2I5-OS-17b-3

人工知能は道徳的になりうるか，あるいはなるべきか

Can or should AI be moral?

久木田水生 *1
Minao KUKITA

*1名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University

Moral decision making seems to be important part of our intellectual activities. Then, can AI perform the task on behalf of humans? In the US, researches are going on to create artificial intelligent systems that can make moral decisions. In my talk, we shall consider technical and ethical difficulties concerning such enterprises, especially difficulties that are different from those concerning AI in general.

1. はじめに

道徳的であるということの中には，自分が置かれた社会的環境の規範を学習し，他者の感情を思いやり，様々な価値を評価し，適切に状況判断して行動に移す能力を身に着けていることが含まれる．これは極めて高度な知的活動である．従って，もしも人工知能が人間の知的活動一般をシミュレートすることを目指すのであれば，それは道徳性もまた実現しなければならない．実際，人工知能やロボットの社会への浸透に伴って，道徳的に判断し，道徳的に振舞うことができる人工知能・ロボットの開発を提唱，推進している研究者たちがいる（例えば [Wallach and Allen 09], [Anderson and Anderson 11] を参照）．この分野は機械倫理，人工倫理，人工道徳などと呼ばれている．

自律的に活動する機械やソフトウェアが社会で実用化され，機械同士あるいは機械と人間が複雑に相互作用する時，それらの振る舞いを正確に予想するのはきわめて困難である．場合によっては破滅的な事故につながることも考えられる．そのような事態を避けるため，機械が生命・健康などの価値を配慮し尊重できるようになることが望ましい，と彼らは主張する．いまや機械が道徳的実践の一翼を担うべき時なのだ，と．

しかしながら人工道徳の開発・実用化にはいくつかの重大な問題がある．本発表ではそれらについて考察をする．

2. 人工道徳の方法論

現在までのところ，人工道徳は伝統的な記号的・論理的人工知能のパラダイムに沿って推進されている．その背景には「道徳的判断はアルゴリズムに従った計算である」という了解がある．例えば 2014 年に米海軍に支援をうけて道徳的に行動するロボットの開発を始めた AI 研究者，セルマー・ブリングジョードは義務様相論理のような倫理的推論のための大枠となる論理と，個別的な倫理的課題の遂行のための新しく考案された論理を使うことを提案している（http://www.kurzweilai.net/can-robots-be-trusted-to-know-right-from-wrong，2015 年 3 月 27 日アクセス）．また「機械倫理」プロジェクトの主導者の一人である倫理学者のスーザン・L・アンダーソンは，倫理は原理的に計算可能であるという前提に彼らが立っていることを明言している（[Anderson 11]）．そして実際に彼らが開発しているシステムは，帰納論理プログラミングという論理的 AI の古典的な手法を使っている（[Anderson et. al.]）．

もう一つ，人工道徳に従事する研究者たちが明示的あるいは暗黙に前提としていることは「道徳的に見える振る舞いで十分とする」という，いわばチューリング・テストの道徳版のような了解である．これは [Wallach and Allen 09] における，意識や意志を伴う「十全な道徳性 full morality」と，あくまでも振る舞い上の「機能的道徳性 functional morality」を分けて，機械が持つことのできるのは後者のみであり，私たちはそれに満足するべきだという主張に顕著である．実際，[Allen, et. al. 00] は機械の道徳性を判定するための「道徳的チューリング・テスト」を提案している．

3. 問題点

ヨーロッパ，アメリカの倫理学においては理性や知性の働きを重視する，いわゆる合理主義の伝統が主流であり，人工道徳，機械倫理はそのような伝統からの自然な帰結である．[Nadeau 06] のように，人間よりも機械の方がより完全な道徳を実現できる，と主張するものもいる．しかしながら，こういった趨勢に対しては反論もある．たとえば [Gunkel 12] は機能的道徳性の追求が道徳の「官僚主義」をもたらす可能性を指摘している．また [Beavers 11] は，もっぱら実装可能性の観点から，どのような倫理を採用するかということが論じられることは，「倫理的ニヒリズム」につながると批判する．

しかしながら人工道徳にはより現実的な問題もある．倫理においては「何を為すか」ということと同様に「誰が為すか」ということが重要なのである．同じ行為でもそれをやる人間によって道徳的評価は異なってくる．責任能力のない主体（子供，認知症患者など）については道徳的評価ができない．道徳的に重大な帰結をもたらす行為には責任が伴うが，ロボットの行為については責任をとる主体が曖昧である．責任能力のない主体に道徳的に重大な行為を行わせることは，それ自体が無責任な行為として批判される．

このような懸念は，アメリカ，イギリス，イスラエルなど，軍事技術大国たちが自動的に（人間による介入なしで）敵を攻撃する自律型兵器，いわゆる「殺人ロボット」の開発を推進している状況で，より深刻化している．戦場においても兵士が守らなければならない法的・道徳的ルールがある．非戦闘員や投降して来た相手を攻撃してはいけないというルールなど

連絡先: 久木田水生，名古屋大学大学院情報科学研究科，名古屋市千種区不老町，052(747)6883，minao.kukita@is.nagoya-u.ac.jp

