クーケルバーグ『AI の倫理学』[1]の概要

文責:鳥居 千朗

本書は、ウィーン大学哲学部で技術の哲学の教授を務めつつ EU 圏の専門家会議の委員として、AI に関する政策提言も行っているマーク・クーケルバーグの手による、AI の倫理的問題への入門書です。AI にまつわる問題としてこれまで何が語られてきたか、それらが実際どれほど現実的で緊急の課題なのか、またそもそも、それらを問題にすること自体に問題はないか、といった重層的な視点から、今日の目の前の倫理的問題に取り組む一般的視点が提供されます。訳者・直江清隆のまえがきにあるように、本書は**「AI の倫理学について概観する最良の手引き」**(iii 頁)となると同時に、公平なアルゴリズムを開発するにはどうすればよいかといった技術面での問題だけでなく、「AI が人間を超える」といった物語の歴史的な由来がどこにあるのかといった人文学的な観点をも踏まえることによって、そこから、**そもそも人間とは何か、正義とは何か、といった、より普遍的な哲学的論点**にも踏み込んでいるという点で、単なる入門書を超える内容を含んでもいると言えるでしょう。

目次

まず、本書の各章のタイトルと、そこでどのような内容が語られているかを確認しましょう。

第一章　鏡よ、鏡……

> 本章の中心的なテーマ・主張：
>
> **AI の技術の現在。それが孕んでいる様々な問題の可能性。**
>
> AI は飛躍的な発展を遂げ、今や生活のいたるところに浸透している。しかしそれは同時に、人間の安全やプライバシー、自律性、平等性を脅かす大問題をも伴っている。本書全体で扱われる問題の提示。

第二章　スーパーインテリジェンス、モンスター、そして AI 黙示録

> 本章の中心的なテーマ・主張：
>
> **AI についてありがちな言説、それに加担することの危険性。**
>
> 「AI が人間を支配するかもしれない」というディストピア的な問題を立てること自体が偏った見方である。そうした態度自体が特定の集団に利益をもたらし、特定の集団を不利な立場に立たせることになり得る。

[1] Mark Coekelbergh, *AI Ethics*, Cambridge, MA: The MIT Press, 2020. (=マーク・クーケルバーグ『AI の倫理学』直江清隆ほか訳, 丸善出版社, 2020.) 以下、本書への参照指示は、原則として訳書の頁数のみ記す。

第三章　すべては人間のこと

本章の中心的なテーマ・主張：

AI による人間の代替可能性をめぐる哲学的議論の紹介と検討。

人間になり代わる知能を持つ AI の実現は技術的にもまだまだ現実的ではない。さらに、そもそも人間の本質は知能ではない。AI は人間に似せて作られる必然性もないし、技術と人間は対立するものなどではない。

第四章　ただの機械？

本章の中心的なテーマ・主張：

AI の道徳的行為者性と被行為者性に関する哲学的議論のまとめと検討。

AI の判断は道徳的問題に関わる帰結をもたらす。では、AI は人間の道徳的価値に適した意思決定を行えるのか。そもそも人間は道徳的価値を十全に理解しているのか。また、AI の方を道徳的に扱う必要があるかどうかは、動物の場合のように、社会的・歴史的にも規定されるものであり、常に流動的に対応しなければならない。

第五章　AI という技術

本章の中心的なテーマ・主張：

AI の歴史と本質的機能、そして倫理的懸念。

AI が生まれた歴史。AI には、フローチャートに則る記号的 AI と、統計処理に基づく機械学習型 AI がある。前者は既に社会の到る所で利用されている。それは一部のメリットをもたらしつつも、特定の人々にリスクを与えている。

第六章　データおよびデータサイエンスをお忘れなく

本章の中心的なテーマ・主張：

なかでも、学習型の AI はどのような仕組みなのか、その問題点。

近年きわめて流行している統計的学習 AI は、人間が関与する仕方に応じて「教師あり supervised」「教師なし unsupervised」等に大別されるが、いずれも計算過程が人間には解釈不可能なブラックボックスである。しかし、学習によるパターン形成は決して普遍・中立的なものではなく、また計算結果が何を意味するのかは最終的に人間の判断によることを踏まえれば、人間による監督がどこかで必要である。

第七章　プライバシーやいつも挙げられるその他の問題

本章の中心的なテーマ・主張：

プライバシーとセキュリティの問題。人間の根本的な脆弱性。

AI は職場や街頭での監視にも利用されており、市民やユーザーの脆弱性と搾取が問題になるため、取集されたデータがどのように使われているのかということに関する透明性と、データプライバシーの諸権利が守られなければならない。また AI は、フェイクニュースや兵器の制御に加担することによっても、人間の自律性や安全を脅かしている。

第八章　責任能力を欠いた機械と説明不可能な意思決定

本章の中心的なテーマ・主張：

責任の問題。AI の行為を理解し釈明する責任を果たすことの困難さ。

AI の責任を開発者などに求めるにしても、誰がどこまで責任を負えるのかを特定することは難しい。ブラックボックスの中身は誰も説明できないし、「説明」とはただ因果関係の連鎖を記述することではないはずであり、人間が説明責任を果たせるような仕方で AI を開発することが、AI 研究者に与えられた困難な課題となっている。

第九章　バイアスと人生の意味

本章の中心的なテーマ・主張：

バイアスと労働の問題。AI は既存の差別・格差を強化しうる。

AI の設計、試験、利用、あらゆる段階でバイアスが発生する。技術的問題だけでなく、アファーマティブアクションといった政策もあるように、そもそもバイアスの無い AI の方が不正義である可能性も考える必要がある。また、どの労働を AI に委託せず人間にとっておくかを考える必要がある。人間が労働から解放されるという言説が富裕層の幻想である可能性もある。

第十章　政策提言

本章の中心的なテーマ・主張：

実際に世界で提出されている AI 政策の紹介

2010 年代以降、様々な国家や国際 NPO 等から AI の透明性とリスクに関する政策が提出されている。特に重要なのが、トレーサビリティ等を搭載するよう設計の段階から組み込まれるべき「デザインが担う倫理」である。

第十一章　政策立案者にとっての挑戦

本章の中心的なテーマ・主張：

「デザインが担う倫理」を導入する際の注意点。原則と現場との架橋

トップダウン式に規範を命じるのではなく、ボトムアップ式に現場や関係者各位の意見を最初から募らなければ民主主義は成立しない。無関心に落ち込むのではなく、積極的に未来の社会を構想する倫理と学際性が必要である。

第十二章　気候こそが重要なのだ、愚か者！　私たちの優先度、人新世、イーロン・マスクの宇宙の車

本章の中心的なテーマ・主張：

環境問題との関わり。技術の発展次第で済む問題ではない

環境問題、SDGs は今最も緊急の課題である。AI 技術もこの課題に貢献し得るが、技術的解決のみを考えていると、結局は格差や人間の自律性、責任といったものを無視した現実逃避に陥ってしまう。政治的・倫理的研究が不可欠である。

こうして一瞥するだけでも、AI に関する倫理的問題の考察が、不可避的に人間存在そのものへの問いかけへと送り返されていくことがわかります。AI の倫理学に真正面から取り組むためには、個別具体的な実態と、人間や世界の普遍的・根本的な問題とを絶えず往復しながら思考しなければなりません。今、世界では何が起こっており、我々は何をなすべきなのでしょうか。そのことを考える助けになるものとして、以下、本書の詳細な要約をまとめました。

第一章　鏡よ、鏡……

2016 年にディープマインド社の AI「アルファ碁」が囲碁の世界チャンピオンに勝利して以来、AI という存在への期待と危機感はより現実味を帯びて「世界」を包んでいる。インターネットでは AI がユーザーを広告へ誘導し、かたや自動運転車の実現に向けた様々な実験が進められ、軍では AI を搭載したドローンや自律型殺傷兵器が運用されている。2018 年には自動で美容室やレストランの予約電話をかけてくれる「グーグル・デュプレックス」が発表され、アメリカ、カナダ、イギリスなど 8 か国で実際に利用されている[2]。アメリカでは AI による犯罪予測が実施され、裁判所でも COMPAS システムが意志決定の材料となっている。もはや AI は日々の生活における便利な助手であるに留まらず、そもそも我々の生活を形作り、あるいはその運命を左右する存在になっていると言えるだろう。

