をなくせば判断能力が欠落し責任能力を欠くとみなされるからである．この考え方に従えば機会に意識がなければ行為主体とみなすことはできず，AI と人間の共同作業の責任は常に人間側にあることになる．

機械の意識があるかどうかについては，先のものも含めて意識がないという立場と，意識があるという立場がある．だがこの対立は見かけ上のものであるという立場をここではとりたい．意識の社会構成説にしたがって機械にも人間同様の意識があるけ

連絡先：東北大学大学院文学研究科，宮城県仙台市青葉区川内 27-1，ymurakam@m.tohoku.ac.jp

れども，その構成プロセスに必要なのは生命ではなくエージェントとして名指しされることである．名前と指示対象の組が共有されて初めて，社会構成員それぞれの心の理論の構成および社会的コミュニケーションで各エージェントが共有することが可能となる．

そもそも，Google 自動運転車に対するように，AI 単体に法人格を付与することが社会的に決定されれば，もはや機械が意識を持つかどうかは議論する必要がなくなる．このようなメタの可能性も含めて，機械の意識の有無は論理的な問題ではなく社会的決定に左右される問題である．

さらに意識の社会構成のためには，主体間のコミュニケーションだけがあれば十分というわけではなく，コミュニケーションがなされていることが他の構成員も含めて認識されなければならない．前者で形成可能なのは相互知識止まりで，主体全員の了解があってはじめて共有知識が形成されるからだ．だからAIが人間と共生する社会では，機械の処理結果を人間に理解可能な提示方法（人間可読化）が必要となる．

4. 行為主体の道徳判断

行為主体は意識の有無によって範囲が決定されるものではなく，行為能力を社会的に付与された主体であれば人間でも機械でも扱える理論展開が必要だ．その意味で自動道徳判断は人間の道徳判断の研究とパラレルだが，AI そのものによる道徳判断の研究はいまだ発展途上だ．

4.1 AI の道徳判断の現状

(1) ムーアの 4 分類

[Moore 2010]では，自動道徳判断システムを以下の 4 レベルに区分し，第 3 レベルの研究が精いっぱいで第 4 レベルへの到達は難しいとしている．

- 第 1 レベル：結果が道徳的に判断される
- 第 2 レベル：道徳判断はデータとして入っている
- 第 3 レベル：自動道徳推論システムを備える
- 第 4 レベル：人間同様の推論能力と責任能力を備える．

その後，自動運転車等の研究が進んでいるが，人間にも判断が困難と思われる第 4 レベルを超えるトロッコ問題について論じられる一方で，第 3 レベルの実装について開発途上である．

(2) 価値判断問題の 4 レベル

メタ判断を含む自動道徳判断システム以前に，個別の価値評価問題そのものも 4 レベルに分けられる．

- 第 1 レベル：すでに価値基準が与えられている状況で最善の選択肢を探索
- 第 2 レベル：与えられた複数の可能性に対して価値を評価
- 第 3 レベル：与えられた問題のスコープを設定し，スコープ内の可能性について価値を評価
- 第 4 レベル：自動道徳推論システムを備える：何が問題なのか設定したうえでスコープを設定し価値を評価

第 1 レベルに関しては，たとえば 1 階述語論理の意味論が与えられたときに評価することに相当しており，機械が得意とする領域である．複雑な問題に関しても人間よりもはるかに高速で計算可能である．複数の可能性の評価を勘案する第 2 レベルに関しても，可能性の範囲と個別の事象の見積もりが所与のものに関してはすでに公理化され[Murakami 2005]，実装されているシステムが存在する[Bringsjord 2007]．

だが，前節第 3 レベル以上相当システムが直面する実世界問題は当節の第 3-4 レベルにあたる．考慮すべき事象のスコープが未設定のこともあるし，スコープが決定済みでも個別事象の発生確率やもたらされる被害・利益があらかじめ見積れるとは限らない．たとえば発生確率が非常に低い事象について統計的に優位な実験を行うためには莫大なコストがかかり現実には困難である．このような「トランス・サイエンス」といわれる領域の問題は実世界では不可避であり，科学的には決定できない点を社会的合意で何とか解決しながら人間は暮らしている．だからこのような現実社会で自律道徳判断システムが機能する前提として，人間を含めたエージェントの社会的合意が必須である．完全な自律道徳システムでは，自分勝手で協調性が欠落しコミュニケーションも不可能な人間と同様に，社会的に受容されないだろう．

(3) 実生活での自動道徳判断における人間の役割

自動道徳判断システムの実世界問題解決にあたり，人間が決定に関与するにも，人間に理解可能な仕方で AI が問題提示する必要がある．機械学習の出力結果は人間では思いもよらないものであることがあり，想定外の問題では改めて社会的判断を待たなければならない．このような出力結果を人間可読な方式で提示することで，人間の理解のシステムと接合させ，また社会的合意形成手続の中に位置づけなければならない．

このとき，AI を含むエージェントを対象とする法制度がなければ，AI を社会の中に取り込むことはできない．つまり，社会的問題に AI による解決を求めるのであれば，法制度・規約は合意形成手続を円滑にするものとして必須である．

5. 結語

情報の物理的実現は従来通り個別化可能であるが，情報そのものの個別化はうまくいかない．とくに行為主体の範囲特定は論理的に決定されるものではなく，社会的合意が前提だ．自動道徳判断システムについても，人間がいない惑星上や工場内などの環境ではなく，現実にこの社会の中で機能させるためには，科学技術による解決だけでは十分ではなく，開発・設計から運用までのすべての局面で社会的合意を取り込む仕組みが必要である．法制度・ガイドラインはその実装の一端である．

一方で AI を適切に使用すれば，これまで現実的に不可能だったリスク個別事象の発生確率や被害の見積が可能となるなど，社会的合意形成を支援できるかもしれない．その際にも AI 単独で判断を行うのではなく，必ず社会構成員である多様な人間の関与が発生する．よって人間可読性の向上が急務である．

シンギュラリティを「独り歩きする AI」のホラーストーリーとしないために，人間と共生できる AI の開発が必須である．AI の潜在的多様性に思いをはせれば，そんな共生社会では人間の多様性の範囲は小さいように見えるかもしれないが，一律に「人間」でくくってしまわずひとつひとつ拾い上げることになろう．

参考文献

[Bringsjord 2007] Selmar Bringsjord,. Philosophical Studies, 136:59-97, 2007. DOI 10.1007/s11098-007-9143-7

[Horty 2001] John F. Horty, *Agency and Deontic Logic*. Oxford University Press. 2001

[Murakami 2005] Yuko Murakami, Utilitarian Deontic Logic. *Advances in Modal Logic*. 5: 211-230 Kings College Publication, 2005.

[Wallach 2010] Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2010.