総務省 学術雑誌『情報通信政策研究』 第2巻第1号
Journal of Information and Communications Policy Vol.2 No.1

特別寄稿

人工知能として認識されない人工知能の埋め込まれる社会に向けて

堀　浩一[1]（東京大学／理化学研究所）

要　旨

人工知能への期待と不安の両方が高まる中で、人工知能の研究開発のガイドラインなどが議論され提案されてきている。しかし、そのようなガイドラインを適切に定めてそれを守って研究開発がなされたとしても、人間社会においていろいろと問題が発生する可能性は残る。なぜなら、人工知能の技術が社会のあらゆるところに埋め込まれ、それらが相互にネットワークで結合されて機能するようになると、それぞれの人工知能技術は健全であってもそれらが相互作用したときに予期せぬ事態が発生する可能性はあると考えられるからである。想定外の事態に対応するためには、人間社会の方でも、従来とは異なる、「責任」や「権利」などのあり方の可能性を検討しておく必要があろう。さらには、その検討を受けて、人工知能のあり方についても再検討が必要になろう。本稿では、そのような議論のサイクルの端緒を示すことを試みる。

キーワード： AIネットワーク、人工知能倫理、社会システム、人工知能設計、液状化社会

1．まえがき

人工知能への期待が高まる一方で、人工知能への不安も高まっている。不安を少しでもやわらげるべく、人工知能の研究開発に関して、ガイドラインの提案などが国内外で行われている。しかし、残念ながら、今のところそれらのガイドラインを実際の人工知能の研究開発に生かすための具体的な道筋は明確には見えていない。その最大の理由としては、人工知能のもたらしうる倫理的な問題に深い関心を抱いている人工知能研究者の数が十分でないことを挙げることができるだろう。その一方で、もう一つの重要な問題も、合わせて考える必要がある。それは、人工知能の研究開発のガイドラインを適切に定めてそれらを守れば済むかというとそうではなく、人間社会の方も変わらざるをえない、その変え方の可能性に関する議論を十分に行わなければ片手落ちになってしまう、という問題である。技術哲学における社会構成主義が教えるところによれば、技術が社会を変えるのでもなく、社会が技術を変えるのでもなく、技術と社会は双方向に相互作用する[1]。本稿では、「社会がどう変わりうるか」の可能性の検討を行い、そこから人工知能に求められることをもう一度考えるという作業を行ってみたい。

以下、まず、そもそも人工知能とは何かについて議論を行い、そのあとで、社会の変化の可能性について検討してみる。そこでは、人間社会を構成するいろいろな重要概念が「液状化」するのではないか、ということを論じる。「液状化」の意味については、本文中で述べる。最後に、その議論を受けて、もう一度、人工知能について考える。

---

[1]東京大学大学院工学系研究科教授、理化学研究所革新知能統合研究センターチームリーダー

２．そもそも人工知能とは

「人工知能とは何か」という質問は、実は、人工知能研究者にとって答えるのが難しい質問である。「人工知能とは」という本[2]では、筆者を含めて 12 人の研究者がそれぞれ異なる答えを書いている。筆者は、この本で、次のような人工知能の定義を書いた。

- artificial intelligence = new worlds of intelligence, which are synthesized artificially,
- a world of intelligence = the whole of the elements and the whole of the mutual relationships among the elements, where
- the elements = human brains, human bodies, tools, problems, solutions, data, information, knowledge, wisdom, values, emotions, languages, machine programs, machine bodies, machine networks, human networks.