すると、**我々人間は AI にどこまで任せるべきなのだろうか。**実際、アルゴリズムに基づいた犯罪予測は、例えば白人よりも黒人の再犯率を高いものと算出してしまうバイアスを抱えていることが報告されている。我々はこの AI の結論をどこまで鵜呑みにするべきなのだろうか。あるいは、殺傷機能を持った機械に自律的 AI を搭載するようなことはあってよいのだろうか。監視システムに搭載された AI は、我々の顔を認識し、そこからプライベートな感情や嗜好まで読み取っている。また深層学習型 AI を利用すれば、超高精度のフェイク動画を作成することで有名人の架空のゴシップを拡散することもできる。このような様々な問題があることを踏まえると、AI 技術の発展による暗黒の未来、といったことを語りたくもなってくる。しかしここで、これらの倫理的諸問題と共に、次のようにも問わなければならない：**第一に、AI 技術を改良することで利益を得ているのは誰で、損害を被っているのは誰なのか？第二に、「AI が人間に取って代わる」といった類のビジョンを語り流布させることで利益を得ているのは誰で、損害を被っているのは誰なのか？**——AI 技術とは一体、市民、政府、警察、被告人、大企業、誰のためのものなのだろうか？本書では、これらの問題が扱われていく。

第二章　スーパーインテリジェンス、モンスター、そして AI 黙示録

AI がどのような未来をもたらすかを示す様々な物語は、SF の題材の定番であると同時に、いまや現実感を伴って様々な立場の人間が語り、各人が自らの人生を考える材料ともなっている。自律的に自らをアップグレードしていくウルトラ・インテリジェント・マシーンという考え方はその実、既に 1965 年、数学者 A. J. グッドが構想したものであった。2005 年、グーグル社のカーツワイルは、人間が機械知能と融合して生物学的制約

[2] Google Help, About phone calls from Google Assistant.
https://support.google.com/business/answer/7690269?hl=en (2021/06/11 閲覧)

を超えることを「シンギュラリティ」と呼び、これの到来を 2045 年頃と予言した。AI の急速な発展は、それが人類の存亡を脅かしているといった発想や、そこから導かれる様々なビジョンを喚起している。例えば、ボストロムは、苦痛を理解できない AI が、生産効率の最適化のために人間身体を材料にする選択を取る危険性を指摘した。このような危機感から、むしろ AI や機械を積極的に取り込んで活用することで人間の能力を増強し、人間を超えていくべきとするのがトランスヒューマニズムであり、AI を制御することを不可能として、火星移住かさもなくば、人類滅亡かの二者択一をつきつけるのがイーロン・マスクのビジョンである。

しかし、そもそも人知を超えて人類を支配する AI、といったものは実現可能なのだろうか。例えば、M.ボーデンはこれをあくまで理論的な可能性でしかなく、全く現実的なものではないと批判している。AI はあくまで機械であって、それをどのように具体的に活用するのかを第一に考えるべきである、ということだ。先ほど確認したようなビジョンの描き方には、様々な問題が含まれているのである：第一に、いつかの未来に実際にスーパーインテリジェンスが完成するかどうかの判断は留保するとしても、そのように**遠未来ばかりを考えること自体が、目の前の近い将来の現実的リスクを隠蔽してしまう**、という点がある。第二に、「人間の知能を超えること」と、「人間を超えること」とは同義なのだろうか。AI による支配をめぐるシナリオは、その前提からして、**人間の本質を知能というものに限定する思想を強化してしまう側面がある**のだ。――ということは、「AI は人間を超えるかどうか？」という、AI と人間の〈競争の物語〉を語ること自体が、全く中立的な選択ではなく、一定の人々に利益をもたらし、他方の人々を不利な立場に追いやってしまう、という事実があることに注意しておかなければならない。

ここに、文学史的・物語論的な研究が AI 倫理学にとって必要であることの理由がある。例えば、無生物から生物を作り出すという発想自体は古来、洋の東西を問わず人類に共通のモチーフだが、人間が過剰な技術を手にすることで災いを招く、という雛型は特に西洋の伝統において受け継がれてきたものであった。ギリシア神話において神々から火の技術を盗んだプロメテウスは岩壁で心臓をついばまれ、中世のラビ物語における賢者は粘土からゴーレムを作り出すもその制御を失ってしまう。こうした流れを汲んで、1818 年の『フランケンシュタイン』では、自らの生み出した人造人間への責任を放棄した結果、復讐を受けて破滅していく科学者の姿が描かれた。西洋式の物語の雛型の一つは、生命を自ら創造すうことへの憧れと、それによって人間が滅ぼされることへの恐れとが入り混じった**フランケンシュタイン・コンプレックス**である。さらにここに、身体・物体の世界から精神を解き放ち英知的な世界へ帰る、という**プラトン主義・グノーシス主義的な観念**と、遠からず破局が訪れ人類が決定的な篩にかけられる、という**キリスト教的終末論**とが結びついて、今日広く流布している AI と人間の〈競争の物語〉、トランスヒューマニズム、シンギュラリティ、といったビジョンが形成されているのである。

しかし恐らく、これだけが唯一の AI と人間との関係の捉え方ではないだろう。第一に、非西洋圏の物語の形に触れることで、例えばアニミズム的思考における機械と人間の友好性のような新たな理解の可能性を開くことができるかもしれない。第二に、古典的な語り方を批判するためにも、実際にいま現在、AI はどのようなものであるのかを落ち着いて確認し、近い将来への展望を開くことが重要だ。

第三章　すべては人間のこと

まず、AI が人間になり代わるという可能性は、どれほど現実的なことなのだろうか。哲学の領域では、いわゆる独仏・大陸哲学の潮流に属する思想家たちがこの可能性に対し懐疑的であり、英米・分析哲学の論者たちが積極的である、という大まかな傾向がある。ヒューバート・ドレイファスは、ハイデガーやメルロ=ポンティの哲学を活用して、人間には記号化・形式化できない身体的な暗黙知があり、これを AI による記号操作・演算によって獲得することはできない、と主張した。**人間においては身体と精神が分かちがたく結びついた抽象化不可能な認知・行動がある一方、AI は記号や命題規則を処理し入力と出力を対応させるだけである、というのが懐疑派**の基本的な論調だ。対して、チャーチランドは脳を再帰的ニューラルネットワークに還元する消去主義的唯物論の立場をとり、デネットも身体の内にないような存在はない、として、人間を一種のロボットに見たてている。**積極派では、人間の心をコンピュータモデルによって完全に理解・再現可能なものと捉える**向きがあるのだ。両陣営とも心身二元論を否定するという点で一致していながら、一方はそこから心身の結びついた具体的で神秘的な人間の姿を描き出し、他方は精神を身体・物体に還元して、通常「精神」と思われているものの抽象化・再現を目指す。しかしどちらの立場をとるにしても、人間になり代わる AI がすぐにでも実現する緊急の課題なのかは別の問題である。

ところで、この論争において前提にされていることは、AI が人間をモデルにしている、という認識である。これは前章でも確認したように、西洋において伝統的に〈競争の物語〉を巡る種々の論争を生んできた根本的なモチーフの一つになっている。「AI は人間になりうるか？」という問いの立て方は、ある意味で、18,19 世紀の科学・合理主義と人間の神秘を重んじるロマン主義との対立の反復であるとみなすことができ、これはそのまま、トランスヒューマニズム(超人間主義)とヒューマニズム(人間主義)との対立に反映されている。この技術をめぐる人間中心主義を克服し、〈競争の物語〉を克服することは、一つの新たな視点を提供するかもしれない。

