特　集「人工知能学会・情報処理学会共同企画―第３部「技術紹介」―」

人工知能と倫理

Artificial Intelligence and Ethics

松尾　豊 東京大学大学院工学系研究科
Yutaka Matsuo Graduate School of Engineering, The University of Tokyo.
matsuo@weblab.t.u-tokyo.ac.jp

西田　豊明 京都大学大学院情報学研究科
Toyoaki Nishida Graduate School of Informatics, Kyoto University.
nishida@i.kyoto-u.ac.jp

堀　浩一 東京大学大学院工学系研究科
Koichi Hori Graduate School of Engineering, The University of Tokyo.
hori@computer.org

武田　英明 国立情報学研究所情報学プリンシプル研究系
Hideaki Takeda Principles of Informatics Research Division, National Institute of Informatics.
takeda@nii.ac.jp

長谷　敏司 SF・ファンタジー小説家
Satoshi Hase Science Fiction / Fantasy Writer.
haseo@white.plala.or.jp

塩野　誠 株式会社経営共創基盤（IGPI）
Makoto Shiono Industrial Growth Platform, Inc.
m.shiono@igpi.co.jp

服部　宏充 立命館大学情報理工学部
Hiromitsu Hattori College of Information Science and Engineering, Ritsumeikan University.
hatto@fc.ritsumei.ac.jp

江間　有沙 東京大学教養学部附属教養教育高度化機構
Arisa Ema College of Arts and Sciences, The University of Tokyo.
cema@mail.ecc.u-tokyo.ac.jp

長倉　克枝 科学ライター
Katsue Nagakura Science Writer.
katsue.nagakura@gmail.com

1. は　じ　め　に

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，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．3 月に，マイクロソフトのチャットボット「Tay」がヒトラーを礼賛したと話題になったことも記憶に新しい．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．本学会では，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 [松尾 15a]．総務省の総務政策研究所が主体となった会合[*1] でも 2015 年から同様の議論が行われ，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．国際的には，スタンフォード大学の AI100[*2]，テスラ・モーターズの CEO である Elon Mask が支援している FLI[*3]，Friendly AI で有名な Eliezer Yudkowsky が創設した MIRI[*4]，Elon Mask や Peter Thiel[*5]，Sam Altman[*6] らが創設しディープラーニングの有力な研究者も所属する OpenAI などの団体，ケンブリッジ大学に

＊1 総務省，AI ネットワーク化検討会議
＊2 One Hundred Year Study on Artificial Intelligence
＊3 Future of Life Institute
＊4 Machine Intelligence Research Institute
＊5 PayPal の創業者で，『ゼロ・トゥ・ワン―君はゼロから何を生み出せるか』の著者としても有名である．
＊6 Y コンビネータというスタットアップのインキュベータを立ち上げたことで有名．

できたCSER[*7]が人工知能の倫理面について議論を行っている．

さて，こうした議論のなかでの論点は，多くの場合共通している．本稿では，人工知能と倫理に関わる問題を次の四つに整理することを試みる．i）人工知能のもつリスク，シンギュラリティの捉え方，あるいは人々の感じ方に関する話題，ii）人工知能を利用あるいは研究開発する人間の倫理に関しての話題，iii）人工知能に関する職業と教育などの社会的インパクトの話題，iv）人工知能に関する知財や権利などの法律，あるいは倫理規範や社会の在り方に関わる話題である．以下では，これを順に議論していく[*8]．

2. 人工知能のもつリスク

まず，人工知能の倫理を語るうえで，最初に議論しなければならないのが，人工知能のも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」，「何が起こるかわからない」と感じる．これは，ハリウッド映画の多く（例えば「ターミネーター」や「2001年宇宙の旅」，最近でも「トランセンデンス」や「エクス・マキナ」など）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．人間が他の動物に対してもつ優位性は，知的能力であるから，これまで競争優位だった点で自らを超えるものに対して危惧を覚えるのも自然かもしれない．しかし，例えば，計算の能力で機械が人間を上回ったのはとうの昔であるし，最近では，インターネットにより知識の量でも人間を圧倒している．機械が部分的に人間を超えるということはすでに起こっている．

