人工知能とバイアス

久木田水生

ITEC セミナー
2019 年 7 月 17 日

概要

ビッグデータや機械学習の発展は、これまでは機械に遂行させるのが困難あるいは不可能だった様々なタスクを自動化することに成功してきた。その中には人間の能力や適性を評価する、労働を管理するといったタスクもある。このようなタスクを遂行するシステム、アプリケーションは人間以上に正確、客観的、公平であると宣伝されるのが常である。しかしそのように宣伝される正確さ、客観性、公平さはしばしば幻想にすぎず、実際にはバイアスを持ったアンフェアなものである。本発表ではいかにして人工知能の判断にバイアスが入り込むか、それが社会にどのような影響を与えているか（与えうるか）を論じる。

1　はじめに

人工知能は現在、最も発展の著しいテクノロジーの一つであり、経済と産業の起爆剤になることが期待されている。その一方で、それは社会の構造、人間の生活、人間同士の関係、さらには私たちの思考さえも大きく変化させるポテンシャルを持っている。そしてその変化は、必ずしも明るいものばかりではない。人工知能には様々な懸念が付きまとっており、それゆえに現在、人工知能の危険性について、あるいは人工知能に関する倫理的指針や法による規制について盛んに議論されている（cf. 村上 [9]；弥永・宍戸 [13]；江間 [5]；平 [11]；オニール [10]）。本講演では特に、ビッグデータに基づいて学習を行い、人間を評価する人工知能の問題を取り上げる。

ビッグデータや機械学習の発展は、これまでは機械に遂行させるのが困難あるいは不可能だった様々なタスクを自動化することに成功してきた。その中には人間の能力や適性を評価する、労働を管理するといったタスクもある。このようなタスクを遂行するシステム、アプリケーションは人間以上に正確、客観的、公平であると宣伝されるのが常である。しかしそのように宣伝される正確さ、客観性、公平さはしばしば幻想にすぎず、実際にはバイアスを持ったアンフェアなものである。本発表ではいかにして人工知能の判断にバイアスが入り込むか、それが社会にどのような影響を与えているか（与えうるか）を論じる。

2　人工知能は人間を対象にしたリスク分析のための革新的なツールである

「人工知能」という言葉は、多様なテクノロジーに対する曖昧な総称である。本稿では近年発展が著しい、ビッグデータに基づく機械学習システムに限定して論じる。以下では断りがなければ「人工知能」という言葉によってそのようなシステムを指すものとする。

人工知能がやっていることは、与えられた大量のデータから共通するパターンを見つけ出し、そのパターンに基づいて（確率的な）分類や予測を行うことである。例えば大量の猫画像から猫に共通するパターンを抽出

し、新しい画像についてそれが猫であるかどうかを識別できるようになる。

これだけだとそのインパクトは分かりづらいが、人工知能について論じる際、それを私たちが置かれている情報環境と切り離して考えることは意味がない。パーソナルコンピュータ、インターネット、検索エンジン、スマートフォン、その上で働く様々なアプリ、動画配信サービス、ネットショッピング、ソーシャルメディア、監視カメラ、スマートスピーカー、などなど。今日、私たちのあらゆるオンラインの行動、そしてますます多くのオフラインの行動のデータが様々なインターフェースを通じて収集・保存されている。人工知能はこの膨大なデータ（ビッグデータ）に基づいて、人々を分類し、人々の振る舞いや属性、性格や能力を予測、推測する。

科学はこの世界に起こる出来事を予測し制御することを主たる目的としている。しかし従来の科学的手法では人間や社会のような複雑なシステムについては、その振る舞いの予測も制御も非常に困難であった。しかし人々の属性や行動についての多種多様かつ膨大なデータが手に入るようになり、そしてそのデータからパターンを見つけ出す人工知能が登場して、状況は一変した。人間の振る舞いはもはや予測・制御不可能なものではなくなった。「これこれの属性を持つ人はかくかくの確率でしかじかの行動をとる」といった形の予測が可能になった。そしてデータが集まれば集まるほど予測できることが増え、また予測の精度も増していく。