第一に、ヒューマニズムにしてもトランスヒューマニズムにしても、それが人間を中心に考えられた技術を前提にしているという点でヒューマニズム(人間中心主義)と呼んでよいとすれば、この意味でのヒューマニズムに、**ポストヒューマニズム**と呼ばれる潮流を対置することができる。この立場は、人間をあくまで非人間的な存在者との水平的な関係の内に置き、どちらか片方を能動的主体と見なすのではなく、どちらも相互に働きかけあって一つの新たな在り方を共に作り上げていく同等な存在者と考えることで、人間/非人間の垣根を存在論的にも政治的にも越境しようとする試みを旨として、B.ラトゥール等によって提唱されているものである。トランスヒューマニズムのように科学技術の側から一方的に境界を踏み越えて既存の人間のあり方を否定してしまうのではなく、人文学・芸術・政治学の知見をふんだんに取り込むによって、人間と技術とがそもそも自立し対立するものではなく、お互いを前提し合って未知の可能性を持つ新たな世界を生み出していく運動を詳細に明らかにすることで、人間/非人間の非対称性を克服するとされている。こういった思想からは、**AI を含む非人間的なものは人間をモデルにされる必要もないし、実際にそうされるべきでさえない**、という示唆を得ることができる。

第二に、ハイデガーの技術論を刷新することから始まった**ポスト現象学**と呼ばれる近年の潮流は、人間と技術との関係を、本来対立させられるべきではないものとして描き直す。ドン・アイディは AI や電子データといっても何か宙に浮いたものではなく、あくまで物質的下部構造(インフラ)に基づいていること、人間によって設計・操作・調整されるものであることを強調し、現代の哲学者は研究開発チームの現場に参画して未来

のデザインを担うことができる、と主張する。ピーター=ポール・フェルベークは、**技術とはそもそも人間に内在的なもの**であり、人間が存在し始めたその瞬間から、人間と世界との間を媒介するものとして作用してきたものであると指摘し、「人間と技術は対立する」、ましてや「人間を技術から守るべきだ」と考えることのナンセンスさを強調している。本当に、そして今から考えるべきことは、「AI 技術は人間と世界をどのように媒介することになる(べき)か」だということになる。

以上のように、いま広く流布している AI 観がどれほど歴史に規定されたものであるかを確認することで、新たな思想的アプローチの方向から、人間と AI との関係を考えるきっかけを得ることができる。しかしこの方向は我々のこれからに対して一般的な指針を与えてくれる一方で、目の前の緊急の具体的な問題に着手するには未だ抽象的なものに留まってしまうであろう。AI に特有の倫理的問題としていかなるものがあるのかを把握しておくことが不可欠になる。

第四章　ただの機械？

そもそも、AI について倫理的問題が発生するということは、AI が何らかの仕方で倫理・道徳に関わる立ち位置にあるということを含んでいる。AI はどのような道徳的地位を持っているのだろうか。

第一に、AI の自動運転車か通常の自動車かを問わず、その技術が介在した出来事の結果には、例えば路上の子供が救われた代りに乗員が犠牲になった、というような道徳的(な問題を孕む)帰結が伴う。しかし、通常の自動車と異なって AI はさらに知能を持って判断を下すとした場合、AI は運転手に代わって道徳的な意志決定を行うことができ、実際にこの帰結に関してもそうだったのだと考えるべきなのだろうか。「**AI は道徳的行為者性 moral agency を有するか**」というこの問いに対し、否定的な論者は感情や自由意志を道徳的行為者性の必要条件とし、人間のみがこれを持つと考える。肯定的な論者は、人間と同じ道徳性を諸々の原則として機械に与えることができ、ともすれば機械は人間よりも合理的にその都度優れた判断を下すことができると考える。また、中間的な論者は、十全ではなく**限定的ではあるものの、一定の「機能的道徳性」を予め実装しておく**必要があると主張する。実際に目の前に子供が飛び出したとき、人間の運転手はほとんど考える間もなく判断を下さなければならない。この熟慮と判断の結果を予め開発段階で組み込んでおく、という選択はあるのだ。

しかし、そのようにして AI に**人間の**道徳的振舞を学習させるやり方自体が、道徳的に正しい選択ではないという可能性はないだろうか。つまり、人間は知的な側面だけでなく感情的な側面も含んだ仕方で振る舞う限り、そのような**人間のあり方は道徳的行為者の手本として不適当なのではないか**、という疑問もあり得るだろう。トランスヒューマニズムの立場からすれば、AI の冷静で自律的な判断の方が人間よりもずっと優れた道徳を持つことになり、ポストヒューマニズムの立場からすれば、災害救助犬やメールのフィルタリングボットといった非人間的存在者も道徳的行為者と捉えられるような仕方で、道徳性概念を抽象化・再構成する必要があるとされる。しかしその反面、このように人間の理解を超えた尺度に託したり、あるいは人間を単なる一つの特殊として含む普遍的なものへ拡張したりされた道徳概念というものが、そもそも我々の問題の出発点としては不適当なのではないか、という懸念もあって然るべきものだ。つまり、**道徳性というものはそもそも人間性・人格性に結びついた限りでしか思考できないもので**

あるのかもしれない。「人間の理解を超えた道徳」や、「人間目線ではない道徳」といったものは、語義矛盾であるかもしれないのだ。

第二に、AI の道徳的地位を問題にする、もう一つ反対の方向がある。つまり、我々は AI を行為の対象とするとき、AI に対してなんらかの道徳的責任を負うのだろうか。高度な知能を持つ AI の電源を切ったりこれを破壊したりすることは、人間に対してそうするように、「殺害」にあたるのだろうか。「**AI は道徳的被行為者性 moral patiency を有するか**」という問いがあり得ることは、例えばボストン・ダイナミクス社のロボット犬「スポット」がスタッフによって蹴られている映像に対し同情的な反応が寄せられたという事実からも見て取れる。もし仮に AI には道徳的に振る舞われるいわれはない、といった結論が導かれたとしても、こういった我々の事実的な経験がある限り、その結論は人間の実情を無視したものになるだろう。したがって、もしお望みならば、カントが「コリンズ道徳哲学」において語ったように、別の推論をたどって、機械や動物に道徳的被行為者性を持たせることも可能であろう。即ち、我々が犬を殺してはいけないのは、犬に対して道徳的責任を負わなければならないからではなく、犬を殺してしまうと、我々の内にある、人間に対する道徳的性格も同時に損なわれてしまうからという間接的理由によるのである、と。人間に対して道徳的に十全に振る舞えなくなる、という点を持ち出すことで、犬を殺してはいけないことを正当化するという方法である。

しかし、AI の道徳的被行為者性をめぐるこのような義務論や徳倫理学は、その成果以前に、我々の道徳に関わる経験の実態を見落としているのではないだろうか。というのも、現に我々は以上のような複雑で間接的な思考を経ることなく、ロボット犬や非ロボット犬に対して攻撃的に振る舞うことを躊躇しているからだ。我々は人間に対して道徳的に振る舞うことを自明としているが、しかしその相手が本当に人間であってロボットではないという保証はない。つまり、普段我々が**相手に道徳性を付与するために、相手の「中身」を知る必要は全くない**のである。我々は、相手の見た目や振舞に基づいて、相手に対し道徳的関係を築こうとしているのだと、我々の知っている道徳とはそのようなものだと、臆面なく言うべきなのではないだろうか？さらに、相手との関係はお互いの**社会的位置によっても変化**してくる。我々は自宅や近所のペット相手には極めて親密な道徳的関係を築く一方、食卓に並ぶ料理の原材料となったどこかの見知らぬ家畜に対しては希薄な関係しか持たない。さらにまた、当然**この関係は歴史的・文化的にも大きく変化し規定されていく**ものである。かつての動物－人間の地位と、現代の動物－人間の地位とは大きく変わっている。