Ray Kurzweilの「シンギュラリティは近い」（英題“Singularity is near”）[Kurzweil 05]はシンギュラリティの概念を広めた．シンギュラリティの定義にはさまざまなものがあるが，Kurzweilによると，遺伝子，ナノテクノロジー，人工知能を含むロボットなどの技術が指数関数的に発展し，特異点を境に急激な進展をすることであり，2045年頃に起こると予想している[*9]．また，Nick Bostromの超知能に関する本[Bostrom 14]，あるいは，テレビプロデューサであるJames Barratの書いた人工知能が人類の終焉をもたらすという本[Barrat 13]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AIというのは，現状の技術ではどうやってつく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「人工知能自体がもつリスク」に対しては否定的な意見がほとんどであった．人々のも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[*10]の編集長であるToby Walshは，IJCAI 2016[*11]のワークショップで“Singularity may never be near”（シンギュラリティはけっして来ないだろう）という題で講演し[Walsh 16]，人工知能脅威論に対する不快感を露わにした．ディープラーニング研究を先導するニューヨーク大学のYann LeCunやモントリオール大学のYoshua Bengioらも，ICML 2015[*12]のワークショップでこうした議論に辟易していると発言し，特に，「生存の欲求や他人を支配する可能性といった，進化に由来する人間の性質と，知能を混同しているように感じる．機械はそうした人間の性質はもたない」という趣旨の発言もしている[*13]．Baidu研究所所長のAndrew Ngは，インタビューに答え「こうした心配は，火星に移住した結果，火星の人口爆発を心配するようなものだ」と述べている[*14]．つまり，こうしたリスクがないと断言するのは難しいが（悪魔の証明であり困難である），今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．例えば，シンギュラリティ大学を卒業したFederico Pistonoは，「邪悪な人工知能のつくり方」を論文にまとめた[Pistono 16]．セキュリティの研究において攻撃側の研究をすることが重要であるのと同様，邪悪な人工知能をつくる方法を研究することが防御につながるという論旨である．そうした邪悪な人工知能のタイプを，開発者の故意である場合，ミスである場合，環境がそうさせる場合，人工知能のもつ学習によりそうなる場合などに分けて議論をした．そうした邪悪な人工知能をつくるかもしれない主体として，政府や

＊7 Center for the Study of Existential Risk（CSER）．リーバヒュームトラストからの助成金により，人工知能が人類の未来に与える影響を研究する．

＊8 なお，本稿の内容は，倫理委員会で行われた議論がベースになっており，本稿は，倫理委員会としてほぼ合意された意見を全員で表明するものである．

＊9 中川裕志氏による以下のまとめも参照されたい．http://www.slideshare.net/hirsoshnakagawa3/ss-64701276

＊10 Journal of Artificial Intelligence Research．人工知能における著名な論文誌．

＊11 International Joint Conference on Artificial Intelligence．人工知能における著名な国際会議．

＊12 International Conference on Machine Intelligence

＊13 Deep Learning Workshop．概要が以下に掲載されている．http://deeplearning.net/2015/07/13/a-brief-summary-of-the-panel-discussion-at-dl-workshop-icml-2015/

＊14 Artificial intelligence imagine the worst to prepare for the worst, http://marketbusinessnews.com/artificial-intelligence-imagine-worst-prepare-worst/135819

軍，企業などをあげ，邪悪な人工知能が取るかもしれないアクションについても列挙している．この論文は話題になったが，やや受けを狙いにいっているきらいもある．

それに対して，Google に所属する Google Brain チームの開発者らは，もう少しまじめに分析をしており [Amodei 16]，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（ネガティブな副作用がある場合[*15]と，報酬のハックを行うことで安易だが望まない結果を生んでしまう場合にさらに分けられる），ii）スケールに起因する問題で，設計者は目的関数をよくわかっているが，その評価にコストがかかるので少ないデータから外挿せざるを得ないために起こる問題，iii）設計者は形式的な目的関数はわ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（強化学習のエージェントが不用意な探索的行動を行ってしまう場合と，学習データにないために「悪い判断」を行ってしまう場合にさらに分けられる）と議論している．また，2016 年 6 月には，Google が人工知能に「非常停止ボタン」を付けたとして報道された[*16]．その内容は，強化学習の際に，どんな学習をしても割込みを回避しないようにする技術の研究を Google DeepMind の研究者が行ったということである [Orseau 16]．いずれも，人工知能の専門家から見ても，「確かに危ないことが起こり得る」と納得感のある内容である．

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，社会がも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．人工知能研究者の果たすべき役割としては，技術に対しての理解を促進する努力をし，その可能性やリスクの表明に対して誠実であることであろう．そして，社会全体では，こうした情報を一つの手掛かりとして，法律の問題や倫理，社会制度の問題などに取り組んでいかなければならない．

FLI 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 [Russel 15]，それに賛同する人は 8 000 人を超えている．そのなかでは，頑健な人工知能のためには検証（verification），妥当性（validity），セキュリティ，コントロールの四つが重要であると述べている．本学会倫理委員会は，全国大会で公開討論会を 2 年連続で開催した [松尾 15b，江間 16]．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．

3. 人工知能に関わる人間のリスク

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，人工知能に関わる「人間の」リスクである．人工知能に人間がどのような目的を設定するかで，さまざまな使い方が可能である．

例えば，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．米軍がイラクなどで使う爆弾処理ロボットに爆弾を装備したもので，自律行動ではなくリモコンなので人工知能というべきでないかもしれないが，実際に警察がロボットによって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もっているが，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．こうした中で，日本は「人工知能平和利用の国」という国際的な立場を築くのも一つの戦略かもしれない．

