論　文

人工知能の倫理とその教育

久木田水生[†]

Ethics of Artificial Intelligence and How to Teach It

Minao KUKITA[†]

あらまし　現在、人工知能は様々な倫理的な問題を引き起こしている。人工知能を健全に発展させるためには、人工知能の倫理的問題とその解決についての教育が重要である。本稿では、人工知能の倫理的問題をどのように教えるべきかを論じる。

キーワード　人工知能、データサイエンス、情報倫理

1. まえがき

現在、人工知能（特にビッグデータに基づく機械学習システム）の発展が著しく、経済や産業を牽引するものとして、人々の生活を便利にする道具として、大きな期待がかけられている。そしてそれを反映してデータサイエンスや人工知能の研究と教育が重視されるようになっている。しかし同時に人工知能は予期されていなかった様々な倫理的な問題を引き起こしてもいる [13]。人工知能を健全に発展させ、社会に役立てるためには、テクニカルな知識とスキルだけではなく、人工知能のはらむ倫理的問題とその解決についての教育もまた重要である。本稿では、人工知能の倫理的問題とそれをいかにして教えるべきかを論じる。

2. 人工知能をめぐる喧噪

2012 年、物体認識コンテスト ILSVRC において Google の開発した、「深層学習」という手法を用いた画像認識システムが従来を大きく上回る精度を記録して優勝した。このシステムは「ディープ・ニューラル・ネットワーク」というアーキテクチャーをベースに、従来とは異なる「教師なし学習」[1]によって、与えられた膨大なデータから自動的に対象を識別するための特徴を抽出する。そのためこの方法によって、コンピューターが自ら猫の顔や人間の身体などの「概念を学習する」とさえ言われた [12]。深層学習の成功を起爆剤に、2010 年半ばごろから人工知能の性能が急激に向上し、その他の様々なタスク、分野に応用されるようになった。そしてこのころからメディアには「AI」や「人工知能」という言葉が溢れかえるようになった。

世間を大きく騒がせた出来事として AlphaGo が思い出される。2016 年に Google 傘下の DeepMind が開発した囲碁ソフト、AlphaGo が、当時、世界最強と言われた棋士、イ・セドルを破った。AlphaGo は 5 番勝負で 4 勝 1 敗と圧倒的な強さを見せた[2]。しかもその手法が驚異的だった。AlphaGo は人間から何が良い指し方なのかを教わったわけではない。過去の対局の記録と、そして自己対局から学習し、いわば独学で強くなった。さらにそのおよそ 1 年半後、2017 年に発表された AlphaGo Zero は、囲碁のルールだけを教えられた後は自己対局のみから学習を行ない、40 日後には過去のバージョンよりも強くなっていた。さらにさらにその二か月に発表された AlphaZero は同じ手法を一般化して、チェス、将棋、囲碁において半日にも満たない期間の学習で既存の最強の AI（囲碁では AlphaGo Zero）に勝利するまでになった。

もっとも高度な知的ゲームと考えられた囲碁において人間がはるかに及ばないパフォーマンスを発揮した

[†] 名古屋大学
Nagoya University
DOI:10.14923/transj.???????????
(注1)：ラベル付けされていないデータセットに基づいた学習。

(注2)：なおイ・セドルが勝利した第 4 局目は NFT（非代替性トークン）としてオークションにかけられ、60ETH（その時のレートで 210000 ドル相当）で落札された。https://cryptheory.org/nft-of-go-players-victory-over-google-ai-sells-for-210000/

ことは大きな驚きをもって迎えられた。このころイーロン・マスク、スティーヴン・ホーキング、ビル・ゲイツといった著名人たちが、「AI は人類の生存にとって脅威になりうる」という AI 脅威論を広めた。哲学者でトランスヒューマニストのニック・ボストロムが描く「超知能」の不穏なビジョンも大きな影響力を持っていた [3]。

人工知能によって多くの仕事が奪われる可能性も指摘され、それによって生み出される失業に対処するための政策の必要性が主張された [4], [8], [9]。また「AI に仕事を奪われないために人間はどんな能力を身に着け、伸ばすべきか」といったことも論じられた [1]。その裏返しで、人工知能が新たな産業革命を牽引するだろうという大きな期待も声高に叫ばれ、多くの政府や企業が人工知能の開発と活用に食指を動かした。日本政府が掲げる「Society 5.0」なる夢のような未来社会のビジョンにおいても、人工知能は鍵となる役割を担うものとされている。さらに 2020 年、日本政府は「ムーンショット型研究開発」という大型のプログラムを立ち上げたが、その達成目標の中には「ゆりかごから墓場まで、人の感性、倫理観を共有し、人と一緒に成長するパートナー AI ロボットを開発し、豊かな暮らしを実現する」というようなものもある[3]。