以上のことから導かれる結論は、あくまで未来における変化に開かれた相対的な時点として、**今日のうちに性急に AI の道徳的地位を決定してしまうことは望ましくない**、ということである。関係の変化によって道徳的地位も変化するという我々の経験の実情を無視して一方的に AI の地位を裁いてしまうことは、抑圧的な暴力として働くことになるだろう。また、このように実情に即することによって、自動運転車やバイアス、セックスロボットなど、一層現実的で具体的な問題も考えることができるようになるはずだ。

第五章　AI という技術

さて、AI という存在をめぐってどのような問題が引き起こされるかを概観したところで、そもそも AI というものがどのような仕組みで、どのように生まれてきたのか、つまるところ、AI とは一体何のことを言っているのかを確認しておこう。

一言で言えば、AI とは人間から見て知的と言い得るものを機械によって再現したものであるとフィリップ・ジャンセンは言っている。科学の一分野として考えればそれは知能という現象の体系的解明のことであり、技術の面で考えればその能力を実際的目的に運用する試みだと言える。その歴史をたどってみれば。1940 年代にデジタル・コンピュータが発明され、ノーバート・ウィーナーが 1948 年に提唱した『サイバネティクス』が生命体の働きと情報通信技術とを統合的に理解し、1950 年にアラン・チューリングが学習と抽象的思考の能力を持った機械の可能性を示唆したところから、1956 年、ダートマス・ワークショップにてジョン・マッカーシーが Artificial Intelligence という言葉を作り出した。

AI 技術というものは、言うまでもなくあくまで機械の一部であり、そうである限りは、**ロボットである必要はないし、ましてや人型である必然性もない**。今日、AI は様々なソフトウェアや自動車の中に、また機械ではないもの同士を遠隔で繋ぎ通信ネットワークを形成する IoT といった技術の中に導入されて、ソフトウェアとハードウェアという区別をかき乱している。また、その機能が自然言語処理や人間的コミュニケーションに特化したとき、AI はアレクサや Siri といったソーシャルロボットの形を取ることになる。

そして当然、計算機械の一部である限りは、AI の基本は何らかのアルゴリズムである。AI の計算過程の仕組みは、1980 年代まで「**記号的 AI**」と呼ばれるものが主流で、これは今日も、診断の作成や治療計画の案出などの場面で運用されている。これは、事前に当該分野の専門家が、自分たちの推論・判断の仕方を指示し、これをプログラマーがフローチャートやデータベースに起こすことで、複雑な因子や膨大な資料に依拠しつつ正確で迅速な判断を再現できるようにしたもの(エキスパートシステム)である。この仕組みの AI は一つのフローチャートを辿る計算であるため、最終的な判断に対し、その過程を遡って、当の AI がどこでどのような思考を行ったのかを理解し評価することができる。ヒューバート・ドレイファスが人間のあり方からはかけ離れているとして批判したのもこのタイプの AI であった。

しかし 1980 年代以降、コネクショニズムという立場が台頭してくると、AI 設計の仕組みは、推論方式のような高次な認知機能を再現することではなく、極めて単純で低次なユニット同士を大量に接続させることによって網を形成(ニューラルネットワーク)し、それによって AI に自律的な学習を行わせることを主軸とするようになった。これはあくまで統計的処理に基づく判断を下すものであり、出力された判断を見ても、人間はそれがどのような過程を経て出てきたものなのか、計算過程の一つ一つを見ても一切理解することができない(**ブラックボックス**)。だから、アルファ碁がどうやってチャンピオンを打ち負かしたのかは、誰にもわからないのである。

その他にも、今日 AI を設計するさまざまな試みは、身体を持って環境と相互作用するヒューマノイドや、人工生命や、遺伝的アルゴリズムによって進化していく AI などの姿をとっている。こういった仕方で再現された知能は、いまや我々の生活の到る所に浸透している。SNS ボット、医療・金融におけるデータ分析、自動運転、ビデオゲームのキャラクター、地方紙の記事執筆、学問分野での活用(デジタルヒューマニティーズ)、パーソナライズ広告、アレクサ、オリジナルレシピの作成……、人類を超越する存在ではなくとも、**一つの分野に特化した AI で言えば、それは既に社会に普及している**。そもそも AI は他の何かしらの技術の中に組み込まれて初めて意味を持つものであって、「AI それ自体」といった何かがあるのではない。しかし AI 技術とそうでない技術との境界が曖昧になっているという事態は、AI の存在自体を我々に対し隠蔽してしまう恐

れもある。我々は現状を正しく認識しなければならない。AI が普及していることによって、プライバシー侵害や貧富の格差の拡大、戦争技術の変化、環境汚染などの問題もあれば、新たなコミュニケーションの創造、人間のタスク・リスクの削減、サプライチェーンの改善、水使用量の削減などの利点も浮かんでくるだろう。大切なのは、**AI の発展が誰に利益をもたらすのか、先進国に利益をもたらすのではないか、誰に被害をもたらすのか、発展途上国に被害をもたらすのではないか**、ということを考え、知っておくことである。

第六章　データおよびデータサイエンスをお忘れなく

とりわけ、今日の先進的な AI 技術の主流である**統計的 AI** の仕組みをより詳しく見てみよう。統計的 AI は「学習」するが、しかしこの機械学習は、あくまでビッグデータから何らかのパターンを認識(マイニング)し予測に活用するという統計的プロセスであって、我々人間の学習のプロセスとは全く類似したものではない。自律的アルゴリズムに目的のみが与えられて、それに適合するソリューションを見つけ出す作業が「機械学習」と呼ばれているのである。そこにおいて人間の役割は、そのつど出力される結果の適合度に応じてフィードバックを与えることであり、記号的 AI と違って、特定の指示や規則を与えることはない。

またこの学習の方式にも、いわゆる「教師あり supervised」と「教師なし unsupervised」、それから「強化学習」がある。教師ありの学習では、前もってプログラマーから既存のカテゴリーが教えられる。例えば、カテゴリーA に属するものとして一群のデータが与えられ、それとは分けられて B に属する一群のデータが与えられることで、そのカテゴリーを学習した AI が次なる予測に向かうのである。教師なしの学習では、ひたすらに与えられたデータから AI が独自のカテゴリーを生成することになる。そこからは統計的にのみ意味をもつ予測計算が構築されるため、最終的に AI が提出する答えは、専門家から見ても理解できない、一見恣意的にしか思われないものになる。強化学習では、そのつど AI が出力する答えに対してその良し悪しを評価し返すことを繰り返し続けることで、最適解を学ばせる方式である。

そもそもこのようなビッグデータを利用することができるようになった歴史的背景としては、技術面で低価格のコンピュータの処理能力が向上したことや、様々な組織において顧客等のデータ収集を行うことのハードルが下がったことなどが挙げられる。今日、我々はネットショップや SNS などのあらゆるデジタル活動を通して組織にデータを提供しているのである。

以上のことを踏まえると、**AI にはあくまで人間の関与が不可欠である**ということも見えてくる。まず、AI が導き出すデータの関連性には、意味が欠けている。現在も医師が AI による治療法の提案を利用しつつも最終的には自らの経験と直観を用いるように、**ある物事が何を意味するのかは人間が解釈することによって初めて言える**ことなのである。また、AI に全く無差別に相関関係を発見させることは、本来は因果的関係を持たない事象を結びつけてしまう疑似相関の危険性を孕んでいる。どの相関関係が有意義で学習に値するものであるのかを判断するのは人間である。そしてそもそも、**データの群の中から一定のパターンを抽出する作業は、決して中立的・普遍的なプロセスにはならない。**一つの土地について地図を作製しようにも様々な観点に則った描き方があるように、アルゴリズムがどのような観点を選択しパターンを抽出しているのか、は人間が批判できるのでなければならない。

第七章　プライバシーやいつも挙げられるその他の問題

さて、以上のことを踏まえた上で、実際に AI が導入されている現場では、どのような問題が考えられるだろうか。まず、AI は例えば**職場や街頭における監視システム**に導入され、画像認識に活用されている。そうして個人のデータが収集されていることを考えれば、ここには当然、プライバシーの問題が関わってくるだろう。より具体的に言えば、個人が、**自らのデータが収集・処理されていることを知る権利、そのデータがどのように利用されているのかを知る権利**など、つまるところ、データが AI によって処理されていることについての透明性の確保が考えられなければならない。これは、社会学の研究者などによって実施されるアンケート調査の場合とはわけが違う。例えば SNS の例をとってみても、ユーザーが自らのデータが何のためにどのように用いられるかを把握することは困難であるし、最終的には、当のサービスの恩恵に浴するためには、その規約に同意する以外の選択肢を剥奪されている。実態はといえば、収集されたデータが企業によって転用・売却される可能性も常にあるのだ。

そうしてデータを提供することに同意した**ユーザーは、極めて脆弱な立場に立たされる**。まず、SNS 等のユーザーは、企業のために無償でデータを産出・採取・提供する労働を行っていることになる。マルクーゼの『一次元的人間』(1962)は、「自由」で「非全体主義的」とされる社会にも、産業体制の内に消費者を取り込み搾取する支配の形態があることを指摘している。2016 年の米大統領選では Facebook ユーザーのデータが政治的に利用されることもあったが、そうした明白な利用に限らず、スマートフォンという名の個人データ収集機を人々が不可欠とするような社会になるように経済が向けられることで、AI を利用した監視社会が一層巧妙な仕方で形成される恐れもある。AI が人間の認知活動を代行することによって批判的思考が失われることの懸念もあれば、そもそも**子供や高齢者**といったユーザーは、最初からより脆弱な立場にある。というのも、我々がプライバシーや搾取を問題にするとき、しばしばユーザーは自律的で若く健康的な成人であるという暗黙の前提が敷かれているからである。そこから漏れ出てしまう子供や高齢者の多くは自らのデータが収集されることに同意する機会も持たず、ましてやそれがどのように利用されているのかを知ることもないのであって、こうした層への目線が欠けている限り、AI 倫理も不十分なものとなるだろう。

一方で、搾取されるのは AI を使う側だけではなく、**AI を作る側の労働**にも目を向ける必要がある。即ち、鉱物採掘者や電子廃棄物処理者、輸送労働者、諸々の仲介業者など、ハードウェアを製造する人々、またアルゴリズムにトレーニングを施す作業をする人々、こういった労働の存在があって初めて、家庭におけるアレクサの「無償の労働」が成立しているのである。

また、AI による成果が直接危険をもたらす可能性もある。AI はインターネットやデジタルメディアの環境と組み合わさって、これまでのどのメディアにも勝って勢力を拡大し、フェイクニュース等による強力な情報操作のために利用される恐れがある。仮に全体主義的支配が成立するのではないとしても、虚偽情報の氾濫は人々の間から信頼と交流を喪失させるだろう。また、AI がその行為者性を拡大し、自動運転車やインフラ、軍事機器などに搭載されていくに伴って、AI がハッキングされた場合の危険性の規模も拡大していく。というのも、高度な専門技術の部門のなかに AI というネットワークへの入り口を作ってしまうことは、外部からのハッキングのハードルを下げることにもなるのである。またそもそも、外部から悪用されるまでもなく、作業効率を最適化する

AI を搭載した機械は、実際の運用にあたって付近の人間の安全をどの程度まで考慮に入れるのかについて判断するようにデザインされることになる。以上のように、メディア、インフラ、軍事機器、工場ロボットやお手伝いロボット、といった形で、我々の生活の様々な部分を AI に委託するということは、すなわち我々の安否を少なからず AI に握らせることなのである。こうして AI の勢力が拡大しつつある現状は、**人間存在そのものが根本的に脆弱であるということを明るみにしている**と言うことができる。

第八章　責任能力を欠いた機械と説明不可能な意思決定

AI に我々の行為を委託することが問題になるのは、まず責任の所在が不明であることに由来する。AI が自らの行為について責任を負うことができないとしたら、何か AI 絡みで問題が発生した場合には、その製造者や利用者の責任を問えばよいのだろうか？このような考え方はひとまず一般的に受け入れやすいものである。アリストテレスの『ニコマコス倫理学』によれば、行為者が自らの行為に責任を持つための第一条件は当の行為がその行為者に起源をもつことであり、第二条件はその行為者が当の行為とその帰結について自覚していることである。この場合、AI は第一条件を満たしつつも、第二条件を満たしていない、と言うことができる。かくして責任の所在が繰り延べされることは、実際に、ペットや幼児の過失についてその保護者が責任を問われたり、組織の責任がその管理者に帰属させられたりしているように、現在も一般的に行われている判断である。

しかし、AI の場合に特に顕著になる問題もある。高頻度取引や自動運転車などにおいて、AI は**人間が介在する余裕もない高速の意志決定**によって重大な判断を行う。この場合、当の判断の責任は利用者に問うことができるのだろうか？仮に問うことができたとして、その責任は利用者にあるのか、提供者にあるのか、製造者にあるのか、どこまで遡ればよいのだろうか？(many hands の問題)また、「AI に問題があった」として、それは具体的に、アルゴリズム、センサー、データ、ハードウェア、ソフトウェア、あるいは学習段階でデータ処理や訓練を施す人間、どこに原因があるのだろうか？そもそも AI というものはこれらの相互関係のネットワークの中で成立しているものであり、**どこからどこまでがどの部門の責任であるということは決定できない**のである(many things の問題)。

また、第二条件のことを改めて考えてみると、そもそも「自覚している」とはどういうことなのだろうか。我々人間の間で起きていることを見てみれば、そこでは常に応答能力と説明能力が問題になっている。つまり、我々が行為者の責任を問うとは、その行為者に当の意志決定に関する説明を求め、これに応答することを期待すること、を意味しているのである。AI は意識を持っておらず、また計算過程や判断のログを残すことはできても、これについて釈明することはできない。自覚能力、反省する能力とはそもそも、他者との関係の中で実現するものなのである。それでは、当の AI に携わる人間が代弁して、AI の判断について他者に説明することになるだろうが、一方、人間でさえ AI の行為については説明できないのではないか、という問題も浮上してくる。即ち、ブラックボックスの問題である。

フローチャートに基づいた記号的 AI については人間による解釈と説明が可能であるのに対し、機械学習 AI 等に関してプログラマーが理解しているのはその全体を動かしている一般的コードのみであり、AI がどのようにして当の行為に至ったのか、各階層でどのような判断を行ったのかは誰にもわからない。このコードを開示したところで何

かを説明したことにはならない。そもそも**ある行為について他者に「説明」することは、その行為が発生するに至った因果の連鎖の全体を提示することではない**のであって、説明を求める人々が求めていることに適切な仕方で応答する、という相互関係論的な次元こそが問題なのである。このことも踏まえると、AI の判断によって誰かに被害が与えられたとき、誰がどのように「説明」することで責任を果たせばよいのだろうか。確かに、人間の価値観にも対応して推論するよう最初からアルゴリズムを設計しておけば、「説明」の際にも、AI の意志決定のプロセスを道徳的観点から解釈して弁解することができる、という考え方もあるが、しかしそもそも価値観に基づいて推論することなど AI に可能なのか、そしてそれをデザインする人間の方でさえ、自らの根本的な価値観について無自覚である部分があるのではないか、という問題が残っている。

それ故、いっそのことブラックボックスを開けて可読性を確保しようとする研究も存在する。その目標が達成されれば、アルゴリズムの改善や人間の知能への還元の道が開かれるであろうが、今のところはまだ不可能なままである。そういうわけで、**AI の責任をめぐっては、その判断の透明性と信頼性が確保されなければならないのであり、具体的にその推進を狙う政策も行われている。**即ち、AI に対する一般的なイメージからターミネーター的悪夢を希釈することや、AI の製造・提供側の説明能力を向上させることが眼目になっている。**しかし、「信頼」というのは人間が機械に対する関係として適切なものなのか**、という問題が存在することには注意しておかなければならない。機械の判断を信頼する社会というのは、資本家やテクノクラートによる支配の拡大に留まらずに、彼らでさえも自分が何をしているのか答えられないようなハイテク社会を実現させる危険性を常に孕んでいる、ということを意識しておく必要がある。

AI をめぐる責任の問題の考察は、具体的で法的な問題の以前に、より一般的な哲学的問題を喚起した。我々はどこまで責任を負えるのか、どこまで AI に依存したいのか、を考えなければならない。

> AI がこれらの理由や説明を**直接に**提供するかどうかに関わらず、人間は以下の問いに答えることが可能であるべきだ。すなわち**「なぜ」という問い**である。**AI 研究者に与えられた課題とは、そもそも AI が意思決定に用いられるとするならば、人間がこの問いに可能な限り答えることができるような仕方で、この[AI という]技術を作り上げることを確実にする、ということなのである。**(104 頁)

第九章　バイアスと人生の意味

AI を利用することに関わる問題は責任の問題だけではなく、AI の判断が実は一定のバイアスを含んでいるという事実もある。例えば、アメリカの司法において採用されている COMPAS システムは、被告の再犯率の予測を行い、これを考慮して裁判官が仮釈放の時期を決定している。しかし実態として、このシステムが再犯率を実際より高く見積もったケースは黒人の被告に偏っており、反対に実際よりも低く見積もったケースは白人の被告に偏っていた。また地域ごとの犯罪発生率の予測を行って警察力配分を助ける PredPol システムは、収入の低い有色人種の地域に強いバイアスを働かせ、そこに監視を偏らせることによって、かえってそのことがその地域の不安を煽って治安を悪化させ、予言を自己成就させてしまう、という批判を受けている。

こうしたAIによるバイアスの再生産は、開発者や利用者の意図に関わらず、あらゆる段階で発生するものである。データセットがそもそもアメリカ白人男性を中心に採取されたものであるのに、その演算結果が多様な民族的背景の男女に適用されることは不適切である。また殺人事件などそもそも母数となるデータが少ないものや、ネット上のテキストなど最初から強いバイアスを含んだコーパスもある。あるいは、研究開発チームの構成員の民族的偏りがAIの設計に影響したり、実際に算出された結果が実は本末転倒な疑似相関である場合もある。例えば、過去のある被告が判決の結果刑務所に送られたという事実と、その親の片方は過去に刑務所に入っていたという事実との間には、統計的に相関関係はあっても、そのことを根拠にして、親の片方が過去に刑務所に入っていたという目の前の被告に対し、刑務所に送る判決を下すことはナンセンスである。

我々は、そもそもAIからバイアスを取り除くことなど可能なのか、ということから問わなければならない。バイアスを抑制することとアルゴリズムの精度を上げることとは、トレードオフの関係にある。そして、そもそも一切のバイアスを含まない不偏不党の公平性こそが本当に目指すべきものなのかどうかも考える必要がある。我々は、AIに**一切の偏りなく現在のありのままの世界を映させるべきなのか、それとも、現に存在している不当なバイアスを相殺するべく、肯定的にバイアスを働かせるアファーマティブアクションをとらせるべきなのか**、民主主義とは多数派の利益を保護することであるのか、少数派の利益を促進することであるのか、といった問題がある。例えば、グーグルの検索アルゴリズムに「数学教授」と入力した場合、アルゴリズムは現状の実態を反映して男性の画像を多く表示するべきなのだろうか、それとも我々の認識を変革するためにも女性の画像を優先的に表示すべきなのだろうか。AIのバイアスに関する問いは、単なる技術の問題ではなく、そもそも我々はどのような世界・正義を望むのか、という問題になってくるのである。

こうしたビジョンの問題に深く関わるのは、労働の問題である。フレイとオズボーン(2013)によれば、やがてアメリカの全労働の47%がAIによって自動化できるとされている。一方で、労働は人間に生きる目的や社会的な絆・帰属意識、健康、責任能力を行使する機会などを与える。他方で、もちろん労働は人生の唯一の価値ではなく、AIによる自動化によって、収入の概念と労働の概念が切り離される可能性も考えられる。ベーシックインカムに基づいた社会経済や、それによってレジャーや価値創造の機会が増した社会を構想することもできる。AIに人間の労働を委託するといっても、その全てを任せる必要はない――そもそもその可能性の現実味は薄い――のであって、考えるべきは、**労働がAIのものになってしまったときに人間がどうするか、ではなく、人間はどの仕事を自らのためにとっておくのがよいか**、である。

ところが、こういったビジョンを語る際にも念頭に置かなければならないのは、歴史上も、こういった技術による労働の軽減というユートピアが語られつつも、それが実現することはなかったということである。というのも、機械は一部の人間の下請けを担うと同時に、他の人間たちの搾取のためにも利用されてきた技術なのであり、社会の階層構造を根本的に変えたことはなかったのである。AIの功罪に関する教養を有している一部の人間がその恩恵にあずかって労働から解放される一方で、AIがもたらす労働や環境的リスクに脆弱なまま晒される人々が残る、という懸念は消えない。

ここまでで、AI倫理は古代ギリシア以来の哲学的リソースや歴史的観点を活用して、問題系を明らかにしてきた。しかしこういった考察は一方で、現代の急激な社会変動のスピードに対応しきれなかったり、実際に運用されるには多様な文化的背景や具体的な政治的プロセスへの視点が不足していたり、といった問題を抱えてもいる。そこで、こ

こからはそういった哲学的理念的な議論から、実際の政治的場面における法的具体的な議論にも目を移してみよう。

第十章　政策提言

実際に政策を考慮する際、様々な角度から問いを立てることができる。①法案をいかに正当化するか？たとえば、人権の原則などが AI に関する新たな法案の根拠となりえる。②いつのために必要な政策なのか？まず、AI に関して言えば、既に社会に浸透し始めている技術が、完全に整備されてしまう前に実際に動き出す必要がある。その上で、新たな政策は 5 年後、10 年後といった近い将来のためのものも必要だし、より長期的な社会のビジョンを示すことも必要である。③新たな法は必要か？政策といっても、新法の制定には限らない。ISO 基準のようなものを定めることもできるし、そもそも現行の法律の解釈を改める範疇内で対応できる、という判断もあり得る。④政策の主体は誰か？新たな政策に関して、具体的に動き出すのは国家のみで済むのだろうか。あるいは、企業や経営者、技術者、市民は、ただ政府が動くのを待つだけではなく、それぞれも行動を取る必要があるかもしれない。⑤そもそも問題をどう定義するか？これまでの章でも見られたように、AI とはどのような点で新たな技術であり、どのような点でこれまでの歴史から連続しているのだろうか。また、我々はこの政策で、AI に関していかなる種類の正義を要請するのだろうか。

さて、では現在、実際に世界で提出されている AI 政策にはどのようなものがあるのかを見てみよう。

国家

国	年	名称・組織	備考
アメリカ	2016	「人工知能の未来への準備」 Preparing for the Future of AI	バイアスの問題など、開発・使用者側の自主的な規制を中心に
中国	2017	新世代人工知能発展規則	プライバシー等の社会倫理リスク最小化必要
イギリス	2018	上院 AI 特別小委員会 House of Lords Select Committee on AI	透明性と説明要求権。産業界の義務（下院）
フランス	2018	「ヴィラニレポート」 Villani report	移民排除や格差拡大の回避が重要
オーストリア	2018	国会諮問評議会 ACARI	人間が責任を負い、説明能力を持つ必要

国際組織

年	名称・組織	備考
2018	データ保護プライバシー・コミッショナー国際会議 IDCPPC	公平性・透明性・差別低減、プライバシー配慮
2018	欧州委員会：人工知能に関する高度専門家グループ AI HLEG	人間の尊厳、幅広い層への説明可能性
2018	科学と新技術の倫理についての欧州グループ EGE	持続可能性、干渉されていることを知る権利
2018	EU：一般データ保護規則	法律強化・罰金規定。AIの情報にアクセスする権利、自分のデータを消去する権利

