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

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

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

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

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

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

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

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

長倉　克枝 Katsue Nagakura
科学ライター
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）

遡ることはできない．しかし，機械学習では何のデータが元になり何が学習されたかが追跡し得る．他人に権利のあるデータを使って学習して得られたモデルの権利は誰のものになるのだろうか．あるいは，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*25．欧州では検索エンジンの進展に対して，2012 年から「忘れられる権利」（現在では消去権）が認められたが，こうした新しい技術に対応した権利はもっとあるのかもしれない．例えば，防犯カメラが街中に設置されたときに，どんな些細な交通違反でも検出されてしまう．我々は少々の交通違反を「見逃される権利」があるのではないか，あるいは少しくらい「悪いこと」をする権利があるのではないか．あるいは，データが取得されたとしても，ある人には良いところだけを見せ，別の人には（嘘をつかないまでも）そのことを黙っている権利があるのではないか．こうしたことは，既存の法律の概念でいえば，プライバシー権，愚行権，自己決定権，黙秘権などの拡大概念などと解釈することもできる．技術の進展に伴い，こうした議論も必要になるのかもしれない．

個人情報保護法では，特定の個人を「識別できる」情報を含む情報をいう．しかし，我々は「識別器（分類器，classifier）」の能力によって，識別できるかどうかの結果が大きく変わることを知っている．意匠の類否判断はどうだろうか．意匠の類似性は，主に物品の類似と形態の類似から行われるが，我々は，類似度は「類似度関数」によって大きく異なることを知っている．すなわち「識別器」や「類似度関数」といったアルゴリズムの概念を入れないままに，人工知能が普及する時代に，こうした法律的な概念がこのまま扱えるのであろうか．あるいは，道路の法定最高速度は，規制速度算出要領により，道路沿いに住宅や店舗があるか，車線数，交差点の数などによりポイントで決められる．しかし，仮に自動運転が導入され，細かい気象条件や人通りの多さが時々刻々と計算されるなかで，より本質的な「安全性」に従った条件を設定することも可能ではないだろうか．

自動運転に関しては，すでにさまざまな議論が行われている．冒頭に書いたようにすでに米国では死亡事故も起きているが，現状のレベルの自動運転車においては，事故を起こした場合の責任は運転者にある．米国運輸省高速道路交通安全局は，2016 年 2 月に，Google に自動運転車に搭載される人工知能を「運転手とみなす」と伝えた*26．国内でも警察庁で自動運転における事故の責任や法律上・運用上の課題に関する議論が始まっている [警察庁 16]．自動運転のレベル 1 ～ 3 までは現行法とあまり変わらないと思われるが，レベル 4 の段階になると，きめ細やかな判断が必要になり，ある場合には事故の責任が製造者責任となり，また保険の仕組みも整備される必要があるだろう [損保 16]．倫理委員会ではさらに進んで，人工知能が法人格に似たある種の格をもつ「人工人」（自然人，法人と対比的に）という概念ができるのではないかと議論した．これと同じ議論が [カプラン 16] でも行われており，将来は，何らかの人工知能に対応する格ができ，（法人が利益を競うのと同じく）安全性を競うようになるのかもしれない．

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 [エドモンズ 16]．サンデル教授で有名な議論であり，他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」*27 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．前者は，どういったルールを採用するとどのくらい人の心に合うか・反するかというコストの設定の問題，後者は，社会の安定や発展という目的に従って諸制度をどのように設計するかという問題である．トロッコ問題でいうと，両方の状況が不可避なときに失う命の数の少ないほうを選ぶのだが，全く事件に関与しない第三者の命を脅かすことは，社会の安定性からして危険である．例えば，五人を守るためといって，いきなりある日殺されてはたまらないので，なくなる命の数が少ないほうが良いという規範より，善意の第三者は命を脅かされないという規範を上にもっていったほうが社会の安定性という観点からは良い．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ややもすれば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，法哲学などを巻き込んだ議論を引き起こしており，言い方を変えれば，人工知能の進展が人文社会学系のさまざまな学問分野に対して，これまで積み上げてきたそれぞれの学問分野での体系に挑戦状を叩きつけ，破壊と創造を迫っているともいえるかもしれない．

