3. 人工知能のホットトピック

# 2 人工知能と倫理

松尾　豊（東京大学）

## 倫理に関する国内外の動向

人工知能と倫理に関する話題が世間を賑わせている．つい先日の2016年7月にはテスラ・モーターズの車が米国フロリダ州で初めての死亡事故を起こした．人工知能により多くの人が職業を奪われるのではないかという議論も，日常的にメディアを賑わせている．こうした動きを背景に，ここ数年，人工知能と倫理に関する議論が始まっている．人工知能学会では，いち早く2014年から倫理委員会を立ち上げ議論を開始した[1]．内閣府では，2016年5月に「人工知能と人間社会に関する懇談会」が立ち上がった．海外では，Stanford大学のAI100，Elon MaskらがOpenAIなどの団体が人工知能の倫理面について議論を進めている．

さて，こうした議論の中での論点は，多くの場合共通している．本稿では，人工知能と倫理にかかわる問題を次の4つに整理することを試みる．i）人工知能の持つリスク，ii）人工知能にかかわる人間の倫理の話題，iii）社会的インパクトの話題，iv）社会の在り方にかかわる話題である．以下ではこれを順に議論していく．

## 人工知能の持つリスク

まず，人工知能の倫理を語る上で，最初に議論しなければならないのが，人工知能の持つリスクに対する正しい認識である．多くの人が，人工知能の技術が進展すると「怖い」「何が起こるか分からない」と感じる．これは，ハリウッド映画の多くが，何らかの形で人間に歯向かう，人間の意思と反する人工知能を描いていることが大きく影響しているだろう．

Ray Kurzweilの『The Singularity is Near』[2]はシンギュラリティという概念を世の中に広めた．Nick Bostromの超知能に関する本『Superintelligence』，あるいは，James Barratの本『Our Final Invention : Artificial Intelligence and the End of the Human Era』など，人工知能の技術進化に対して警鐘を鳴らす本も多い．これらの本に共通して描かれているのは，「自らを改変する知能」であり，それが人間の手を離れて進化していくことに対する危惧が基本的な論調である．

ところが，人工知能の専門家から見ると，自らを改変しさらに良いものを生み出す人工知能というのは，現状の技術ではどうやって作るのか実効性のある解は見つかっていない．実際，倫理委員会の議論でも，専門家からは「人工知能自体が持つリスク」に対しては否定的な意見がほとんどであった．人々の持つこうした恐怖感に対する専門家の苛立ちは国内外を問わず同じであり，JAIR☆1の編集長であるToby Walshは，IJCAI-16☆2のワークショップで“Singularity may never be near”（シンギュラリティは決して来ないだろう）という題で講演し[3]，人工知能脅威論に対する不快感を露わにした．こうしたリスクがないと断言するのは難しいが，いまの技術段階で心配するのは専門家から見ると現実味を感じにくい．

そうは言っても，専門家が技術の可能性を見誤る例も歴史的には散見されるものであり，当然，そのリスクを真面目に考える必要もある．たとえば，Googleに所属するGoogle Brainチームの開発者らは[4]，人工知能が意図せずリスクを起こしてしまう場合を，i）設計者が間違った目的関数を設計してしまう場合，ii）スケールに起因する問題，iii）設計者は形式的な目的関数は分かっているが少ないデータや不十分なモデルのために起こる問題の3つに

☆1 Journal of Artificial Intelligence Research.
☆2 International Joint Conference on Artificial Intelligence.