大学・実業界・NPO

年	名称・組織	備考
2009[3]	デジタル・ヨーロッパ Digital Europe	透明性、解釈可能性。法改正には消極的
2013[4]	「殺人ロボット制止の国際キャンペーン」 Campaign to Stop Killer Robots	
2016	NPO：AIについてのパートナーシップ Partnership on AI	IT・SNS各社。Appleは外的機制の必要を強調
2017	モントリオール大学「責任あるAIのためのモントリオール宣言」 Montreal Declaration for Responsible AI	ほかにもサンタクララ大学「マークラ応用倫理センター」など
2017	Future of Life Institute「アシロマAI原則」 Asilomar AI principles	利益性の維持、人間による制御、価値観の調和

[3] European Broadcasting Union, 《EICTA rebrands itself as 'DigitalEurope'》, https://tech.ebu.ch/news/eicta-rebrands-itself-as-digitaleurope-12mar09. (2021/06/26閲覧)

[4] Campaign to Stop Killer Robots, https://www.stopkillerrobots.org/action-and-achievements/, (2021/06/26閲覧)

2019[5]	米国電気電子学会 IEEE「自律的な知能システムの倫理に関する国際指針」 Global Initiative on Ethics of Autonomous and Intelligent Systems	トレーサビリティの実装、自動運転システムに「免許証」を発行する。 倫理的デザイン

これらの提言の中で、AI の倫理的課題に応答する具体的な方策として特に注目すべきものは、「倫理的に調和したデザイン」、**「デザインが担う倫理 ethics by design」**というアイデアである。これは、AI の設計の段階から倫理性を考慮した仕組みを取り入れることで、一定の責任を果たそうとする試みである。例えば、プロセスを逐一記録する装置などによってトレーサビリティを確保することで、それを基にした説明や制御の可能性につなげることが考えられる。記号的 AI ではどのような計算モデルが構築されたのかを説明すること、統計的 AI ではどのようにデータを集め訓練したのかの方式を説明すること、などが求められることになる。では、この倫理的デザインという方法を実際に実現するとすれば、どのような点に注意する必要があるだろうか。

第十一章　政策立案者にとっての挑戦

改めて強調するならば、デザインが担う倫理とは、既にある技術に対する後からの倫理的対策ではなく、現在進行形で開発されている技術に組み込まれるべき倫理的提言である。そうである限り、一層、それを具体的に実施する現場のプロセスに注意深く分け入らなければならない。多くの政策提言は倫理的原則に基づくトップダウンの規範になる傾向があり、幅広いステークホルダーからの意見聴取が不足しているために、各現場で具体的にどのように実施すればよいのかが曖昧であることがある。この翻訳作業は現状、政策の受け手側に委ねられてしまっている。しかし、民主主義社会であるならば、むしろ**政策立案の時点からボトムアップ式に意見を集め反映させることが、まさに必須の要件であって、「できればした方がよいこと」などではない**のである。第一に、開発者、エンドユーザー、リスクの大部分を負わされる人々、技術に伴う負の産物と共に生活することを強いられる人々から意見を集めることが必要であり、第二に、少数企業にデータプロファイルや権力を集中させてしまわないことが必要である。例えばアメリカにおける環境問題を見てみれば、一部の勢力の政治的圧力によって問題の存在自体が否認されたり、アクションが抑圧されたりといった事態がある。AI の倫理に関しても、行動を起こすにあたってはこのような障害にぶつかることが予想されるのであり、問題を具体的に捉えることを緊急の要件としないことが一体どの勢力に利益をもたらすのか、ということを常に意識しておく必要がある。

ただし、もちろんデザインが担う倫理の実施に含まれる根本的な問題もゼロではない。第一に、デザインに倫理的価値を組み込むためには我々人間がその価値を明示化できなければいけないが、倫理や民主主義を原則の集合や推論によって汲み尽くすことができない以上、それには限界がある。第二に、「倫理的にデザインする」という理念には、「デザインする」ということ、即ち当の技術を開発すること自体は前提されている。場

[5] Springer, https://link.springer.com/chapter/10.1007%2F978-3-030-12524-0_2, (2021/06/26 閲覧)

合によっては、その技術の開発自体が停止されるべきだという批判を行うこともできる余地が残されていなければならない。

いずれにしても、デザインが担う倫理や AI 技術の水準確保のためにも、文理総合的な学際性が必要である。人文学・自然科学、両方の人材を研究開発に参加させるべきであり、若手にもベテランにも積極的な教育が必要である。人文学出身者は新技術が具体的に何をしているのかを、自然科学出身者は技術が持つ倫理的・社会的側面を、さらに理解しなければならない。理想的な状態は、**「AI をやってます」や「データサイエンスをやってます」という言葉の中に、当然のように AI 倫理も含まれているような社会**である。倫理とは、技術的実践にあとから付け加わる周辺部分ではなく、その本質的部分を成しているのである。

そのうえ倫理とは、何かを禁止するという消極的な形をとるとは限らない。善き社会についてのビジョンを与える前向きな倫理が必要なのである。もちろん、「何が価値あることなのか」という問いは、自由民主主義の時代・社会にあって、あくまで個人に委ねられたものになり、それによって一定の戦争が回避され、一定の安定と繁栄がもたらされていることは事実である。従って、確かに「どのような生き方が善き生き方なのか」ということを国家や国際組織が一義的に決定するべきでは決してない。しかし、もしそれに伴って、AI と共にいかなる未来を築くべきか、という問題自体が単に各々の理念の問題として放置され、その結果、公的に議論されずに無視されるのであれば、それはそれで、各々の利権に基づいた AI の利用の抗争を成るに任せる無責任にほかならないであろう。善の理念を提示する権力を少数に委ねることではなく、「テクノクラシーと参加型民主主義との間の程よいバランスをとること」(148 頁)が必要なのである。その際には、非西洋圏における価値観も参照されなければならないであろう。何よりも避けるべき最悪の事態とは、教育格差、無知、無関心が技術的リスクの着実な蓄積をもたらし、倫理・社会・経済・環境を崩壊させていくことである。

> 危険であるのは、知識を伴わずに、(したがって)責任感も伴わずに、権力が行使されることであり、他の者たちがその権力に従属したときに、さらに危険は大きくなる。もし、邪悪というものがこの世に存在するのだとしたら、それは、20 世紀の哲学者ハンナ・アーレントが突き止めた例の場所に住んでいる。つまり、陳腐で日常的な作業や判断における無関心の中に住んでいるのである。(151 頁)

第十二章　気候こそが重要なのだ、愚か者！　私たちの優先度、人新世、イーロン・マスクの宇宙の車

ここまでで、AI の倫理的問題の重要性と、それに対処する具体的な方途を探ってきた。最後に、この「AI の問題」がどれほど急を要するものであるのか、とりわけ、「環境の問題」が第一の緊急の課題として考えられる現代において、我々はどのように優先順位をつければよいのかについても注記しておこう。

まず、AI を問題にするにあたっては、ここまででも見られたように、常に人間を問題にしなければならないにしても、過度の人間中心主義に陥ることは避けなければならない。同時に、AI が地球の環境や人間以外の動物に与える影響も考慮する必要がある。さらに、2015 年以来の SDGs なども考え合わせると、国内・国家間格差、戦争、貧困・栄養失調、水のアクセス、民主主義制度の不足、高齢化、感染症、原子力、子供の人権、

ジェンダー、移民、気候変動、干ばつ、生物多様性など、我々人類が抱える緊急の課題というものは、AI 云々以前にもこれだけあるように思われる。我々は、その実、AI の倫理よりも真っ先に考える問題を抱えているということになるのだろうか？というよりも、**AI の開発が既存の社会問題や環境問題を悪化させる可能性や、反対にそれらの解決に資する可能性もあることを考えれば、これらの課題を考える上では、AI の問題も重要である**と言うべきであろう。しかし、もっと具体的に見てみよう。