あまり注目されていないが，重要なリスクの一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など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[*17]，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，例えば，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．例えば，ある人の情報を網羅的に調べることで，ある商品を特定のやり方で提示すれば絶対に買うことがわかっている．これを提示してもよいのか．2016 年 5 月に放送された NHK スペシャル[*18]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，あるいは研究開発をする人間が，（どのような価値観をもつべきかという議論はまだ難しいとしても）少なくとも

＊15 Nick Bostrom は，クリップの生産を最大化するために人間までクリップの材料にしてしまう「クリップ・マキシマイザー」というシナリオで同じことを表現している．

＊16 米 Google，AI に「非常停止ボタン」暴走防止，日本経済新聞（2016 年 6 月 8 日掲載）

＊17 ソニーのロボット犬 AIBO のお葬式が行われている（AIBO の「お葬式」…解体・再利用へ，71 体を供養，朝日新聞（2015 年 11 月 19 日掲載））．古くは，対話システムの Eliza（1967）に人々は没頭していた．

＊18 NHK スペシャル「天使か悪魔か　羽生善治・人工知能を探る」（2016 年 5 月 15 日放映）

適切な倫理観をもつことは重要である．本学会倫理委員会では，そのための第一歩として，人工知能に関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*19．綱領案は

1．人類への貢献，2．誠実な振舞い，3．公正性，4．不断の自己研さん，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 [江間 16]．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うえ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．

4. 失業など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．オックスフォード大学の研究者が今後10年でなくなる仕事が約半数であるという論文を発表し [Frey 13]，また日本では野村総合研究所が2015年に同様の調査を発表し [野村 15]，大きな話題となった．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・インカムなどの経済システムと合わせた議論も行われている [井上 16]．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 [若田部 16]．

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*20がなくなる」という論である [Chui 16]．自動改札機ができても，従来，改札で切符を切っていた駅員さんという職はなくならず，その仕事の内容が変わったように，タスクがなくなることによって仕事の内容が再定義されるということを，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*21．また，失業の議論をする以前に，人工知能技術を国や企業としての競争力に生かせるかどうかという論点も重要である．経済成長する国の中での失業の話をするのと，経済的に停滞する国の中での失業の話をするのは大きな違いである．前者であれば，所得の再分配を行うためのさまざまな政策オプションが可能になる．本稿の著者の一人である松尾は，特に，ディープラーニングとものづくりの掛け合わせによる，日本の産業競争力向上の可能性を主張している [松尾 15c]．

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などの技術の進展で，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*22（これは人工知能の問題というよりは，イノベーションの進展の速度の問題である）．これまでのように，人生の最初の時期に学習し，残りの期間で仕事をするというライフスタイルから，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．また，人工知能の技術が進展しても，課題発見の能力やコミュニケーション能力といった能力は時代を超えて重要性が高いだろう．人工知能の時代だからこそ，改めて「人間力」，「社会力」[西田 14] が問われるようになるのではないだろうか．

5. 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もつ*23．創造性の定義にもよるが，すでにピカソ風の絵を描く [Gatys 15]，作曲をする，新聞記事を書くなどは実現されている．創造性が，多くの過去からの模倣と，（別領域からの知識の転移による）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．コンテンツの創作活動に詳しい株式会社ドワンゴの川上量生は，「近年の人工知能の進展を背景に，創造性は理屈で説明できないから難しいと思われてきたが，実は簡単だったというのがわかってきたのだと思う」と述べている*24．これと似たようなことは，過去にAlan Turingが「ラブレス夫人への反論」という論文のなかで，「コンピュータにも独創的なことはできないが，人間もまた独創的でない」と述べている [伊庭 16]．

人工知能が創造性をもった場合に，人工知能が創作した作品の権利はどうなるのか．現実的に，人工知能がつくった作品，人工知能を「道具として」人がつくった作品，あるいは純粋に人がつくった作品，この三つを外見上で見分けることは困難である．その場合，人工知能がつくった作品まで，人による創作と同じに取り扱われるならば，膨大な知的財産の独占が起こるかもしれない．こうしたことが，内閣府の知財戦略本部で昨年から議論されている [知財 16]．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．人間は，現在の自分を構成するものが何かを明示的に覚えていない．ある発明をしたとしても，その発明の根拠を

---

*19 NHKを含む多くのメディアで報道された．例えば，「人工知能学会，AI開発の倫理綱領案 安全確保強く求める」日本経済新聞（2016年6月6日掲載）

*20 原文ではwork activityだが，わかりやすくタスクと訳している．

*21 例えそうだとしても，仕事の内容の変化に対応できない人をいかに救済するか，教育の機会や支援をいかに提供し，新しいセーフティネットを構築するかは社会全体の課題である．

*22 山内祐平：人工知能に負けない子供，どう教育するか，日経BizGate（2015年11月16日）

*23 ここでいう創造性は，2章に述べた「自らを改変する知能」に必要な創造性とは大きくレベルが異なる．

*24 NHKクローズアップ現代＋「絵画・音楽・小説まで…人工知能は創造でも人間を超える？」（2016年7月12日放送）