一昔前ならば荒唐無稽なおとぎ話と言われたかもしれない。しかし AI の成功はこのような言説が説得力を持って聞こえてくるほどに顕著だった。

3. 人工知能の倫理

人工知能のすごさが喧伝される中で、人工知能の負の影響に対する懸念、それに対処する必要性も主張された。2017 年にはアメリカの NGO、Future of Life Institute が「Asilomar AI Principles」を発表した[4]。これは Research Issues、Ethics and Values、Long-term Issues の三分野に分かれて、全部で 23 の原則がある。Ethics and Values の分野では安全性、透明性、責任、価値との調和、プライバシー、自由、利益の共有、人間による制御、社会的市民的プロセスの尊重、AI 軍拡競争などに関する原則が挙げられている。

また同じ年に日本の人工知能学会も倫理指針を発表している[5]。この倫理指針は全部で 9 条からなり、「人類への貢献」、「法規制の遵守」、「他者のプライバシーの尊重」、「公正性」、「安全性」、「誠実な振る舞い」、「社会に対する責任」、「社会との対話と自己研鑽」、「人工知能への倫理遵守の要請」について定めている。

現在まで様々な学会、NGO、企業、政府機関、国際機関が人工知能に関する倫理的原理や規制を定めている[6]。こういった原則の中では、人類への貢献、公平性、透明性、包摂性、安全性、説明責任（アカウンタビリティ）、プライバシー、ウェルビーイングなどの重要性が説かれている。

こういった原則で書かれていることは抽象的・一般的、かつ総花的なものが多い。そのために人工知能に限定せずテクノロジー一般についても言えることがほとんどであるように思われる。これは人工知能という言葉がそもそも非常に多様なテクノロジーを指す、曖昧な用語であるということ、また深層学習のような基礎技術はありとあらゆる用途に応用ができるということに起因しているのだろう。かつ「AI が人間を超える」とか「AI が人間を滅ぼしかねない」といった漠然とした不安を背景としていることも手伝っているのかもしれない。

人工知能に関連する倫理問題についての論文や書籍は枚挙にいとまがないほど出ている。そしてこれらの中ではより具体的な問題に則した詳細な議論がなされている。人工知能が現在のように騒がれる前から、「コンピューター倫理学」や「ロボット倫理学（robot ethics/roboethics）」、あるいはこれらを包含するより包括的な「情報倫理」といった応用倫理学の分野が存在しており、そこで扱われた問題の多くは人工知能倫理学にも受け継がれている。例を挙げれば、コンピューターを介した行為に関する責任の所在について、自律的兵器の是非について、ソーシャルロボットの欺瞞性について、ロボットの道徳的行為者性・被行為者性について、ロボットの権利や責任について、などである。

特に現代の人工知能に特有、かつ喫緊の重要性を持つ倫理的問題としては、ビッグデータに基づく機械学習によるプロファイリングが挙げられるだろう。現在、機械学習によって個人の様々な属性や行動傾向を予測することが可能になっている。典型的な使用例はオンラインショッピングでの商品の推薦であるが、企業などの人事、裁判、警察、保険、金融などでも活用が広まっている。しかしこれらは、人工知能が設計者や社

（注3）：https://www8.cao.go.jp/cstp/moonshot/sub3.html
（注4）：https://futureoflife.org/ai-principles/
（注5）：http://ai-elsi.org/report/ethical_guidlines

（注6）：江間 [6] は、これらの倫理的原理や規制の分かりやすい解説と、様々な原則の比較を提供する。

会の持っている差別的な偏見を学習してしまうという深刻な問題を引き起こしている[7]。

4. 人工知能の倫理をどのように教えるか

この節では私が授業や一般向けのセミナーなどで人工知能の倫理について話した際の経験に基づいて、どのように人工知能の倫理について教えるのが良いかということについての私の見解を述べる。ただし授業の効果についてきちんとした調査・分析をしたわけではないので、個人の印象と経験則に留まることはお断りしておく。

上述のように人工知能、AI という名で呼ばれるテクノロジーは多様であり、かつ特に昨今のブームの中では AI という言葉が濫用されている。そこで人工知能の倫理について教える時には、まずどのようなテクノロジーについて話をしているのかを明確にする必要がある。