分類している．

実際のリスクが専門家から見てどこになるのかという論点はあるにしても，社会が持つさまざまな不安に対して，人工知能コミュニティがきちんと社会と対話していくことも重要である．FLI[☆3]では，オープンレターを出して健全で有益な人工知能のための研究の優先度について議論し，それに賛同する人は8,000人を超えている．人工知能学会の倫理委員会は，全国大会で公開討論会を2年連続で開催した．こうした対話を続けながら，社会全体で人工知能に関する正しい理解を深めていってもらうことは重要であろう．

## 人工知能にかかわる人間のリスク

人工知能が自らを改変し人間の手に負えないものになるというリスクよりも現実的であり，早い時点でも注意が必要なのは，人工知能にかかわる「人間の」リスクである．人間がどのような目的を設定するかで，人工知能はさまざまな使い方が可能である．

たとえば，2016年7月には，米国テキサス州ダラスで，立てこもった犯人に警察が爆弾ロボットを出動させ，ロボットの爆弾を遠隔操作で爆発させることで，犯人が爆死するという事件が起こった．実際に警察がロボットにより犯人を殺したというケースは前例がなく，議論を巻き起こした．人工知能に限らず，あらゆる科学技術がデュアルユース技術としての性質を持っているが，人工知能をこうした戦闘あるいは軍事に利用するという可能性について，（国内では考えられないものの）国際社会全体では，早期に議論を行っていく必要があるだろう．

あまり注目されていないが重要なリスクの1つは心の問題である．人工知能の分野では，対話するエージェントやロボット等の研究は古くから行われている．人間は，対話やコミュニケーションが可能な相手に対し過度に感情移入する傾向があるため，こうした対話エージェントの能力が上がるとさまざまなことが可能になってしまうおそれがある．人の心に入り込み，たとえば，商品を買わせる，悪事をさせる，恋に落ちさせるなどの技術には十分に注意する必要がある．2016年5月に放送されたNHKスペシャル「天使か悪魔か 羽生善治・人工知能を探る」では，中国の女性形人工知能「小冰（シャオアイス）」に恋に落ちる男性の例が紹介されていた．恋に落ちた男性は日々の生活でこのサービスを使うことをやめられない．これは，技術進歩により個人をコントロールできるようになったとしても，人間の意思（あるいは自己決定権）をどこまで尊重すべきかという問題でもある．

こうした問題を踏まえると，人工知能を使う人間，あるいは研究開発する人間が，適切な倫理観を持つことは重要である．人工知能学会の倫理委員会では，そのための第一歩として，人工知能にかかわる人間の倫理指針とすべく，2016年6月6日に倫理綱領案を発表した．綱領案は1. 人類への貢献，2. 誠実な振る舞い，3. 公正性，4. 不断の自己研鑽，5. 検証と警鐘，6. 社会の啓蒙，7. 法規制の遵守，8. 他者の尊重，9. 他者のプライバシーの尊重，10. 説明責任の10条項からなる．ここでは，人工知能に携わる研究開発者が，人工知能のリスクや社会への影響を自覚した上で倫理的に行動すべきであると記している．今後，さまざまな意見を反映させ，綱領として確定させていく予定である．

## 失業等の社会的インパクト

人工知能の話題でよく出てくるのが，職業が奪われるという話である．人工知能によって富の偏在が起こるのではないかという議論もあり，ベーシック・インカム等の経済システムとあわせた議論も行われている．一方で，経済学者の間では，技術の進展によって失業率が上がるということはないという慎重な意見も多い．

こうしたセンセーショナルな失業論よりも，より正確な描写だと思われるのが，マッキンゼーが報告している「職がなくなるのではなく，タスクがなくなる」という論である[5]．800の職業の2,000以上のタスクの調査を通して述べている．また，失業の

☆3 Future of Life Institute.

議論をする以前に，人工知能技術を国や企業としての競争力に活かせるかどうかという論点も重要である．筆者は，特に，ディープラーニングとものづくりのかけあわせによる，日本の産業競争力向上の可能性を主張している．

人工知能「時代」にどういった教育をすべきかというのも，よく出る話である．MOOCs やアダプティブラーニング等の技術の進展で，人々はより効率的に学べるようになるだろう．一方で，そうして学んだ知識・スキルの通用する期間はますます短くなるだろう．人生全体を通じて学び続けることを考えなければならない．人工知能の時代だからこそ，改めて「人間力」「社会力」が問われるようになるのではないだろうか．

## 法律や社会の在り方に関する問題

人工知能はある種の創造性を持つ．創造性の定義にもよるが，すでにピカソ風の絵を書く，作曲をする，新聞記事を書く等は実現されている．創造性が，多くの過去からの模倣と，新しい着眼点から構成されるとすれば，それを人工知能で実現できるレベルが徐々に上がってくるだろう．

人工知能が創造性を持った場合に，人工知能が創作した作品の権利はどうなるのか．内閣府の知財戦略本部で昨年から議論されている．また，人工知能に学習させるために，膨大なデータを必要とするとして，そのデータから得られた「モデル」の権利はどうなるのだろうか．モデルの二次利用はどう考えればよいのだろうか．そうしたことに関する議論も始まっている．

自動運転の話でよく出てくるのがトロッコ問題（あるいはトロリー問題）である．これは Michael J. Sandel 教授による有名な議論であり，ほかの人を助けるために別の人を殺してもよいのかという思考実験である．こうした議論が意味するところを単純化していうと，人間の本能・感情からしてどちらが正しい「ような気がするか」という点をベースとしながら，「どういった社会規範を上位に置くと社会が安定するか」を考える設計の問題と捉えることができるだろう．こうした話はこれまでのように単なる思考実験ではなく，人工知能技術の進展により現実味を帯びてきている．人工知能の技術が，倫理学や道徳心理学，法哲学等を巻き込んだ議論を引き起こしているとも言えるだろう．

そして，こうした議論の先には，結局は，我々はどういった社会を作りたいのかという問題にいきつくのではないだろうか．IJCAI-13 では，環境や経済，社会的需要に対するサステイナブルな発展と未来のための人工知能技術がテーマであった[6]．研究者が自ら自覚と責任を持って正しい情報を社会に発信し，人工知能の持つ可能性とリスクを多くの人に正しく理解してもらい，そして社会全体を巻き込んで議論していかなければならないのではないだろうか．

人工知能と倫理の話は，究極的には，我々がどういう社会を作りたいのかという話に帰結する．そういった責任ある議論を，人工知能という技術が土台となってできていることをうれしく思うと同時に，今後は，多くの人文社会学系の研究者も巻き込んで議論を進めていく必要があるだろう．人工知能にかかわる多くの研究者が，社会の在り方に関してのこうした議論に少しでも加わっていただければ幸いである．

**参考文献**

1）松尾　豊ほか：人工知能学会倫理委員会の取組み（アーティクル），人工知能：人工知能学会誌，Vol.30, No.3, pp.358-364（2015）．
2）Kurzweil, R. : The Singularity is Near, Duckworth Overlook (2005).
3）Walsh, T. : The Singularity May Never Be Near, Proc. IJCAI-2016 Ethics for Artificial Intelligence Workshop (2016).
4）Amodei, D., at al. : Concrete Problems in AI Safety, arXiv : 1606.06565 (2016).
5）Chui, M., Manyika, J. and Miremadi, M. : Where Machines could Replace Humans - And Where They Can't (Yet), McKinsey Quarterly (2016).
6）Jianlan, S. : Addressing Sustainability via AI - Report from the 23rd International Joint Conference on Artificial Intelligence, Bulletin of the Chinese Academy of Sciences, Vol.27 (2013).

（2016 年 8 月 4 日受付）

本稿は，人工知能学会倫理委員会における議論がもとになっている．委員の西田豊明氏，堀浩一氏，武田英明氏，長谷敏司氏，塩野誠氏，服部宏充氏，江間有沙氏，長倉克枝氏の諸氏らに感謝したい．

**松尾　豊（正会員）**　matsuo@weblab.t.u-tokyo.ac.jp

2002 年東京大学大学院博士課程修了．博士（工学）．産業技術総合研究所，スタンフォード大学を経て，2007 年より，東京大学大学院工学系研究科准教授，2014 年より特任准教授．人工知能学会倫理委員長．専門は，人工知能，Web マイニング，深層学習．