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つくりたいのかという問題に行き着くのではないだろうか．社会全体での目的関数が決まれば，人工知能の活躍する範囲は広い．しかし，社会全体の目的関数と簡単にいっても，非常に難解で複雑な議論が必要であることは論をまたないし，そのことは人類の歴史が雄弁に物語っている．一方で，人間が社会的な「動物」であることを考えると，人類という種を存続させることは，（ひとまず総論としては）人類全体の合意になり

---

*25「学習済みモデル」は知財か？―産業技術総合研究所，AI 市場形成で価値・権利認めるルールの検討開始，日刊工業新聞（2016 年 5 月 25 日）

*26「人工知能は運転手」米当局，Google 自動運転に見解，朝日新聞（2016 年 2 月 14 日）

*27 進化倫理学あるいは神経倫理学などの分野での議論に近い．

得るのではないかと思う．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[Song 13]．人工知能を，人類のサステイナブルな未来のために活用するというのは，人工知能と倫理という文脈の上では，一つの究極の目標ではないだろうか．

やや空想的な議論になることをあえて恐れずにいうと，人工知能技術を活用した新しい形の社会が誕生する可能性もあるかもしれない．例えば，ディープラーニングによる認識技術により，労働者の「努力」が正当に評価されるようになれば，努力に応じた富の配分も可能になる．それは社会主義国家の致命的な欠陥であったフリーライダーを防げないという問題を解決し，理想の社会主義的な国家を実現してしまうかもしれない．格差を減らし，あるいは資源の消費を抑え，人々の多様性を重視しながら，人々が自由に生き生きと活躍するようなサステイナブルな社会を実現することも可能になるかもしれない．これは少し楽観すぎる物語だろうか．

最後はかなり楽観論になってしまったが，人工知能の技術により我々の未来をどうしたいか．これは，研究者が自ら自覚と責任をもって正しい情報を社会に発信し，人工知能のも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．

6. お わ り に

SF作家のIsacc Asimovは，小説のなかでロボット三原則を考案した．ロボット三原則は必ずしも完璧なものではなく，小説の中でもさまざまな問題とともに描かれているが，コンセプトとしてはシンプルでわかりやすい．それに対抗するような，シンプルでわかりやすい「人工知能三原則」はないだろうか．ずっとそう思ってはいるが，本稿での議論を見てもわかるとおり，論点が多岐にわたり，とても簡単にまとめられそうもない．しかし，この議論が成熟し，より多くの人にわかりやすい形で情報発信ができればと思う．

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つく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．そして，人工知能に関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 ◇

[Amodei 16] Amodei, D., et al.: Concrete Problems in AI Safety, arXiv:1606.06565（2016）

[Barrat 13] Barrat, J.: *Our Final Invention, "Artificial Intelligence and the End of the Human Era.,* Thomas Dunne Books（2013）

[Bostrom 14] Bostrom, N.: *Superintelligence,* Oxford University Press（2014）

[知財 16] 知的財産戦略本部：次世代知財システム検討委員会報告書（2016年4月）

[Chui 16] Chui, M., Manyika, J. and Miremadi, M.: Where machines could replace humans — and where they can't (yet), *McKinsey Quarterly*（2016）

[エドモンズ 16] ディビッド・エドモンズ：太った男を殺しますか？—「トロリー問題」が教えてくれること，太田出版（2016）

[江間 16] 江間有沙, 長倉克枝：公開討論「人工知能学会倫理委員会」（特集「2016年度人工知能学会全国大会（第30回）」, 人工知能, Vol. 31, No. 6（2016，掲載予定）