ビジネスにおいては人々の行動を正確に予測できることは大きなアドバンテージである。いつ、どこで、どういった需要が、どれくらいの確率で発生するかを正確に知ることができれば、企業はその分だけ無駄なく効果的なアクションを起こすことができ、より多くの利益を上げることができる。さらには適切なタイミングで適切な情報を与えれば、人々の行動を誘導することもできる。何億というユーザーを抱える企業であれば、予測の精度がほんのわずかでも上がれば、それによって得られる利益も莫大なものになる。だからこそ Amazon、Google、Facebook などの巨大 IT 企業は貪欲にデータを収集し、そして人工知能の開発に巨額の投資を行っている。人工知能はこういった大金持ちが超大金持ちになることに、現在最も有効に活用されている。

しかし人工知能の活用に関心を持つのは、そういった巨大 IT 企業ばかりではない。人工知能の開発者は自らの利益関心に沿って、他者について人工知能に様々な推測を行わせることができる。典型的な例を挙げれば、労働者としての能力や適正、犯罪を犯す可能性、ローンを返済する能力、特定の病気に罹る（あるいは罹っている）可能性、交通事故を起こす可能性、配偶者としての相性、ある商品にどれくらいの金額を支払いそうか、ある政党を支持しているかどうか、等々を確率的に推測するために人工知能が使われている。変わったところでは、東京大学で開発されている女性の顔の「魅力」（その顔がどれだけ好まれるか）を推測して数値化する人工知能[*1]、スタンフォード大学で開発されている人々の性的指向を推定する人工知能[*2]などというものもある。

人工知能の開発者たち、あるいは利用者たちはなぜこういったことに関心を持つのか。純粋な好奇心によって動機づけられている場合もあるかもしれないが、多くの場合、理由は人工知能が与えてくれる情報が彼らの利益と損害に直結しているからである。ある商品の需要を知ることは生産者・販売者にとって極めて有益である。同様に、雇用者にとって有能な社員になりそうな人間を知ること、警察にとって潜在的な犯罪者を特定すること、ローン会社にとって誰が破産しそうかを知ること、保険会社にとって病気に罹りそうな人間を知ることは、極めて有益である。

[*1] 「AI で女性の顔の“魅力”も数値化――東大で研究中の「魅力工学」とは？」、『IT メディアエンタープライズ』、2018 年 06 月 06 日。https://www.itmedia.co.jp/enterprise/articles/1806/06/news006.html

[*2] 「スタンフォード大、顔画像からその人の「性的指向（異性愛、同性愛など）」を推測する機械学習を用いた手法を論文にて発表。精度は男性 91 %、女性 83 %で区別」、『Seamless』、2017 年 9 月 8 日。https://shiropen.com/2017/09/08/28032

意思決定を行う際に、ありうる選択肢のそれぞれに伴う潜在的な利益と損害を見積もることを、工学や経済学の分野では「確率論的リスク分析」あるいは単に「リスク分析」と呼ぶ。リスク分析においては、ある選択肢をとったときに、どのような事象がどれくらいの確率で生起するかを見積もることが必要になる。しかし従来は人間や社会のような複雑なシステムについて、特定の振る舞いの生起確率を正確に予想することは難しかった。それゆえに人間に関するリスク分析はしばしば多くの仮定に基づいた不正確なものにならざるを得なかった。しかしビッグデータと人工知能は人間の振る舞いについての従来よりはるかに正確な確率的予測を可能にした。人工知能は、何よりもまず、人間や社会を対象にした確率論的リスク分析のための革新的なツールなのである。