やや大き過ぎる定義ではあるが、今後さまざまな新しい人工知能の研究が進んだ時にも、適用可能な定義であると考える。

図 1 (a) 人間の一人を機械に置き換えた人工知能

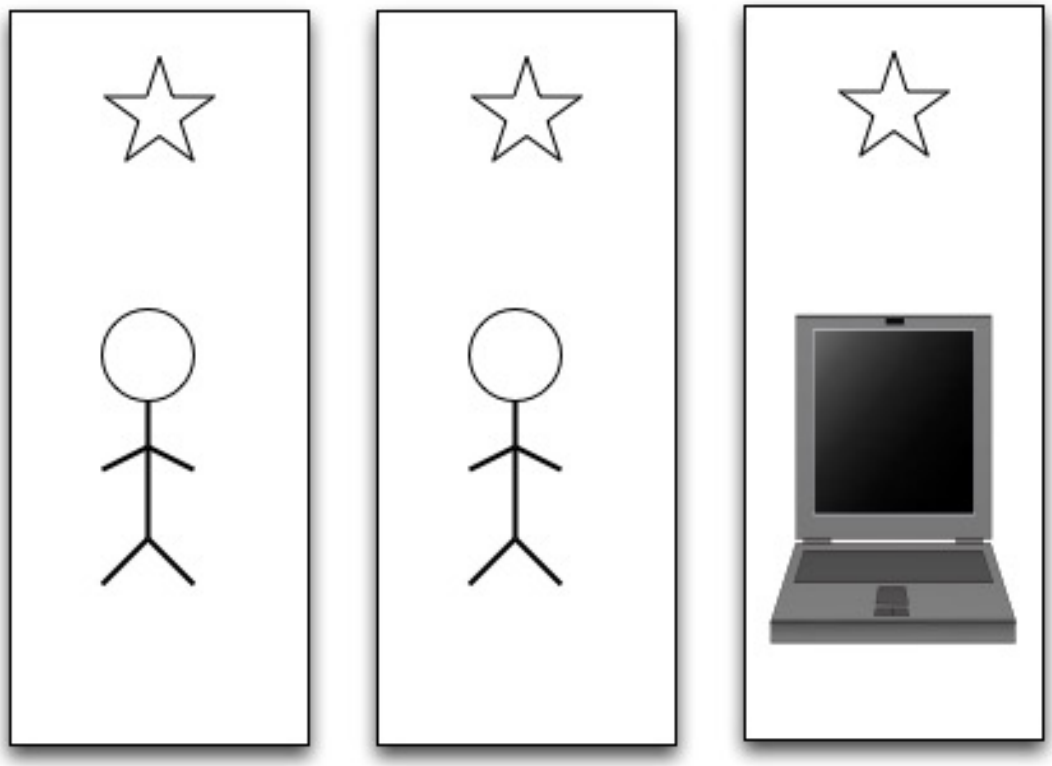
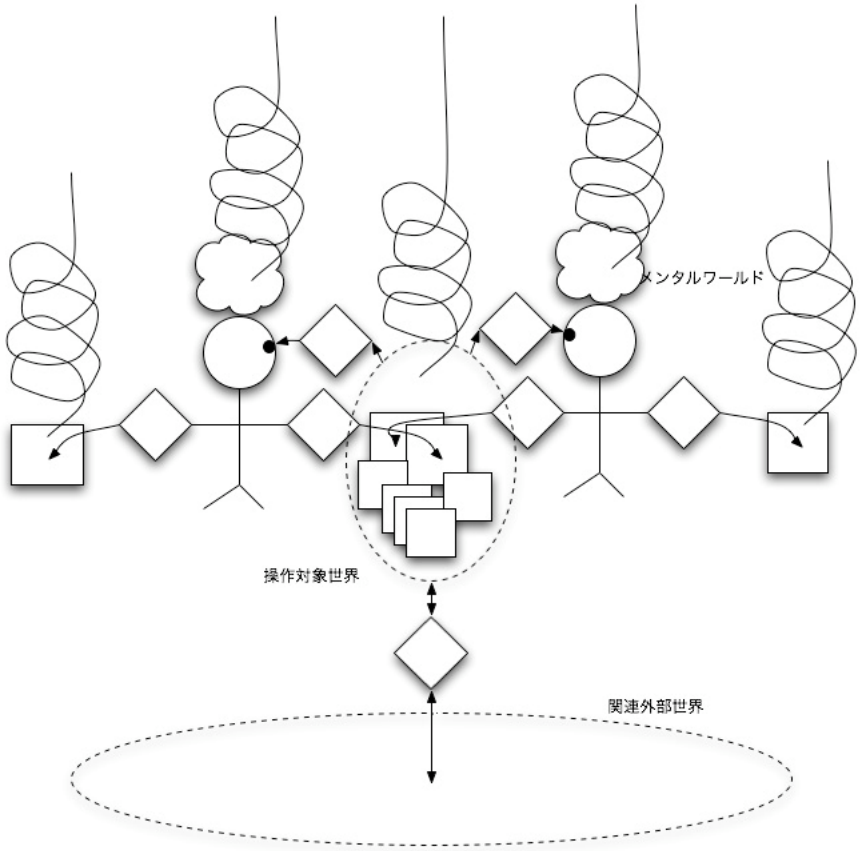


