The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

倫理的行動を促進するAIを社会で活用するための課題

Challenges for AI which promotes humans' ethical behaviors

福原 慶子[*1]
Keiko FUKUHARA

[*1] 名古屋大学大学院情報学研究科
Graduate School of Informatics, Nagoya University

In this paper, I consider challenges we would be faced with if an ethical Artificial Intelligence (AI) suggests ethical behaviors for humans. When a human agent follows an ethical AI and takes any action suggested by it, is the behavior or the agent ethical? Through tackling on this question, we will advocate a view of the relationship between humans and an ethical AI in which they are not separated a subject and an object but together constitute an agent as alliance. If we can design an ethical AI that not only teaches humans ethics but also learns it from humans, it would promote the more ethically developed society.

1. はじめに

1.1 背景

ソーシャルメディアなどインターネット上での反社会的な発言を、人工知能(AI)を使って発見しようとする研究がおこなわれている。たとえば、東フィンランド大学ではヘイトスピーチや暴力を示唆する書き込みを見つける機械学習モデルを考案している[Science Daily 2017]。国内では、クーロン株式会社が機械学習・自然言語処理・行動分析に基づいた AI が、書き込みの内容を自動的に分析し、必要と判断すればフィルタリングを施す「QuACS（クアックス）」というシステムを提供している [クーロン 2015]。

これらのように、AI を使って人々の言動の倫理的向上を目指す動きが国内外で進められる。個人には言論の自由が認められているとはいえ、度を越した差別や罵倒のヘイト表現をインターネット上でむやみに表明することは、やはり社会通念上好ましくない。このような書き込みによってユーザが精神的苦痛を被ったり他のユーザとの対立やトラブルへと発展してしまったりということは今や日常的に起きている。そのようなことが起きたときのサービス提供者の責任を考えると、ユーザの行き過ぎた放埒な振る舞いを律することは望ましく思われる。

だが一方で、AI によって行動が示唆されること、特にそれが倫理的な判断を伴っている場合の検討や議論はまだ途上である。倫理的な判断をどう実装するかといった技術面だけではなく、社会的・倫理的側面からも十分検討する必要がある。

1.2 本稿の目的

本稿では、人々の倫理的行動を促進するような AI を今後社会的に広く利用していく場合の課題について考察する。「人々の倫理的行動を促進する AI」と一口で言っても、その実現は方法も程度もさまざまだが、ここでは、我々人間の行動について倫理的に善いとみなされる行動を提案してくれるようなシステムを考える。たとえば SNS に投稿しようとした際に、それがヘイトスピーチに該当するかを判断して投稿を却下したり修正を要求したりするようなシステムがこれに相当する。悪いことを防ぐだけではなく、逆に善い行いを推奨するような応用も考えられるだろう。これ以降、本稿ではこのような機能を持つ AI を「倫理 AI」と表記する。

本稿では、倫理 AI を社会で使うためには何が必要かを検討し、人間と倫理 AI が一体となって倫理的行為をしているという考えを強調する。さらに、両者が相互作用的な関係であることで、より倫理的に発展していく社会を展望する。

1.3 本稿の構成

主張をするために、まず既存の倫理理論では限界があるということを示す(2 節)。次に、この限界を乗り越えるため、ピーター=ポール・フェルベークが支持する、人間と人工物が一体で道徳的行為を成すという主張を説明する(3 節)。最後に、この主張を適用したことから受ける利点と社会への展望について述べる(4 節)。

2. 既存の倫理理論の限界

倫理 AI がさらに開発され、インターネット上のより多くの場面で使われるようになると、悪意あるインターネットユーザによる悪辣な発言が抑えられていくことが予想される。社会全体としては、善い方向へと進んでいるように思えるが、現在でも既に、ビッグデータを AI で解析して、ユーザにとって最適と思われる商品や就職先や結婚相手を示すことに対して、自由が奪われていると感じる人もいる（[デュガン他 2017],[山本 2017]など）。さらに、内部のアルゴリズムがわからなければ、倫理 AI の出力結果に不信感を持つだろう。

こうした不信感の元では、倫理 AI に従っての行為は、自由な行動ではなくただ押し付けられていると考えられるだろう。この行為や行為者は倫理的行為・倫理的行為者ではないと考えるかもしれない。

そうした人々は、倫理学の一つの立場である義務論的アプローチから倫理 AI を評価しているだろう。このアプローチでは、行為の動機などの行為者の内面状態から、道徳的善悪が判断される[品川 2015, p.80]。義務論の代表的論者であるイマヌエル・カントによると、道徳的に振舞うためには、その行為者が自分で普遍的道徳法則を思い浮かべそれに従って行為を決定する自律性が必要である。そしてこうした自律性を持てることに、人間の尊厳があるという。逆に、自分の理性を働かせずに、自身の欲望や、他人や神の指示・命令に従って生きるのは他律である。

連絡先：福原慶子, fukuhara.keiko@g.mbox.nagoya-u.ac.jp

また別の立場である帰結主義では、倫理 AI に従っての行動は、結果が善ければ倫理的に善い行為であると結論する。なぜなら、このアプローチは義務論とは反対に、行為の善悪をその行為がもたらした結果から判定するためである[品川 2015, p.107]。倫理 AI に提示された行動は善いことでありそうすべきだという理解が行為者にはなかったにせよ、行動の結果が善にむすびついた、というところに着目する。

倫理 AI の指示に従った行為は倫理的に善いという考えを推し進めていくと、究極的には、行為者自身が倫理について何も考えなくなっても問題ないということになる。その状態にまでなってしまうと、ただ行動を倫理 AI に押し付けられ、強制されているだけになる。上に示したようなそれは自由な行動ではないと感じる人からは、倫理 AI を利用することへの理解がますます得られないだろう。

倫理 AI からの単なる押し付けを避けるために、複数の選択肢を提示させ、ユーザが自分の意思で行動を選択できるようにしようと設計者は考えるかもしれない。たとえば、SNS に投稿しようとした内容が、倫理 AI によってヘイトスピーチと判断されたとしよう。そのまま倫理 AI が投稿を禁止してしまえば、それは他者からの押し付けであるし、投稿者自身の自律性が侵害されている。だが、もし、投稿内容の修正あるいは取りやめをするか、それともそのまま投稿するかをユーザが選択することができるようになったらどうだろうか。これならば投稿者の意思を反映する余地がある。選択の見せ方として、[セイラー他 2009]で考えられているような、デフォルト状態を一番望ましいものにしておくというデザインも適用されるかもしれない。デフォルトでの選択はあるが、それが変更できないわけではない。選択肢として他の行為もまた、行為者自身の意思で選べる。

だが、完全に放任でも強制でもないリバタリアン・パターナリスティックなシステムやデザインを採用したとしても、与えられたものの中から選ぶというのはやはり倫理 AI、あるいはそのデザイナーやプログラマにより与えられた中での他律的な行動でしかない。

以上のように、倫理 AI に従っての行為あるいは従った行為者は倫理的であるかという問いを義務論・帰結主義からとらえようとすると、それぞれ倫理的ではない・倫理的であるという反対の答えを出す。そしてこの枠組みでとらえているうちは、我々は倫理 AI に従うか拒否するかのどちらかを選ぶことになる。

3. 人工物の道徳

ピーター=ポール・フェルベークの分析では、これまでの倫理学の主要な立場は、人間が主体であり、世界の事物が客体であるという主体と客体の二分法に基づいている。カントのような義務論は道徳的行為の主体に、帰結主義は行為の客体にそれぞれ焦点を当てている[フェルベーク 2015]。

フェルベークは、技術の倫理的問題はこうした主客二分法によっていては適切に扱えないと批判し、「倫理は人間-技術連合体の問題として考えられるべき[フェルベーク 2015, p.26]」という主張を擁護する。なぜなら人工物は客体にとどまらず、ときに人間の道徳的判断や行動に介入してくるからである。たとえば学校付近の道路や駐車場の敷地内にあるスピードバンプは、運転手にここでは減速しなければならないということを思い出させている。