3 「人工知能だから公平です」という人間を信用してはいけない

人工知能を開発・販売する側は、人工知能が人間と異なり、バイアスがなく、公平で、正確である、と宣伝するのが普通である。確かに人工知能は、人間と異なり、その時々で判断が変わるということはなく、基本的には同じ入力に対しては同じ結論を出力する[*3]。しかしだからと言って人工知能にバイアスがなく公平であり、正確だというのは大きな間違いである。「人工知能だから公平だ」、「人工知能だから正確だ」と言っている人間がいたら、おそらくペテン師であるか、人工知能のことを何も知らずペテン師に騙されているかのどちらかだろう。いずれにせよそんな人間の言うことは信じるに値しない。

では人工知能はどのようにして不公平な判断を下すのだろうか。バイアスの源泉は主に二つある。一つは判断のためのアルゴリズム、もう一つは学習のもとになるデータである。ざっくり言うと、アルゴリズムには作成者の偏見や先入観や意見が、データには社会の持つそれらが反映される。

まずアルゴリズムについて。人工知能は予測したい属性を、その他の属性から予測する。統計学の用語では前者を「目的変数」、後者を「説明変数」と呼ぶ。例えば、顧客の将来の需要を予測するのに、購買履歴と住所を使うとする。このとき将来の需要が目的変数で、購買履歴と住所が説明変数である。説明変数として何を使うかを選択するのは、予測システムを作る人間である。そしてその選択に人間のバイアスが反映されるのである。

ネットショッピングにおける購買行動であれば、予測に使った説明変数が本当に妥当なものであったかを検証して、システムの改善のためのフィードバックを得ることができる。しかし例えば人事採用のシステムであればそうはいかない可能性がある。例えば採用システムを作っている人間が、特定の大学を出ていることをポジティブあるいはネガティブに評価するようにアルゴリズムを設計したとしよう。それによって不採用になった人間が、本当に社員としての能力に問題があったかどうかは確かめようがない。

アルゴリズムのみではなく、人工知能が学習するのに用いられるデータにもバイアスがかかりうる。例えば過去の犯罪のデータは、その社会が特定の人種に対して差別的である場合は、公平なデータではない。なぜなら社会がその人種の人間を犯罪予備軍と見なして、ほかの人種の人間よりも厳しく監視していたために、同じ犯罪をしてもその人種の人間は捕まり、ほかの人間は見逃されるということの結果かもしれないからである。このようなデータにに基づいて予測を行い、さらにその人種の人間を厳しく監視した結果、ますますその人種の人間の検挙率が高まるということになる。実際に、アメリカで再犯の可能性などについて推測するために使われ、量刑の決定にも影響を与えている COMPAS というシステムでは黒人の再犯率を白人の再犯率よりも高く見積もると指摘されている（平 [12]）。さらに COMPAS は、素人によるあてずっぽうの推測にも劣る程度

[*3] もちろん学習の過程では判断の基準は変化する。

の精度しかなかったことも報告している（Young [14]）。

HireVue はデータとアルゴリズムの両方にバイアスがかかっているアプリケーションの分かりやすい例である。HireVue という人事採用サポートシステムは、面接の様子を録画して、面接を受けに来た人間の語彙やジェスチャーや態度を評価し、その会社で優秀とされている社員の語彙やジェスチャーと比較する。そして優秀な社員たちとどれだけ近いかという尺度でランク付けをする。しかし身振りや使う言葉にはその人間の社会的階級やジェンダーや民族性などが強く反映される。たとえばある会社で女性や有色人種が不当に低く評価されていれば、このシステムはそのバイアスを踏襲して、女性や有色人種の候補者を低く評価だろう。ジェンダーや人種を理由として人の評価を下げることはアンフェアであり、場合によって違法でもある。しかしこのようなシステムは、表立ってはそのような差別をしていないように見せかけて、裏口から差別を導入することを可能にする。Arvind Narayanan は、Twitter で「@hirevue によるこのアプリは，社会的バイアスを定着させるために作られているとしか思えない AI の例だ」[*4]と述べている。