ただしこれは「人工知能」、「AI」という言葉の正確な定義を与えるという意味ではない。おそらく人工知能を過不足なく定義するなどということは不可能であり、試みても不毛である。とはいえ導入においては簡潔な言葉で人工知能とは何かを説明してほしいという需要はあるだろう。そこで私は、ジェリー・カプラン [10] の、人工知能分野の発展は単なる「絶え間ない自動化の進展」（p. 17）であり、「知能」という言葉にとらわれるのは無用な混乱のもとである、という見解を紹介することにしている。

具体的に人工知能のイメージをつかむためには、やはりこの分野においてどのようなアプローチが試みられ、どのようなテクノロジーが生み出されてきたのかを大まかに（しかし不正確にではなく）スケッチする必要がある。もちろんこれまでの人工知能の歴史を網羅的に紹介することはよほどの時間を費やさなければできないので、代表的な例をいくつか選ぶことになるだろう。時間があり、受講生の意欲がついてくるようなら記号論理学、計算論の基礎の話からしても良い。記号的 AI、エキスパートシステム、包摂アーキテクチャ、ニューラルネットワークなどは時間をとって紹介する価値がある。もちろん最も時間を割いて紹介するべきなのはビッグデータにもとづく機械学習システムである。なぜならこれが現在最も広く使われ、最も大きな利益を上げており、そして同時に最も深刻な問題を引き起こしている人工知能の種類だからである。

人工知能について説明する際には特にそれぞれのアプローチの原理、強みと限界、期待されたことと達成できなかったことを示すことが有益である。そのことは人工知能についての幻想（ポジティブなものであれ、ネガティブなものであれ）を取り除くことにつながる。

カーツワイルのシンギュラリティ論やボストロムの超知能論をどう扱うかはなかなか悩ましい問題である。これらは人工知能についての（特に欧米の）人々の捉え方や感情に大きな影響を与えている。しかしこれらについて詳しく話をすると、話の一部だけが印象に残って、「人工知能は何だかものすごいらしい」と間違って覚えて帰ってしまう受講生もいる。私の友人は、クリティカル・シンキングの授業の中で、疑似科学の例として「水からの伝言」を紹介したところ、コメントペーパーに「やっぱり言葉は大切なんですね、感動しました！」と書いてきた学生がいたそうである。なのでシンギュラリティ論や超知能論を教える時は、しつこいくらい「こういう説を唱えている人もいるという話で、現実の人工知能はまだまだこれには程遠いのだ」ということを強調する必要がある。

シンギュラリティ論や超知能論に触れるのであれば、マーク・クーケルバーク [5] のやり方に倣うのがよいかもしれない。クーケルバークは人工知能をめぐるこれらの言説を、西洋の伝統的なナラティブの系譜の中に位置づけることを試みている。クーケルバークによれば、古代ギリシャのプロメテウスの神話、中世のユダヤ教のゴーレム伝説、近代の『フランケンシュタイン』や『ロッサムの万能ロボット会社』、現代の『2001年宇宙の旅』や『マトリックス』など、有用であるはずの技術が制御できないものとなり、それを使うものに災厄をもたらすというナラティブは、西洋の精神文化において脈々と語り継がれてきたおり、AI 脅威論はその現代版である。またシンギュラリティ論やトランスヒューマニズムにはユダヤ・キリスト教的な終末思想、プラトン主義的な魂についての思想が現れている。このように現代の AI をめぐる言説を伝統的なナラティブの系譜に位置づけることは、極端な言説をより適切に理解することに役立つ、とクーケルバークは言う。

しかしやはり私たちは、データやアルゴリズムのバイアス、それによって生じる不公正、個人データとプロファイリングの濫用、フェイクニュースの拡散といった、差し迫った現実的な脅威に重点を置くべきである。

(注7)：これについては例えば [2], [7], [14] を参照。

その際には抽象的一般的な話だけではなく、不正、不合理、非倫理的な実践、すなわちバッド・プラクティスの例に沿って教えるのが効果的である。それはまた現在の情報環境、特にインターネットやスマートフォンの上に成り立つプラットフォーム・ビジネス、それらに対して脆弱な人間の心理と情動というより広い文脈の理解を伴っていなければならない。さらに人間がなぜそのような心理と情動を発達させたのかということを、人間にとっての情報やコミュニケーションの価値というより根本的な観点にまでさかのぼって考えられるとなお良いだろう。