である．アメリカは，このような戦場におけるルールを守りながら戦闘を行うロボットの開発を進めているのである（cf. [Arkin 09]）．しかしこのような動きに対しては，自律型兵器の犯した戦争犯罪に対して責任を取れる主体がいない，という問題が指摘されている（cf. [Sparrow 07]）．

また心理学者のシェリー・タークルは Paro[*1] のような，相手を気にかけているように見せかけて，ユーザーの感情的反応を引き起こすように設計されたロボットを，私たちの社会的関係の「真正さ authenticity」を損なうものとして批判している（[Turkle 12]）．ユーザーはロボットの外見と振る舞いによって「ダーウィンのスイッチ」を押されて，ついロボットが自分を気にかけてくれていると感じ，そして自分もロボットを気にかけるようになる．しかしここでの感情は一方通行であり，人間同士の社会的関係のような相互性がない．そのような関係に甘んじることを許容する風潮が一般化することにタークルは警告を発している．

同じことは道徳的ロボットについても言えるだろう．欧米の主流の倫理学者はそうは考えないが，私は道徳において感情は本質的な要因であると思っている．単なる記号操作として計算するのではなく，相手にとっての価値を理解し，思いやり，共感することが道徳の基礎である．言い換えると，人工知能・ロボットが扱う記号は，人間的な価値に接地していなければならない．チューリング・テストでは道徳性は計れない．そして機能的道徳を十全な道徳の代用とすることは，道徳の真正さを損なう可能性がある故に道徳的に問題である．しかし「倫理的記号接地問題」[*2] は，通常の記号接地問題よりもはるかに解決が難しいであろう．「リンゴ」という記号に対応する対象は厳然として存在するが，「幸福」，「尊厳」などの記号に対応する対象は私たち人間にとっても捉えどころのないものだからである．[Winograd and Flores 86] がかつて人工知能について指摘したように，彼らの言葉が意味を持つためには，社会に対する完全な「コミットメント」，すなわち共通の運命を有するものとして私たちと共に生きることが必要なはずだ．

4. AGI に対する含意

AGI が実現するならば，それは道徳的な善悪や美的価値についての評価・判断もするものになるだろう．それがどのような方法でなされるかはわからない．本発表では主に記号的・論理的 AI のアプローチによってなされる人工道徳の問題点を指摘したが，しかしここでの批判の一部はより一般的な射程を持っている．ニューラルネットを用いたものであれ，ウェブ上のビッグ・データの検索に基づいたものであれ，そういったまったく新しいプレイヤーに道徳的価値判断を下す権威を委ねる，そしてそのような慣習が社会に根付くということは，私たちの道徳的実践に劇的な変化をもたらすだろう．少なくとも私たちはそのことに対する警戒をしておかなければならない．

参考文献

[Allen, et. al. 00] Allen, C., Varner G., & Zinser, J.: “Prolegomena to any future artificial moral agent.” *Journal of Experimental and Theoretical Artificial Intelligence*, 12: 251-261 (2000) .

[Anderson and Anderson 11] Anderson, M. and Anderson, S. L. (eds.): *Machine Ethics*, Cambridge University Press, (2011).

[Anderson et. al.] Anderson, M. and Anderson, S. L., and Armen, C.: “Toward machine ethics,” American Association for Artificial Intelligence, http://aaaipress.org/Papers/Workshops/2004/WS-04-02/WS04-02-008.pdf (2004)

[Anderson 11] Anderson, S. L.: “Machine metaethics.” In [Anderson and Anderson 11], 21-27 (2011).

[Arkin 09] Arkin, R. C.: *Governing Lethal Behavior in Autonomous Robots*, Chapman and Hall/CRC (2009).

[Beavers 11] Beavers, A. F.: “Moral machines and the threat of ethical nihilism. ’’ In P. Lin, K. Abney and G. A. Bekey (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics*, The MIT Press, 333-344 (2011).

[Gunkel 12] Gunkel, D.: *The Machine Question: Critical Perspective on AI, Robots, and Ethic*, The MIT Press (2012).

[Kukita 13] Kukita, M.: “Can robots understand values?: Artificial morality and ethical symbol grounding,” *Proceedings of 4th International Conference on Applied Ethics and Applied Philosophy in East Asia*, 65-76 (2013).

[Nadeau 06] Nadeau, J. E.: “Only androids can be ethical.” In Ford, K. F. and Glymour, C. and Hayes, P. J. (eds.) *Thinking about Android Epistemology*, IAAA Press, 241-248 (2006).

[Sparrow 07] Sparrow, R.: “Killer robots,” *Journal of Applied Philosophy*, 24(1), 62-77 (2007).

[Turkle 12] Turkle, S.: *Alone Together: Why We Expect More from Technology and Less from Each Other*, Basic Books (2012).

[Wallach and Allen 09] Wallach, W. and Collin, A.: *Moral Machines: Teaching Robots Right from Wrong*, Oxfort University Press, (2009).

[Winograd and Flores 86] Winograd, T. and Flores, F.: 『コンピュータと認知を理解する—人工知能の限界と新しい設計理念』，平賀譲訳，産業図書（1986）．

*1 産総研の柴田崇徳が開発しているアザラシ型のセラピーロボット．

*2 「倫理的記号接地問題」については拙論 [Kukita 13] を参照されたい．