AC Global Risk 社が開発している「遠隔リスク評価（Remote Risk Assessment; RRA）」というシステムには人工知能による人間評価の様々な問題点が顕著に現れている（Kofman [7]）。このシステムは電話越しの 10 分程度の会話（決められた質問に対して母国語で答える）のみによって危険人物であるどうかを判定するという。「移民を徹底的に精査せよ」というトランプ大統領の要求に対する答えとして、AC Global Risk 社は RRA を「アメリカや他の国が現在、直面している monumental refugee crisis に対する究極のソリューション」と喧伝している。AC Global Risk 社はソフトウェアの詳細に関する質問に答えることを拒否しているが、公開されている材料をレビューした専門家はこれを「たわごと bullshit」、「いかさま bogus」と呼んでいる。音声感情認識の権威である Björn Schuller は The Intercept の取材に対して「何らかの精度で声だけから嘘を見破れるという印象を与えることは倫理的に問題がある。そんなことができると宣伝する人間がいたら、彼らこそがリスクだとみなされるべきだ」と述べている。アメリカの入国審査では、人を捜査したり入国拒否したりする際、喋り方や見た目が口実として使われている。RRA はそのようなバイアスを「ルーティンとして蔓延させ、一見「客観的」に見せるかもしれない」と専門家は危惧している。

仮に人間の持つ既存のバイアスのようなものを反映していない人工知能だったら公平なのだろうか。必ずしもそうとは言えない。それは現在の主流の人工知能は、統計的な推論を行っていることの必然的な帰結である。人間評価アルゴリズムはデータに基づいて人々を分類、クラスタリングする。そしてそのクラスターに属する人間に対してラベリングを行う。例えばあるクラスタに属する人間の 8 割が暴力的な傾向を持つと判断される。そうするとそのクラスタに属する各々の人間が 80 %の確率で暴力的だと推論される（図**??**）。このような判断が体系的に適用されれば必ず一定の割合の人間は、実際には暴力的でないにもかかわらず、暴力的である確率が高いクラスタに属しているという理由で排除されるのである。

もしこのような判断の理由が人種や性別や出身地である場合には、それは差別として非難され、場合によっては法律によって禁止されている。しかし人工知能が新たに生み出すラベリングによる差別は現在のところ禁じられていない。このような推論に基づいて、リスクが高いと判断された人間は例えば企業で採用を見送られる、長い刑期を課される、入国を拒否される等々、不利な状況に置かれることになる。

しかし人工知能による人間の評価は現在社会のあちこちで使われている。それらは宣伝されるほど正確でもないし、バイアスを免れているわけでもない。そしてそれはしばしばすでに困難な状況に置かれている社会的弱者、貧しい人々、差別されている人々により不利な判断をする。アマゾン社が秘密裏に開発していた人工知能による人材評価システムは女性を低く評価するバイアスを持っていたことが判明し、開発チームがその問