図１(b) 人間と人間、人間と外界、および外界中の存在どうしの間にさまざまな人工的プロセスが埋め込まれることにより構成される新しい知能世界



これを絵に描いて説明すると、図１のようになる。マスコミの報道などにおいては、人工知能をなんらかの独立した物として扱うことがある。それを図１(a)に示している。そこでは、一人の人間が一つの人工知能に置き換わっている。そのような人工知能のあり方もないわけではないが、実際には、図１(a)のような姿ではなく、図１(b)のような姿の人工知能のあり方のほうが多く、今後ますますそうなっていくであろうと予想される。図１(b)においては、なんらかの物としての人工知能は存在していない。存在しているのは、図において菱形で示されているいろいろなプロセスである。例えば、図の中で、人間の目と外界との間に菱形が置かれているが、これは、人間が外界を見るときに、なんらかのプロセスが働いて、外界の見え方を変えることになることを示している。それは、VR 技術のようなものによって物理的に見えるものを変えうることから、データマイニングによってデータの中からなんらかの知見が見えるように概念レベルで見え方を変えうることまでを、含む。同様に、人と人の間で情報や物を受け渡すプロセスに変化を起こすことができるし、人の手と外界の間に起こるプロセスに変化を起こすことができるし、外界の物や情報どうしの間に成立するプロセスに変化を起こすこともできる。そのプロセスを変化させるための技とし

ては、これまでに人工知能の分野で研究されてきた、自然言語処理や、画像処理や、機械学習や、オントロジーや、マルチエージェントシステムや、インタラクションデザインや、知的コミュニケーションシステムや、知的故障診断や、知的設計支援や、知的制御や、記号論理や、創造活動支援、等々、さまざまな研究成果が応用されることになる。さらに、今後は、IoT(Internet of Things)やVR(Virtual Reality)の研究成果なども組み合わされて用いられることになる。

図1(b)に示したような人工知能の姿は、一般には、まだあまり知られていないかもしれない。大量のデータ（いわゆるビッグデータ）から知識を獲得する機械学習が近年注目を集めているが、それだけが人工知能ではない。また、囲碁や将棋をプレイするシステムなど、人間と同じ仕事をして人間に置き換わる人間代替型のシステムも人々に驚きを与えた。しかし、それらも人工知能のほんの一部にすぎない。今後、人工知能の技術は、人間社会のあらゆるところに埋め込まれて行く。しかも、それらは、目に見えない形で埋め込まれ、相互に接続されて、機能するようになる。

従来、図1(a)のような人間に置き換わる形の人工知能に脅威を感じる人は多かったが、図1(b)のような形態は、artificial intelligenceというよりも、intelligence amplifiersと呼ばれることが多く、人を助ける道具の集まりのイメージが強かったので、脅威を感じる人は比較的少なかったように思われる。しかし、実は、この図1(b)の形態も、難しい問題を引き起こす可能性がある。次章ではその問題を検討する。