国連が環境問題を「現代における最大の課題の一つ」と言うように、例えば現に地球の温度と海面は既に上昇しているのであり、人類は今すぐに動かなければならない。というのも、**環境問題によるリスクを無くすにはもはや手遅れであり、我々にできることは、これを軽減することだけ**だからである。この危機意識を明瞭に示すのが、近年の「人新世」(P. クルッツェン)というアイデアである。いまや人類は地球に対し絶大で決定的な影響力を持っており、我々は今まさに、更新世、完新世に次ぐ、新たな地質年代を刻み形成している、という認識を持つことで、環境問題というものが実に重大で差し迫ったものである、という理解が共有される。こうした環境問題に対して、我々はどのような対策をとればよいだろうか。まず、しばしば持ち出される二つの「解決策」を検討してみよう。しかしそれらは、もっぱら技術的解決のみを考えたものであり、それが孕む深刻な問題がやがて露になるだろう。

第一に、「**持続可能な AI**」を開発し活用することで、環境の保全にとって最適な資源の運用などのソリューションを得ることができるかもしれない。ただし、そのような AI の開発自体が環境に対しどのような影響を与えるかは常に考慮しなければならない。さらに、もしもこの解決法が唯一の方策として執られた場合には、一層深刻な問題が現れてくる。スーパーインテリジェンスは最適な資源配分のために人類の生命の配分を管理する。これが地球工学と十分なだけ結びつけば、地球全体が生命配分のための資源として人類(機械)の設計図の下に計画・利用されることになる。「我々は何をすべきか」について、人知を超えた AI が親切にも人間の代りに判断し、我々の直観に左右されることなく、全てを最も効率的な資源運用のために動員する。これは、確かに人新世が喚起した問題を「解決」することではあろうが、それは、AI の判断に依存したテクノクラシーを完成させることによって、倫理的問題意識を我々人間が持つ必要のないものとしてしまって、「そもそも地球資源を利用することなど問題ではない」という仕方で問題を消滅させることでしかない。そしてやがて自律性を失った人類の姿は、もはや AI に奉仕される神、「ホモ・デウス」(ハラリ)として思い描くことなどできず、ただ AI の計算過程の中で自由に処分できない邪魔な遺物としか言えないだろう。

第二に、いっそのこと地球から脱出してしまう、というシナリオがある。2018 年、イーロン・マスクは宇宙空間に漂うスポーツカー「テスラ」の映像をビジョンとして示した。彼は火星の植民地化計画を構想しているのである。確かに、宇宙へ人類が進出すること自体は、サバイバビリティの観点から地球上の問題を研究する上でも大きな進展に繋がるだろう。しかし、もし反対にそれが地球上の問題を無視することのために進められるなら、それは見過ごすわけにはいかない。アーレントが『人間の条件』で述べたように、数学と技術の過度の純化によって、人間は物事を抽象化することで自らが本来持っている身体的・自然的条件に属する問題や責任を端的に無視していく傾向に飲まれる恐れがあるのだ。つまり、人間の倫理や政治といったものは常に地上の現場と言論の中で行われなければならないし、その外では無意味になるのだが、科学者が取り組み提示する世界は、まさに言論のない非人間的な世界そのものなのである。科学技術が可能にする道を唯一の解決と見なすことは、実は我々の倫理的責任を全く無視して現実逃避す

ることの言い換えでしかない。人類がその住処としていた地球から脱出してしまうことは、地球上の他の生物や環境を、人類が散々に改変した状態のままに放置して立ち去ることを意味する。これは人類の倫理を端的に無視した選択ではないだろうか。

こういったトランスヒューマニスト的・トランスアーシスト的空想を急速に育んでしまう危険性を孕んでいるのが、AI 技術なのである。AI 技術は我々の生活や問題解決能力の大きな助けとなる一方で、こうした特定の逃避的言説とまぜこぜにされて新たな問題を生む可能性があるのだ。実際、地球からの脱出の構想は、人間内部に分断を生じさせる。権力者や富裕層など地球外へ移ることが容易な層と、そうでない層とが存在するのであり、こういった言説を流布することで利益を得るのは常に前者なのである。最終的には、前者が地球から脱出する一方で、後者が汚染された地球上に取り残されるということさえ考えられる。そうでなくとも少なくとも、宇宙開発のために投じられる資金は、地球上の他の問題には割り当てられないことになった資金である、ということは確かである。このようなシナリオは何も誇張したものではなく、デリーや北京における大気汚染と空気清浄技術を見てみればわかるように、すでに一般に、**技術というものは裕福な人々ばかりの生存の道具となる傾向にある**のだ。

このような問題がある以上は、AI に全ての判断を委ねたり、いっそ地球から脱出したりといった仕方で環境問題から目を反らすことはできない。具体的に地球上で AI が環境問題にどのような仕方で資するのかを考えなければならない。AI 技術がもたらす利点は、スマートグリッドやスマート農業などの実現によって環境対策を飛躍的に推進させることができるかもしれない、ということであり、反対に問題点は、AI 技術の開発に伴う廃棄物等が環境問題を悪化させるかもしれない、ということである。さらに言えば、AI の判断が環境問題の解決に最適なものになったとしても、あるいはそうなってしまったときにこそ、AI の助言によって我々の日々の行動がその最適な方向へ、誰もそれと気づかずに規律させられることは、人間の自由と自律性を失わせる新たなパターナリズムに繋がるであろうし、あるいは全てがデータ処理の下にあるという世界観・人間観を支配的なものとしてしまうことにも繋がるであろうことは常に念頭に置かなければならない。

人類が抱える問題への対策を講じるにあたって**避けるべきなのは、技術的解決至上主義**なのである。問題解決のためには AI を使用すればよい、という発想は、最終的な解決が存在すること、そして技術だけがそれを達成できるということ、こういった前提を含んでいる。**間違いなく、技術や数学のみで問題が解決できるということはない。そこには必ず政治的で社会的な問題が存在しているのであって、その限り、答えは常に人間の手に委ねられなければならないだろう。そして人文学や社会科学に目を向けることは、そもそも「最終的解決」という発想自体に危険が含まれている可能性も教えてくれる。**我々は、実際に人類の未来を拓くために、学際的な視点を持って行動しなければならないのである。AI が得意とする抽象的知能と、人間が併せ持つ実践的知恵と、この 2 つを上手く統合する方向に道を探るときが来ている。

読書案内

最後に、翻訳書巻末に添えられている「日本の読者のための読書案内」(168-170 頁)をまとめておきます。本稿や本書をきっかけにして AI 倫理の知見を深めるための助けとなれば幸いです。

AI の現実的な危険性について

・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.

・山本龍彦『おそろしいビッグデータ　超類型化 AI 社会のリスク』朝日新書，2017.

・平和博『悪の AI 論　あならはここまで支配されている』朝日新書，2019.

AI の倫理について

・西垣通，河島茂生『AI 倫理　人工知能は「責任」をとれるのか』中公新書ラクレ，2019.

倫理学的な背景を知るために

・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.

・ウェンデル・ウォラック、コリン・アレン『ロボットに倫理を教える　モラル・マシーン』岡本慎平.久木田水生訳，名古屋大学出版会，2019.

AI と社会、法について

・江間有沙『AI 社会の歩き方　人口知能とどう付き合うか』DOJIN 選書，2019.

・稲葉振一郎ほか編『人工知能と人間・社会』勁草書房，2020.

・宇佐美誠『AI で変わる法と社会　近未来を深く考えるため』岩波書店，2020.

技術哲学について

・ラングドン・ウィナー『鯨と原子炉　技術の限界を求めて』吉岡斉，若松征男訳，紀伊国屋書店，2000.

・ピーター=ポール・フェルベーク『技術の道徳化　事物の道徳性を理解し設計する』鈴木俊洋訳，法政大学出版局，2015.

・久保明教『機械カニバリズム　人間なき後の人間学へ』講談社選書メチエ，2018.

機械学習について

・山本一成『人工知能はどのようにして「名人」を超えたのか？最強の将棋 AI ポナンザの開発者が教える機械学習・深層学習・強化学習の本質』ダイヤモンド社，2017.