[*4] https://twitter.com/random_walker/status/901851127624458240

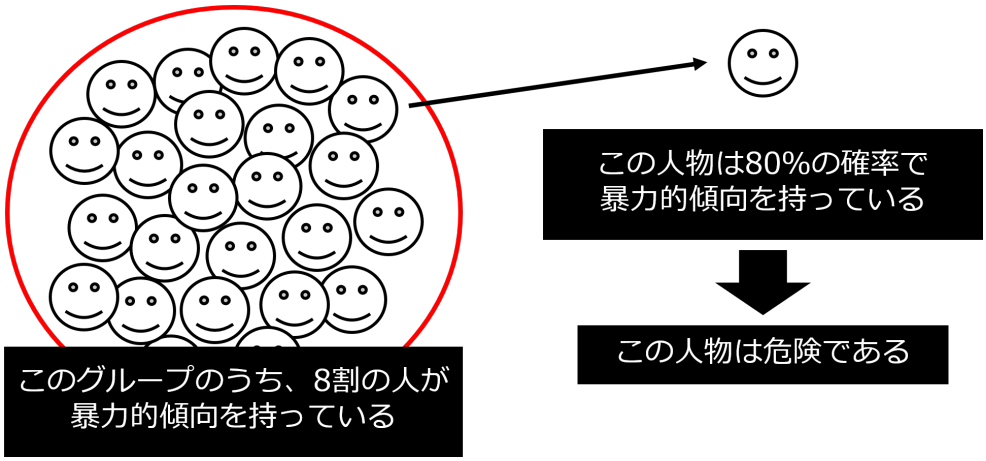

図 1　アンフェアな推論

題に対処することができなかったために破棄された（Dastin [4]）。Google の検索エンジンは、ユーザーが女性の場合、男性よりも収入の低い職業の求人広告を出すというバイアスを持っていた（Borgesius [1]）。インターネットショッピングサイトは、都市に住んでいるカスタマーよりも、実店舗で買う選択肢を持たない田舎のカスタマーに対してより高い価格を吹っ掛ける（*ibid.*）。こういった問題が次々に明らかになっているにも関わらず、企業や政府はアルゴリズムで人を評価することに熱心である。なぜならそれは複雑で難しい問題に対する、単純で効率的な「解決」であるからだ。

4　人工知能のメッセージ

メディアとは情報を伝える媒体であり、私たちが直接に経験していない現象や対象について、何らかの認識を持つことを可能にする手段である。そしてこの意味において、人工知能は文字通りメディアである。ある種の技術哲学においては、テクノロジーはすべて使用者と環境の間の仲介をするものであり、それゆえにメディアだと考えられる。しかし人工知能はあらゆるテクノロジーの中でも、もっともメディアらしいテクノロジーの一つだと言えよう。そしてマーシャル・マクルーハン [8] がかつて主張したように、メディアそれ自体が「メッセージ」であるとするならば、人工知能それ自体もまた何らかのメッセージであるに違いない。それでは人工知能とはいかなるメッセージなのだろうか。

伝達される情報の形式や様態（モダリティ）、あるいは私たちがその情報を消費する仕方は各々のメディアの持つ技術的特性によって規定される。例えば新聞ならば伝えられる情報は文字だけ、ラジオならば音声だけである。新聞は私たちが情報を消費する間、そこに注意を集中することを要求するのに対して、ラジオは私たちが他の仕事をしながら（運転しながら、本を読みながら、食事をしながら、ジョギングをしながら）聞き流すような消費の仕方を可能にする。また新聞の情報は保存がきくのに対してラジオはそうではない、音声や映像は文字に比べて感情を喚起しやすいなどなどの違いがある。これが「メディアはメッセージである」といわれるゆえんである。

では人工知能はどのような形式・様態で情報を伝え、私たちがそれをどのように消費することを促しているのだろうか。新聞やラジオ、テレビのような伝統的なメディアと人工知能が大きく異なるのは、前者が基本的には事実と確認されたことを伝える（誤報の可能性はあるが）のに対して、人工知能は一般にデータに基づい

た確率的な推測を伝えるという点である。もちろん伝統的なメディアも全く推測をしないわけではない。新聞には毎日天気予報が掲載されるし、政治や経済の先行きや、ペナントレースの行方について専門家の予想が掲載されることもある。しかし人工知能が伝える情報は、すべて推測、しかもそれは個々の私人（あるいは私人の集団）の行動や性格や能力や思考についての確率的な推測なのである。この点が人工知能というメディアの際立った特徴である。ではこの特徴から私たちはどのようなメッセージを読み取ることができるだろうか。