３．人工知能が埋め込まれた社会における人と機械の境界の不明瞭化

前章で述べたように、人工知能の技術が社会のあらゆる場所に埋め込まれると、どういうことが起こりうるだろうか。

現在においても、テレビの天気予報において、気象予報士が「コンピュータが予想した雨雲の動きです」という言い方をすることがあるが、これについて読者はどう思われるだろうか。筆者は、「コンピュータが予想した」ではなく、「コンピュータを使って予想した」と言ってほしいと思ってしまう。電車の駅のホームのアナウンスについても同様で、「ドアが閉まってまーす」と言われると、「ドアを閉めています」と言ってほしいと思ってしまう。しかし、筆者の周辺の人々に尋ねた限りでは、「コンピュータが予想した」という言い方に違和感を覚えない人は多いらしい。雨雲の動きを予想したのは、コンピュータだろうか、気象庁の予報官だろうか、種々のデータを集めたセンサたちだろうか、予想プログラムを作成した人だろうか。人工知能が入り込んだ近未来社会ではなくとも、すでに現在において、天気予報のような身近な機能において、人間と機械とデータの役割は、複雑に入り組み始めていると言ってかまわないだろう。「気象衛星からの情報やアメダスなどの観測点からの情報などを入力として、数十年前から気象庁において改良を続けている予報プログラムを用いて計算した結果、次のような雨雲の動きの予想が出力されました」というようなややこしいことを言わなくても、すべてをコンピュータのせいにして、「コンピュータが予想した雨雲の動きです」と言っても、不思議でない世の中になっているらしい。

しかし、天気予報システムの場合は、まだ、人とプログラムとデータの役割の境界は、見定めようと思えば、見定めることができるであろう。おかしな予想が出力された時に、予報官のミスなのか、プログラムのバグなのか、データの誤りなのか、調査することはできそう

である。人工知能が埋め込まれた社会においては、そのような調査が難しくなる場合も増えることが予想される。それは、社会のいろいろな重要概念の変化を迫ることになるかもしれない。それを次章で考える。

４．人文社会科学における諸概念の液状化の可能性

一つ仮想的な例題を考えてみよう。実際にそうなってもおかしくない近未来の日常生活を想像してみたい。ある日新しい冷蔵庫を買ったら、「冷蔵庫の中の品物の情報をあなたの健康管理のために役立たせるサービスを提供します。そのサービスを受けることに同意しますか？」と言われて、同意したとしよう。すると、冷蔵庫の中の品物の情報が、随時、健康管理サービス会社に送られて、健康管理サービス会社のシステムは、いろいろなプログラムを使って、その冷蔵庫のユーザの食生活の改善方策を推論するとしよう。その推論結果をユーザに直接教えればいいのだが、健康管理に問題のあるようなユーザは直接アドバイスをしても従わないかもしれないと親切に考え、健康管理サービス会社は、直接ユーザにアドバイスを提供する代わりに、食品販売ネットワーク会社と組んで、そのユーザがより健康になるような食品を知らず知らずのうちに買いたくなるような仕掛けを提供するというような複雑なこともやるようになると想定してみよう。例えば、もっとほうれん草を食べさせた方がいいとわかったら、どういうわけかそのユーザがスーパーマーケットに行くと、ほうれん草を食べたくなるようなポップ広告がそのユーザにだけ見えるのである。ユーザは、なんとなくそのポップ広告につられて、ほうれん草を買って食べ、健康改善に役立つらしい。

さて、これは、人工的に作った例題で、いろいろと設定に無理はあるが、この健康管理サービスを使ったせいで、ユーザがかえって病気になってしまった、という事態が出現したとしたら、どういうことになるか、ということを考えてみたい。この例題で登場するステークホルダーは、すべて善意の人々であり、善意に基づいて作られた人工システムたちである。それでも、そのせいで病気になったとしたら、その責任をどのように考えることができるであろうか。健康管理会社は、何を食べろと、直接は指示していない。どういうわけかそれを食べたくなるような、ナッジ(nudge)を仕組んだだけである。ユーザは、自由意志に基づいて、食べたいものは自分で決められたはずである。しかし、そのユーザに特化したナッジが巧妙に仕組まれていたことに問題はないだろうか。健康管理会社の使ったシステムはどうだろうか。一体どういうデータとどういう知識を使うシステムによって、健康増進のアドバイスを出力していたのだろう。そもそも、健康管理サービスと健康あるいは病気との因果関係を証明するのは難しいのではないか。しかし、そのサービスを使った人たちが、使わない人たちに比べて、統計的に有意に多く、健康が阻害された、とわかったとしたら、どういうことになるだろうか。

もう少しうまい例を作れればよかったのだが、指摘したいことは、人工知能の技術が社会のあらゆるところに埋め込まれ、それらが相互に接続されて働くようになったときには、自由意志、責任などという重要概念が揺らぐかもしれないという問題である。上の例は、単一サービスだけを想定したのでつまらない例になってしまったが、今後、さまざまな機能を目に見えない形で結合し、さらに、いろいろなナッジを組み合わせて仕掛けることにより、本人が意識しないような形で本人の行動を変化させることが可能となったとき、誰がどのように責任を取ることができるであろうか。刑事責任と民事責任とでは異なるであろうし、人

命に関わるような問題とそうでもない問題とでは異なるであろう。筆者は人文社会科学の分野は素人であるので、何も確たることは言えないが、人文社会科学の専門家の方々とともに、今後、下に述べるような問題をできるだけ急いで検討していく必要があるのではないかと考えている。

それは、人文社会科学が扱ってきた諸概念が、人工知能技術が社会に入り込むことによって、「液状化」するのではないかという問題である。あえて「液状化」という用語を使うのは、筆者が人工知能の研究領域において「知識の液状化と結晶化」という問題を扱ってきた[3]からでもあるのだが、社会学者の Bauman が早くから指摘していた「社会の液状化」[4]の問題が人工知能技術の浸透にともなってますます顕在化すると考えるからである。液状化の意味は、ここでは、「概念の境界が固定されず変化し、それが適用される範囲が流動的な広がりを持ち、他の液状化された概念と混じり合う可能性があること」としておきたい。

重要な概念が液状化する可能性があると言うと、やや怪しげであるが、一般の方々にわかりやすく説明するために筆者が時々使う例は、「オリンピックとパラリンピックの位置付けが逆転し、パラリンピックが主でオリンピックが従であるような社会」を想像して欲しい、という例である。もともと完全な人間などというものは存在しなくて、いろいろな不足点を道具で補うのは自然な姿である。義足を履いた人にとっては、その人の身体の境界線は人体と義足の間には存在しなくなり、義足と一体化して身体として認識されるようになる、とのことである。Polanyi の「盲人の杖」の例[5]を思い出す人もいるだろう。同様に、さまざまな人工知能技術を活用するようになったときには、人の知能と機械の知能の境界は明確でなくなり、人間と機械が一体となって知的活動を行うようになる可能性がある。それは、未来の SF 的な話ではなく、現在においてもすでに、スマートフォンがないと目的の場所にたどり着くことができない若者たちに見られる姿であると言ってもよいであろう。

液状化する可能性について検討すべき概念として次のようなものを挙げてみたい。すでに人文社会科学の研究者の間でさまざまな議論が行われている[6][7][8][9]ので、人工知能研究者もこれらを急いで勉強する必要がある。

- 「道具」という概念： 受動的な道具（金槌、ピアノなど） から 能動的に行為を行う道具（自動作曲、自動作詞など）までの間の連続的な広がり。
- 「道徳的被行為者としての他者」という概念： 人形・ぬいぐるみから、ペットの動物、ロボット、人間までの間の連続的な広がり。
- 「道徳的行為者としての自己」という概念： 人と機械の組み合わさった自己の可能性。
- 「自律的な個人の自由意志」という概念： カント的な意味で自律した（自由意志から道徳的義務を理由に行動できる）一人の人間として統合された個人から、分人、ナッジなどによって無意識の相互作用から生まれる意志、物理的因果に基づく決定論、さらにはカオスや random phenomena のような非決定論的現象までの連続的な広がり。これは「責任」という概念の議論に直結するだろう。
- 「責任」という概念： 「自律的な個人の自由意志」に基づく道徳的行為者という概念の液状化にともない、液状化され分配される責任を考える必要が出てくるかもしれない。
- 「説明可能性」という概念： 自由意志に基づく行動の理由の説明から、本能的行動の

後付けの説明、さらには説明不能の行為までの広がり。[*注１]

- 「権利」という概念： 権利をもつ者の境界線が明確でなくなる可能性。

ここに挙げた概念の変化を検討し、それに対応した社会制度を準備するのは大変なことになるであろうが、人工知能の研究開発者としては、社会におけるその準備がなされていないと、安心して人工知能の研究開発を行うことができなくなる恐れがあると考える。「コンピュータの予想した雨雲の動き」の情報のせいで大きな損失を被った人が、その予想プログラムの作成者を訴える、ということに相当するような、従来とは異なる、いろいろと複雑な訴訟が起こるようになったとしてもおかしくはない。例えて言うならば、雨雲の動きの予想が外れた時の責任を、様々な人間と様々な機械と様々なデータとで分担して負う、すなわち責任を液状化して負わせる、というような制度が欲しい。

５．液状化する社会における人工知能開発の留意点

「透明性」、「説明可能性」、「制御可能性」などの人工知能への要請[10]については、すでに多くの議論がなされ、ガイドラインなどが提案されている。

本稿で問題にしたのは、それらのガイドラインを守っても、予期しない問題が発生しうる、という問題である。

前章では、その問題に備えるために、人間社会の側で、いくつかの重要概念の液状化に対応した社会制度を準備して欲しいという要望を述べた。そのような社会制度の整備を可能にするためには、今度は、再び、人工知能の作り方にも要請がなされることになるであろう。本章では、それを考えてみたい。

前章の最後に例題として述べた、「いろいろな責任を、様々な人間と様々な機械と様々なデータとで分担して負う」というような社会制度を設計することは可能であろうか。あるいは、特に何もしなくても、現行制度のままで、なんとかなるのであろうか。すでに述べたように、それは、刑事と民事では異なるであろうし、対象領域ごとでも異なるであろう。人文社会科学における今後の議論の進展に期待したいが、その議論との間で往復運動をする形で、人工知能研究の側でも、人工知能の設計の見直しを検討する必要が出てくると考える。

例えば、前章で述べたように、「道徳的行為者としての自己」という概念が液状化して、「人と機械の組み合わさった自己」のような、従来の「自己」よりも広がりを持った「自己」という概念を認めて、社会制度が作られるようになった場合のことを考えてみよう。これまた、先に述べた例を使って、義足を履いた走り幅跳びの選手が義足を履かない走り幅跳びの選手よりも良い記録を作った、としてみよう。その時に、その好記録が、本人の努力の成果なのか、義足の性能のおかげなのか、と議論することに意味があるかどうかは、そもそもよくわからなくなるだろう。本人にとっても、それはよくわからないことかもしれない。が、競技の公平性などを持ち出す時には、どうしても、本人の果たした役割と義足の果たした役割との分担の見定めを行いたくなるであろう。最近の人工知能をめぐる議論の中で、人工知能の説明可能性を問題にし、人工知能が自らの行いを自分で説明できるようにすべきであるという主張がなされることがあり、実際、そのような研究も盛んになり始めている。しかし、これは、義足に向かって、「義足よ、お前の果たした役割を自分で説明せよ」

と言っているようなものかもしれない。この義足の例でも明らかなように、「人と機械の組み合わさった自己」における人と機械の役割分担を同定したければ、どうしても、その「自己」の外側から観察を行って、人と機械の果たした役割を分析するという作業を行う必要が出てくるであろう。その観察を精密に行いたいならば、選手の身体と義足にさまざまなセンサを取り付けて計測を行う必要があるだろう。この例を、人工知能が埋め込まれる社会の問題へと拡張することは、一つの可能性としては、考えてみてもよいのではないだろうか。すなわち、社会における重要概念が液状化した時に、液状化した「自己」や液状化した「責任」や液状化した「権利」の中身を同定したくなるならば、なんらかの観察系を別途構築する必要がある、という考え方である。

さまざまな人工知能技術の埋め込まれた社会において観察を行い、悪い現象の予兆を発見して警告を出したり、「自己」や「責任」の中身を同定したりするための系というのを考えるとしたら、その系も人工知能技術を活用して構築せざるを得ないし、また、その系も分散的に社会に埋め込まざるを得ないであろう。

そうなると、その観察系がやることを観察するもう一つ別の系も必要となって、問題が循環してしまわないだろうか。確かに、なんらかの系があって、その系を観察する系がその上にあって、と考えると、どんどん上の系を作る必要が出てきてしまうかもしれない。また、下手に作ると権力による国民監視システムにつながってしまうかもしれない。そういうことを避けるとしたら、さまざまな系が同じレベルで行為や観察の相互作用をしていて、全体としては、人間社会にとってプラスとなるような平衡状態に常に引き込まれて落ち着く、というような系の設計を行う必要が出てくるであろう。

誰かがその全体設計を行うというのは、実質的には困難であるが、現実的には、人間社会でいろいろなコミュニティが有効に機能しているのと同様に、いろいろな種類の数多くの草の根 AI ネットワークを構築し、それらを社会に投入する、というのが一つの有望な解であろうと筆者は考えている。

６．むすび

本稿では、まず、人間代替型ではなく、さまざまな人工知能技術があらゆるところに埋め込まれ相互に接続されているという人工知能活用社会の可能性を述べた。そのような社会においては、いろいろな意味で、人と機械との境界が明確でなくなる可能性を指摘した。人と機械の境界があいまいになった社会においては、人文社会科学のいくつかの重要概念も変化させる必要があるのではないか、と考えた。さらに、それらの概念の変化に対応して、人工知能技術として何が求められるかを考えた。

本稿での議論は、それらの可能性の端緒を示すにとどまり、具体的な社会制度の設計の例や具体的な人工知能システムの設計の例を示すことまではできなかった。人工知能を巡る議論は盛んになる一方であるが、実際に社会の何をどうしたいのかの議論はまだまだ不十分であり、実践を伴った提案が少しずつ出てくることが期待される。いろいろな立場の人々が、いろいろと異なる解の可能性を提案し、それらをうまく組み合わせることが望まれるであろう。

[*注１] ここではAIの動作の説明可能性(explainability)ではなく、AIを活用した（あるいは AI に誘導された）人間の行為の説明およびその責任を問題にしている。これは、accountability の問題であると言い換えることもできる。本稿の脱稿後に、G7 Multistakeholder Conference on AIという会議が開催され、その中の「Accountability in AI – Promoting Greater Societal Trust」というセッションでは、accountabilityとtrustの問題について、突っ込んだ議論がなされた。この会議のためにカナダと日本で協力して作成したdiscussion paper [11]に、accountabilityとtrustについて緻密な議論が展開されているので、その問題に興味のある方はぜひ参照されたい。

参考文献

[1] 村田純一： 技術の哲学、岩波書店、２００９.
[2] 中島秀之・西田 豊明・溝口 理一郎・長尾 真・堀 浩一・浅田 稔・松原 仁・武田 英明・池上 高志・山口 高平・山川 宏・栗原 聡: 人工知能とは、近代科学社, 2016.
[3] 堀　浩一：　創造活動支援の理論と応用, オーム社, 2007.
[4] Zygmunt Bauman: Liquid Modernity, Polity, 2000, 2012. 森田典正訳：リキッド・モダニティ　液状化する社会、大月書店、2001.
[5] Michael Polanyi, The Tacit Dimension, Routledge and Kegan Paul Ltd., 1966. 佐藤敬三訳: 暗黙知の次元—言語から非言語へ , 紀伊国屋書店, 1980.
[6] 弥永 真生, 宍戸 常寿　編: ロボット・AIと法, 有斐閣, 2018.
[7] 平野 晋: ロボット法--AIとヒトの共生にむけて, 弘文堂, 2017.
[8] 久木田水生, 神崎宣次, 佐々木拓： ロボットからの倫理学入門, 名古屋大学出版会, 2017.
[9] 山本龍彦　編著：　ＡＩと憲法, 日本経済新聞社, 2018.
[10] 堀　浩一, 人工知能の研究開発をどう進めるか - 技術的特異点（シンギュラリティ）を見据えて, 情報管理, Vol. 58, No. 4, pp. 250-258, 2015.
[11] Jason Millar, Brent Barron, Koichi Hori, Rebecca Finlay, Kentaro Kotsuki, Ian Kerr: Accountability in AI - Promoting Greater Societal Trust,
G7 Multistakeholder Conference on Artificial Intelligence,
Montreal, Canada, December 6th, 2018.
https://www.ic.gc.ca/eic/site/133.nsf/vwapj/3_Discussion_Paper_-_Accountability_in_AI_EN.pdf/$FILE/3_Discussion_Paper_-_Accountability_in_AI_EN.pdf