同様に、倫理 AI も人工物の一つであり、人間と倫理 AI は、それぞれが主体と客体の関係ではなく、一緒になって道徳的行為をおこなう連合体であると考えてみよう。

倫理 AI は人間に行為を促すものであり、人間と一体になって道徳的行為者となる。ただし、倫理 AI はスピードバンプとは違い、機械学習によって促す行動を変化させる可能性がある。その可能性を広げるために、人間と倫理 AI が互いの倫理性を高め合うという関係を次に考える。

4. 利点と展望

我々人間は、最初から道徳や倫理を知っていたり、誰の手も借りずに自力で身につけたりしているわけではない。自律的に道徳的な行動をするには、外部に頼った訓練がある程度は必要である。小学校で道徳の教科書を読んだり、悪戯をして大人に怒られたり、浅慮な言動で自分自身がひどい目に遭ったり、といった経験を重ねて行為の善悪を学ぶ。そして理解して内在化していき倫理的行為者となるものである。

Microsoft が開発したチャットボット Tay は、言葉を教えてもらったが、それが倫理的に善いかどうかまでは教えられなかった。このシステムは Twitter を通じて機械学習で言葉を学習し、ユーザとコミュニケーションを図る目的で開発された。話しかけられた内容を繰り返すという方法で学習していったが、差別的な内容や汚い言葉遣いばかり覚えていき、すぐにそのアカウントは停止された。

これに陥らないためには、倫理 AI にも人間のほうから倫理を教えることができるインタフェースデザインも必要になる。人間と倫理 AI とが相互作用でき、倫理 AI が作り上げる倫理体系を永続化・固定化せず、変化可能にする。ただしそれは、恣意的に結果を変えるという意味ではない。倫理を教えるにあたり、設計者やプログラマのバイアスも入り込むことがありうるし、偏った教え方では結局 Tay の時と同じようになってしまう。したがって、倫理 AI に関する課題は、一部の技術者や有識者だけでなく、このシステムに関わる全ての人の関心事へと変えていかなければならない。

倫理が可変的であるということは、行動の指針としようとした場合に頼りないものに思われるかもしれない。だが、ここには一つ利点がある。社会が変われば倫理もまた変わるものである。倫理をアップデート可能にしておけるということは、いつまでも旧態依然とした悪い意味で伝統に縛られた価値観からの脱却をいくらか容易にするかもしれない。政治への参加、性別や人種に関わらない平等など、人類が自由を獲得してきた道筋のひとつは、そのようなものであった。もし、倫理 AI と人間の相互作用がうまくいったとしたら、倫理的善の発展を一層促す社会を構想することができるだろう。

5. まとめ

本稿では、倫理 AI を社会で利用するための課題を検討した。人間が主体であり人工物は客体であるという主客二分法を超え、人間と倫理 AI が一体となり倫理的行為者になるという考えを強調した。さらに、こう考えることで、人間と倫理 AI が互いに倫理を教えあうという相互作用を生み、善をより促進する社会を展望した。

参考文献

[デュガン他 2017] マルク・デュガン、クリストフ・ラベ、鳥取絹子訳:『ビッグデータという独裁者:「便利」とひきかえに「自由」を奪う』, 筑摩書房, 2017.

[セイラー他 2009] リチャード・セイラー, キャス・サンスティーン著、遠藤真美訳:『実践行動経済学：健康、富、幸福への聡明な選択』, 日経 BP 社 , 2009.

[フェルベーク 2015] ピーター=ポール・フェルベーク, 鈴木俊洋訳:『技術の道徳化：事物の道徳性を理解し設計する』, 法政大学出版局 , 2015.

[品川 2015] 品川哲彦:『倫理学の話』, ナカニシヤ出版, 2015.

[山本 2017] 山本龍彦:『おそろしいビッグデータ 超類型化 AI 社会のリスク』, 朝日新聞出版, 2017.

[クーロン 2015] クーロン株式会社:プレスリリース, https://www.quelon.co.jp/release_20150707_01/

[Science Daily 2017] Science Daily : New machine learning models can detect hate speech, violence from texts, https://www.sciencedaily.com/releases/2017/04/170412091222.htm.