[Frey 13] Frey, C. B. and Osborne, M. A.: *The Future of Employment: How Susceptible are Jobs to Computerization,* http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf（2013）

[Gatys 15] Gatys, L. A., Ecker, A. S. and Bethge, M.: A Neural Algorithm of Artistic Style, arXiv preprint arXiv:1508.06576（2015）

[伊庭 16] 伊庭斉志：人工知能と人工生命の基礎，オーム社（2016）

[井上 16] 井上智洋：人工知能と経済の未来 2030 年雇用大崩壊，文藝春秋（2016）

[カプラン 16] ジェリー・カプラン：人間さまお断り—人工知能時代の経済と労働の手引，三省堂（2016）

[警察庁 16] 警察庁交通局：自動運転をめぐる最近の動向と警察庁の取組について（2016年6月）

[Kurzweil 05] Kurzweil, R.: *The Singularity is Near*, Duckworth Overlook（2005）

[松尾 15a] 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，Vol. 30, No. 3, pp. 358-364（2015）

[松尾 15b] 松尾 豊 ほか：公開討論「人工知能学会倫理委員会」（〈特集〉2015 年度人工知能学会全国大会（第29回）），人工知能，Vol. 30, No. 6, pp. 754-755（2015）

[松尾 15c] 松尾 豊：人工知能 自動運転：ものづくりが変わる 日本の製造業に勝機あり，エコノミスト（2015年10月6日号）

[西田 15] 西田豊明：人間力・社会力を強化する情報通信技術 人工知能を中心に，情報管理，Vol. 57, No. 8, pp. 517-530（2014）

[野村 15] 野村総合研究所：日本の労働人口の 49%が人工知能やロボット等で代替可能に，プレスリリース（2015年12月2日）

[Orseau 16] Orseau, L. and Armstrong, S.:Safely Interruptible Agents, https://intelligence.org/files/Interruptibility.pdf（2016）

[Pistono 16] Pistono, F. and Yampolskiy, R. V.: Unethical research: How to create a malevolent artificial intelligence, *Proc. IJCAI-2016 Ethics for Artificial Intelligence Workshop*（2016）

[Russell 15] Russell, S., Dewey, D. and Tegmark, M.: Research priorities for robust and beneficial artifcial intelligence, *AI Magazine,* winter（2015）

[Song 13] Song, J.: Addressing sustainability via AI — Report from the 23rd International Joint Conf. on Artificial Intelligence, *Bulletin of the Chinese Academy of Sciences,* Vol. 27（2013）

[損保 16] 日本損害保険協会：自動運転の法的課題について（2016年6月）

[若田部 16] 若田部昌澄：経済学者は人工知能の夢を見るか：第2次機械時代の経済社会構想，総務省ICTインテリジェント化影響評価検討会議（2016年2月2日）

[Walsh 16] Walsh, T.: The singularity may never be near, *Proc. IJCAI-2016 Ethics for Artificial Intelligence Workshop*（2016）

2016年8月22日 受理

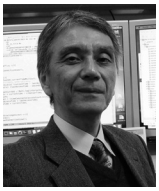
著 者 紹 介

松尾　豊（正会員）

1997 年東京大学工学部電子情報工学科卒業．2002 年同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 10 月より，東京大学大学院工学系研究科准教授．2012 ～ 13 年度本学会編集委員長，2014 年より本学会倫理委員長．専門は，Web マイニング，Deep Learning，人工知能．

西田　豊明（正会員）

1977 年京都大学工学部卒業．1979 同大学院修士課程修了．1993 年奈良先端科学技術大学院大学教授，1999 年東京大学大学院工学系研究科教授，2001 年東京大学大学院情報理工学系研究科教授を経て，2004 年 4 月京都大学大学院情報学研究科教授，現在に至る．会話情報学，原初知識モデル，社会知のデザインの研究に従事．本学会会長（2010 ～ 11 年度），日本学術会議連携会員（2006 年～）．情報処理学会フェロー．電子情報通信学会フェロー．JST CREST「共生社会に向けた人間調和型情報技術の構築」研究領域総括（2014 年～）．AI & Society 誌 Associate Editor.



堀　　浩一（正会員）

東京大学大学院工学系研究科教授．2008 ～ 09 年度本学会会長．2010 年～本学会顧問．知識処理システム構築のための知識獲得の困難に直面したのち，研究すべきは，知識の獲得ではなく知識をつくり出すことの支援であると考えるようになり，創造活動支援システムの研究を行ってきた．人間がもっと創造的に仕事するための道具としての人工知能の研究を行ってきたが，単なる道具では済みそうになくなってきたので，最近は人工知能倫理の問題にも興味をもっている．

武田　英明（正会員）

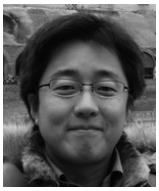
1986 年 3 月東京大学工学部卒業．1988 年 3 月同大学院工学系研究科修士課程修了．1991 年 3 月同博士課程修了．工学博士．ノルウェー工科大学，奈良先端科学技術大学院大学を経て，2003 年 5 月より国立情報学研究所教授．2006 年 4 月～ 2010 年 3 月同学術コンテンツサービス研究開発センター長，2005 年 12 月～ 2010 年 3 月東京大学客員・特任教授．知識共有，セマンティック Web などの研究に従事．AAAI，Design Society, 情報処理学会各会員．

長谷　敏司

1996 年 3 月関西大学文学部卒業．2001 年角川スニーカー大賞にて作家デビュー．2009 年より SF を本格的に書き始め，2009 年『あなたのための物語』, 2012 年『BEATLESS』で日本 SF 大賞ノミネート，2014 年『My Humanity』で日本 SF 大賞受賞．作品は近未来を舞台に AI が登場するものが多く，2014 年ニコニコ学会参加をきっかけに，AI が現実社会に実装されるとき，社会の変化を望まない人々にまで触れてしまうインパクトについて考えるようになる．日本 SF 作家クラブ会員．

塩野　　誠

1998 年慶應義塾大学法学部政治学科卒業．2008 年ワシントン大学ロースクール法学修士．株式会社経営共創基盤取締役．国内・海外にて企業，政府機関に対して戦略立案・実行のアドバイスを行っている．テクノロジー関連の事業開発，企業提携や M&A，企業危機管理の実績を多く有する．



服部　宏充（正会員）

2004 年 3 月名古屋工業大学大学院博士課程修了．博士（工学）．2004 年 4 月～ 2007 年 3 月日本学術振興会特別研究員（PD）．リバプール大学，およびマサチューセッツ工科大学客員研究員を経て，2007 年 4 月～ 2014 年 3 月京都大学大学院情報学研究科助教．現在，立命館大学情報理工学部准教授．マルチエージェントシステム，社会シミュレーション，デザインに関する研究に従事．本学会の「表紙問題」における社会の反応から, 人工知能と社会の関係に関心をもち, 有志とともに AIR（Acceptable Intelligence with Responsibility）の活動を開始.

江間　有沙（正会員）

2007 年東京大学教養学部卒業．2012 年同大学院総合文化研究科博士課程修了．博士（学術）．2012 年京都大学白眉センター特定助教，2015 年 4 月より東京大学教養学部附属教養教育高度化機構特任講師．2014 年より人工知能と社会の関係について考える AIR（Acceptable Intelligence with Responsibility）研究会を有志とともに開始．専門は科学技術社会論.

長倉　克枝

2007 年北海道大学獣医学部卒業．同年，日本経済新聞社入社（科学技術部，証券部記者），2011 年退社．文部科学省系独立行政法人勤務などを経て 2015 年からフリーのライター・編集者．テクノロジーと社会，医療などを中心に取材・執筆．朝日新聞出版 AERA 編集部所属（記者）.