またビッグデータや機械学習の負の面ばかりを強調してしまわないように気を付けるべきである。そういった人工知能は人間の生活を向上させ、経済を成長させるために使うこともできるし、実際に良い使い方もされている。やはりここでも具体的な例、グッド・プラクティスに即して説明するのが良いだろう。しかしこの時、ナイーブなテクノロジー中立論——およそテクノロジーというのはそれ自体としては価値中立であり、使い方次第で善くも悪くもなるという考え——のように受け取られないようにすることも重要である。テクノロジー中立論は非常に人気のある考え方で、学生はほぼデフォルトでそれを信じていると想定しても構わないほどである[8]。従って私たちは意識的に学生からナイーブな中立論を取り除くようにしなければならない。個々のテクノロジーには悪用しやすさの度合いにおいて大きな違いがあるということ、人工知能やビッグデータは実際に濫用・悪用が容易なテクノロジーであること、そして人工知能の悪用を防ぐには多大な努力が必要であることを受講生に理解させれば上出来である。

5. おわりに

本稿では、人工知能にはどのような倫理的課題があるかということ、およびそれらをどのように教えればよいかということについて論じた。私は人工知能の倫理について教える際には以下の点に留意をすることが有益であると考えている。

- 人工知能がどのようなテクノロジーであるかを、実際に提案されてきたアプローチに即してざっくりと理解させる。
- シンギュラリティ論や超知能論について話すときには、それはまだ実現されていない技術についての想像であることを強調する。
- 差し迫った現実的な問題について、具体的な例に即して話す。
- 現在の人工知能を、より大きな情報環境と社会環境の中に位置づける。
- 人工知能の良い使いかたにも目を向ける。
- 学生からナイーブなテクノロジー中立論を取り除くようにする。

謝辞 本研究は、JST ムーンショット型研究開発事業 JPMJMS2011、JST CREST JPMJCR20D2、JSPS 科研費 JP19H00518、JP18H00608 の支援を受けたものです。

(注8)：そしてこの原稿を読んでいる研究者・教育者の中にもテクノロジー中立論の何が間違いなのか、と疑問に思っている方がいるだろう。そう思った方は [11] をご一読されたい。

文　献

[1] 新井紀子、『AI に負けない子どもを育てる』、東洋経済新報社、2019 年。
[2] Ruha Benjamin, *Race after Technology: Abolitionist Tools for the New Jim Code*, Plity Press, 2019.
[3] ニック・ボストロム、『スーパーインテリジェンス——超絶 AI と人類の命運』、倉骨彰訳、日本経済新聞出版社、2017 年。
[4] エリック・ブリニョルフソン、アンドリュー・マカフィー、『ザ・セカンド・マシン・エイジ』、村井章子訳、日経 BP 社、2015 年。
[5] マーク・クーケルバーク、『AI の倫理学』、直江清隆・久木田水生・鈴木俊洋・金光秀和・佐藤駿・菅原宏道訳、丸善出版、2020 年。
[6] 江間有沙、『AI 社会の歩き方——人工知能とどう付き合うか』、化学同人、2019 年。
[7] Virginia Eubanks, *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*, St. Martin's Press, 2017.
[8] 井上智洋『人工知能と経済の未来——2030 年雇用大崩壊』、文藝春秋社、2016 年。
[9] ジェリー・カプラン、『人間さまお断り』、安原和見訳、三省堂、2016 年。
[10] Jerry Kaplan, *Artificial Intelligence: What Everyone Needs to Know*, Oxford University Press, 2016.
[11] 久木田水生、「人工知能と人間のよりよい共生のために」、『RAD-IT21 WEB マガジン』、2020 年。https://rad-it21.com/ai/kukita-minao_20200317/
[12] Quoc V. Le and Marc'Aurelio Ranzato and Rajat Monga and Matthieu Devin and Kai Chen and Greg S. Corrado and Jeff Dean and Andrew Y. Ng, "Building high-level features using large scale unsupervised learning", arXiv:1112.6209v5 [cs.LG], 2012.
[13] 村上祐子、「人工知能の倫理の現在」、『電子情報通信学会基礎・境界ソサイエティ Fundamentals Review』、11 巻 3 号、p. 155-163、2017 年。
[14] キャシー・オニール、『あなたを支配し社会を破壊する AI・ビッグデータの罠』、久保尚子訳、インターシフト、2018 年。

（xxxx 年 xx 月 xx 日受付）

久木田水生

2005 年、京都大学大学院文学研究科より博士学位（文学）を取得。2017 年より、名古屋大学大学院情報学研究科准教授。専門は哲学、倫理学。著書に『ロボットからの倫理学入門』（共著、名古屋大学出版会、2017 年）など。

Abstract Currently, artificial intelligence is causing a variety of ethical problems. In order to develop artificial intelligence in a sound way, it is important to educate students about the ethical issues of artificial intelligence and its solutions. This paper discusses how the ethical issues of artificial intelligence should be taught.

Key words AI, data science, information ethics