上述したように、人工知能は様々なデバイスから取得された機械可読なデータから、人間の性格や能力、適正、危険性などなどを確率的に予想、推測する。その予測は、確率的リスク分析のために利用される。リスクが高いと判断された人間は、しばしばその判断が正しいかどうかを検証されることもなく、排除される。ビッグデータと人工知能によって提供された情報は、その情報を利用する人間をこのような行動に導くのである。要するに人工知能は、他者をウェブ上に残されたデータの集積、そしてそこから推測される種々の属性の束として扱い、特定の利益関心に沿ったリスク分析を行った結果として、「この人はこれだけの利益／損害をもたらす見込みがある」という情報を伝えるメディアなのである。

人工知能が様々な局面で応用されるにつれて、人間をリスクとして扱い、利益／損害という観点から評価する慣行、そして不利益をもたらすと判断された個人を排除する風潮が広がるだろう。そのような判断は間違っていることもある。それは人工知能の判断が確率的であることからの不可避の帰結である。しかし例えば一部の人たちを誤って切り捨てたとしても、全体として利益が向上するならば、人々は人工知能の判断を採用するだろう。

企業がそのような判断をすることはある程度仕方がない。企業と金銭的利益の最大化を第一の目的とするものだからである。リスク分析は基本的に政策決定者や経営者が大局的な観点から効率的なマネジメントを行うための道具である。そこで重要なのは統計上の数字であり、全体として利益が上がるならば切り捨てられる個人は顧みられない。

しかし個人間の人間関係には、数値化できて機械で測れる利益とは異なる価値がある。例えば私にとって家族は、まさに彼らであるということに価値があるのであり、その価値を数値化したり他人と比較したりすることは意味をなさない。人工知能による人間の評価が、あらゆる人間関係の成立以前に介入してくることは、このような個人的に特別な関係が築かれる可能性を阻害する。また人の性格や能力は決して固定された、客観的に測れるようなものではなく、特定の人間関係の中で発達していく可能性を持ったものである。人間は信頼されれば、その信頼に応えるようと努力する。つまり信頼がきっかけになって、その人が本当に信頼に足る人間になるということがある。単純に頻繁に顔を合わせるだけで好意が生まれることもある[*5]。そしてそうやって生まれた信頼や好意に基づいた協力関係がお互いに利益をもたらし、そのことがさらなる信頼と好意を促進させる。だがこのような良い人間関係を促進するサイクル、ポジティブなフィードバックループは、ウェブ上で入手できるデータに基づいた自動評価システムからは始まらないだろう。

一言でまとめると、メディアとしての人工知能が持つメッセージは、次のようなものである。第一に、人間は様々なデバイス、アプリケーションを通じて取得されるデータと、そこから確率的に推測される属性の集まりとして理解できる。第二に、他者はあなたにとってリスクであり、そのリスクは前もって見積り、回避することができる。人工知能が社会に浸透するということは、このようなメッセージに私たちが知らず知らず曝され続けるということである。そのことが人々の人間観や、成立しうる人間関係に与える影響について、注意しておかなければならない。

[*5] 心理学ではこれを「単純接触効果」と呼ぶ。

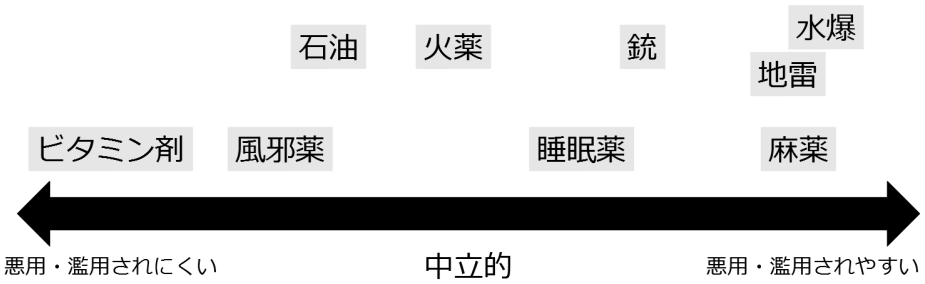

図 2　テクノロジーの持つバイアス

5　やる、なