明治学院大学

AIと倫理

櫻　井　成一朗

1．はじめに

AI社会を迎えようとしている二十一世紀，AIスピーカーや自動翻訳機等，AIが人々の日常生活に浸透しつつある。日本政府は，2019年3月29日に人間中心のAI社会原則（https://www8.cao.go.jp/cstp/aigensoku.pdf）を策定した。人間中心のAI社会原則では，7つの原則を定め，「AIを効率的かつ安心して社会実装するため，AIに係る品質や信頼性の確認に係る手法，AIで活用されるデータの効率的な収集・整備手法，AIの開発・テスト・運用の方法論等のAI工学を確立するとともに，倫理的側面，経済的側面など幅広い学問の確立及び発展が推進されなければならない」とある。AIの倫理的側面について，2019年に西垣通名誉教授と川島茂生准教授が「AI倫理」を上梓した。西垣名誉教授のいうAI倫理とは，AIが遵守すべき規範を定めるものではなく，社会で人間の共感に基づき社会規範をつくるという倫理思想である。本論文では，このAI倫理を踏まえて，道具としてのAIを利用するものの倫理について考察する。

2．AIの定義と倫理の必要性

2.1　強いAIと弱いAI

強いAIとは，サール[(1)]によって定義されたもので，自由意思を持つAIであり，まさに人間の代わりとなり得る存在のことをいい，強いAI以外のAIを弱

い AI という。弱い AI は，特化型 AI あるいは専用 AI とも呼ばれ，特定の領域に対してのみ機能する AI をいう。特化型 AI は，深層学習や強化学習等の機械学習技術の進歩により，特定の領域においては人間の能力を凌ぐようになったが，あくまでも道具に過ぎない AI である。しかしながら，特化型 AI は人間のような汎用的な知能ではないので，特定領域以外に対してはほぼ無力となるため，あらゆる場面に対応できる汎用知能をもつ AI として汎用 AI（Artificial General Intelligence）が活発に研究されるようになった。本論文では，汎用 AI が自由意思を持つとは限らないので，強い AI とは区別する。

特化型 AI をネットワークで緩く結合すれば，汎用 AI が実現されるという楽観的な考え方もある(2)が，大量の特化型 AI を緩く接続しただけでは，知能の汎用性は生まれないであろう。特化型 AI を膨大な専門領域，たとえば，医学，物理学，化学，法学，情報学，社会学，経済学等におよぶ，何百万あるいは何千万の特化型 AI を接続しただけでは，検索エンジンを通じて，適切な回答が得られるわけではなく，AI における未解決問題の一つである，フレーム問題(3)を解かねばならず，膨大な候補の中からどの特化型 AI を選択するのかは結局人間に委ねざるを得ないであろう。特定の特化型 AI を選択する作業こそが，むしろ知能の汎用性の源泉であるのであるから，それこそがこれから開発すべき，汎用 AI の核となる要素に他ならないのである。

コンピュータパワーの急激な拡大と機械学習技術，特に深層学習の進歩によって，汎用 AI の実現可能性は高まっていると見ることもできよう。仮に汎用 AI が実現できたとしても，汎用 AI において自由意思あるいは意識を実現することは更に困難が予想される。したがって，強い AI は実現不可能であるかもしれないことに注意が必要である。

2.2　自動運転と倫理

AI 社会における倫理の重要性について自動運転を例にして考察する。自動

運転の際にも取り上げられる代表的な倫理の問題として以下のトロリー問題（トロッコ問題）がある。

> *暴走した路面電車が線路の分岐点に近づいている。もしその線路のまま路面電車を走らせておくと，5 人の作業員が命を落とす。もし運転手が車両の進路をもう一つの支線に変えると，命を落とす作業員は一人だけで済む。この路面電車を運転しているのがあなただったら，あなたはどうするだろうか？この路面電車を運転しているのがコンピュータやロボットだったら，コンピュータやロボットはどうするだろうか？*[(4)]

トロリー問題は，そのまま自動運転自動車にも適用可能であるため，設定を僅かに変更した上で，AI がどう判断すべきかという問題として取り上げらることが多い。また，以下の日経新聞の AI の医療応用への記事でも，倫理について触れられている。

> *ただ人間の脳神経細胞（ニューロン）を数理モデル化した仕組みの研究開発などがさらに進めば，将来的には AI ドクターが誕生し，やがては人間の医師と同じ，もしくは超える存在になる可能性を秘めている。新薬開発を効率化し医薬品価格を引き下げる，医師や看護師ら医療従事者の負担を減らす，効率的な治療法を提唱し患者の生活の質（QOL）を向上させる―。その際は自動運転技術と同様に，AI ドクターが医療事故を起こした場合，誰が責任を取るのかなどの法的問題が出てくる。人間の命にかかわる判断を任せていいのかなどといった倫理問題についても，時間をかけて社会全体で丁寧に議論する必要があるだろう*[(5)]。

上記の記事では，自動運転技術だけでなく，医療現場におけるAIの誤判断の責任問題が提起されている。

AIが様々な場面で導入され，従来は人間にしか許容されていなかった判断，たとえば，人間の生命に関わるような判断もAIに委ねられることになれば，倫理面の検討が求められることになるのである。AIが自律的に行動しているように観察されれば，AIを行為主体としてみなす考え方も出てくるようになる。岡本裕一朗教授は，以下のようにAIを行為主体とみなしている。

> *動物は，自然的有機体として，人間から自律的に活動できます。それでも，知能的な面からいって，動物に倫理を要求することはありません。それに対して，人工知能は人間によって製作された機械であるにもかかわらず，今や行為主体と見なされ，倫理が要求されるようになったのです*(6)。

行為主体の必要条件については，次節で検討する。

3．AIと自由意思

3.1　意思と法システム

大屋雄裕教授は法制度と意思の関連について以下のように述べている。

> *だからこそ，責任もそのような自己決定から生じるものと理解されてきた。民事法における過失責任主義は，個人の意図的な選択の帰結（故意）あるいは理性的な存在者であればなし得たはずの注意を怠ったこと（過失）に対する責任のみを追及しようとするものであり，逆*

に言えばいかなる自律的存在にも防ぎ得ず（無過失），したがって自律的な選択の結果と考えられない事態は免責されてきた。刑事法の世界においても，応報刑論はまさに人格を有する個人を前提として，にもかかわらず悪と評価される行為をあえて選択したことを根拠として罪責を問う立場であった。これらの制度に一貫しているのは，〈意思―行為―責任〉という連関によってひとびとの織り成す関係を読み解こうとする姿勢である。意思をもたない存在の動作・意思に基づかない行動は自己決定の一部としての「行動」ではなく，それに対する責任を問う余地もまたないのである[(7)]。

もし自由意思をもつ強いAIが実現できれば，法的責任を問う余地が出てくることになるが，既に述べたように汎用AIの実現までには，まだ時間がかかると予想され，さらには自由意思を持つ強いAIは実現できないかもしれない。

ある存在がこれから実行する行為を，（推定できても）完全に予測できないことはその自由意思をもつことの必要条件である。そこで，こういうことになる。―ある存在が実行する行為の予測困難性という条件のもとで，その理論的自律性さらに自由意思がみとめられ，そのもとではじめて，責任を問われる道徳的主体が出現するのだ，と。そして，そういう道徳的主体だからこそ，実践的自律性をもつといえるのだ，と[(8)]。

西垣名誉教授によれば，プログラムされたAIの行為は予測可能であり，それゆえ自由意思を持ちえないということになる。

また脳科学者の茂木健一郎理学博士はマインドアップローディング[(9)]に関して，以下のように述べている。

もちろん，純粋に理論的な立場から言えば，将来コンピュータのメモリの中に人間の記憶が移植されることが原理的にあり得ないということではない。その一方で，そのような移植が，新しい「意識」の誕生を保証するわけでもない[(10)]*。*

すなわち，カーツワイルがいうところのマインドアップローディングができるようになったとしても，直ちに意識も移植できるわけではないということである。

3.2　意思と法的責任

大屋教授が指摘するように，現行法制度が〈意思―行為―責任〉という連関に依拠している以上，意思をもたない人工知能は責任主体とはなりえない。すなわち，AIが弱いAIであり続ける限り，AIがどんなに賢く振舞ったとしても道具に過ぎないということを忘れてはならない。したがって，道具を製造したものに対する責任や道具を使ったものに対する責任を問うことはできても，AIに対する責任を追及することはできない。しかしながら，AIが自然言語を扱うようになると，疑似人格を認めてしまうような人も出てくるであろう。大屋教授も以下のように述べている。

もちろん我々はすでにAIが言語によって我々の問いかけに答え，ロボットが我々と会話する社会に生きているのであった。言語行為こそが我ら人類の社会に参入する資格を形作るのだとすれば，ロボット・AIの場所はその内部に求められるはずだという意見も，当然ながら現れてくるだろう。

深層学習の登場により，機械翻訳の品質や音声認識の精度が格段に向上し，

現実に AI スピーカーでは自然な会話が行われている。1960 年代に作られた対話プログラム Eliza でさえ，多くの人がコンピュータプログラムであるとは信じられなかったが，現代の AI スピーカーは，インターネットに接続されたことで，より豊富な情報に基づいて応答するので，AI スピーカーを疑似人格として捉えてしまう人も少なくないであろう。これについては，西垣名誉教授が以下のように述べている。

> *だが，もともとコンピュータは，生物とはちがって身体的に世界の意味を解釈する存在ではないから，どうしても限界がある。定形的な応答程度の出力がせいぜいなのだ。機械翻訳にしても，簡単な日常的文章ならともかく，小説のような複雑で非定型な文章には歯が立たないのである。これは「記号接地問題（symbol grounding problem）[(11)]」と呼ばれ，AI の根本的難題として位置づけられている。*

現在の自然言語処理は意味を扱わないという西垣名誉教授の指摘は否定できないが，自然言語の単語や文章を word2vec[9]や word2vec を文章に拡張した doc2vec によりベクトル表現に変換する手法が注目を浴びている。これらの手法により直ちに意味解釈が可能になるとまではもちろん言えないが，深層学習の特徴を発見する機能を活用することで，より人間のもつ意味に接近できる可能性はあるだろう。そうなれば，AI の応答がさらに人間に近づいていくことが予想できるのである。

4．求められる倫理

4.1　西垣名誉教授のAI倫理

西垣名誉教授は，現在のAIが自由意思をもたないが，疑似人格をもつように見える存在としてのAIを前提としてAI倫理[3]を提唱している。西垣名誉教授の提案するAI倫理は，AIがその動作中に遵守すべき倫理規範のみを指すのではない。人間の共同体である社会を個人の心と同じようにオートポイエティックな閉鎖システムとみなし，「社会規範／行動／道徳観」の三項関係からとらえられることになる。

> *社会規範の生成をもたらす原動力は，社会を構成するメンバーが，自らの心的状況／利害得失だけでなく，他のメンバーの心的状況／利害得失にたいして想像力を働かせる「共感」の能力である。共感から，他者の人格の幸福や尊厳をまもる社会規範を設定するという方向性がうまれる*[(12)]。

このように社会に参加する個人の共感のもとで，社会規範の生成や更新にAIを活用していくのである。そして，AI倫理とは以下のように社会で社会規範をつくるという倫理思想なのである。

> *人間は，それぞれ自分のやり方で世界を観察し，意味解釈し，他者に共感しつつ，取り換えのきかない固有の世界のなかで生きていく。その経験から，はじめて道徳観がうまれるのだ。だからこそ，そういう道徳観をもつ人間一人ひとりの人格を，かけがえのないものとして尊*

重せよという主張が正当性をもつ。そして，個々の道徳観をあわせ，集合知[13]として社会規範をつくっていかなければならないのである[14]。

次に，このような AI 倫理の下で，社会に参画する人々に必要とされることについて考える。

4.2 弱い AI と倫理

西垣名誉教授の AI 倫理においては，AI 導入に際して，AI が社会規範を遵守していることのチェックの重要性が以下のように述べられる。

社会規範を AI エージェントに守らせるように設計することは重要である。AI エージェントは指示通りに作動し，自ら道徳観を形成することはないから，その作動は社会規範に正確に沿うことになる。とはいえ，AI が自ら学習して AI を修正発展させていくプログラムの場合は AI の詳細な作動の予測は設計者にとってさえ容易ではなくなり，AI がつねに社会規範を遵守しつつ作動しているか否かをチェックすることは困難である[15]。

現在は，全脳エミュレーションのための素子レベルの開発が始まったばかりであるが，全脳エミュレーション[3]が実現されるようになると，西垣名誉教授が指摘するように，社会規範遵守のチェックはますます困難になる。幸いにも，全脳エミュレーションが実現されるまでには，多くの技術的課題が残されているので，実現に成功するとしてもまだ時間的余裕があると思われる。したがって，当面は AI 社会とは特化型 AI による AI 社会になることが予想される。そうすると，現時点で緊急性が予想されるのは，疑似人格とみなされる特化型 AI が悪用されることである。

西垣名誉教授のAI倫理においては，社会に参画する個々人が自らの自由意思に基づいて行動することが求められる。人々がAI社会に慣れてしまうことで，AIからの指示を鵜呑みにしてしまえば，真の自律的行動からかけ離れた行動をしてしまう危険性が生じてくる。道具に過ぎない特化型AIを悪用すれば，様々なことが可能になる。たとえば，人々がAI推薦システムに慣れていけば，AIの推薦を鵜呑みにして行動決定するようにならないだろうか。AIの指示のままに，PCをクリックするか，スマホをタップしてしまうことで，実は興味のない本や商品を購入してしまうことがないだろうか。本や商品を購入している程度であればまだ良いが，国の政策決定等の重要な意思決定までAIの指示のままに従うようになれば，国の政策を誤り兼ねない。また，AI化によって公正化や公平化が図れるような場合であっても，与える学習データを恣意的に選択してしまえば，逆に不公正や不公平な結果となってしまう危険性がある。道具であるAIを人が正しく使わなければ，AIは人を不幸にしてしまうのである。それゆえ，AIの道具としての危険性について，AIを使う経営者や為政者に対する倫理教育を行うことが重要となる。

社会規範を生成・修正するのは，社会に参画する人間だけなのであるから，経営者や為政者としてAIを使う人間は，AIの危険性を納得した上で，AIを使わなければならないし，説明を求められれば，説明できなければならない。また，末端利用者としてAIを利用するのであれば，AIの管理者が適正な学習データを用いていることを監視しなければならない。管理者が恣意的に学習データを与えれば，AIは末端利用者に不都合な判断をしてしまうからである。それゆえ，AI社会における人間は，管理者としても利用者としても自律的行動主体であらねばならないのである。

5．おわりに

本論文では，西垣名誉教授の分析を踏まえ，AI 社会で生きる人間の倫理について検討した。来るべき AI 社会における人間は，社会規範を生成・修正するのであるから，自律的行動主体として自らの行動に責任をもたなければならない。

汎用 AI[8]や強い AI についての倫理問題は，AI 自体が実現できない可能性もあるが，実現不能性が証明されたわけではないので，今後の課題としたい。

参考文献

[1]岡本裕一朗：人工知能に哲学を教えたら，SB 新書，2018。
[2]ロビン・ハンソン：全脳エミュレーションの時代，NTT 出版，2018。
[3]西垣通・川島茂生：AI 倫理，中公新書，2019。
[4]弥永真生・宍戸常寿編：ロボット・AI と法，有斐閣，2018。
[5]茂木健一郎：記憶の森を育てる　意識と人工知能，集英社 e 学芸単行本，2016。
[6]カーツワイル：シンギュラリティは近い（エッセンス版），NHK 出版，2016。
[7]W. ウォラック／C・アレン著岡本慎平／久木田水生訳：ロボットに倫理を教えるモラルマシーン，名古屋大学出版会，2019。
[8]人工知能学会監修：人工知能とは，近代科学社，2016。
[9]人工知能学会監修：深層学習，近代科学社，2015。

注

（1） Searle, John. R.（1980）Minds, brains, and programs. Behavioral and Brain Sciences 3(3): 417-457。
（2） たとえば，参考文献[1]参照。
（3） 解くべき問題に関連するものだけを選択することが実は非常に難しいという未解決の問題であり，人間にも解決できないと考えられている。
（4） 参考文献[7]，15 頁。
（5） 日経新聞朝刊，2019.11．23
（6） 参考文献[1]，53 頁

（7） 参考文献[4]，63頁。
（8） 参考文献[3]，57頁。
（9） 参考文献[6]で紹介された人間の記憶をコンピュータに移植するという技術のことである。
(10) 参考文献[5]，22頁。
(11) コンピュータ内部で処理される記号と現実の存在を結びつけるという問題。たとえば，猫の画像を分類できたことが，「猫」を学習できたとはならないということ。
(12) 参考文献[3]，122頁。
(13) 集団の意見や知識を集めると新たな知性を見出せるということ。
(14) 参考文献[3]，148頁。
(15) 参考文献[3]，124頁。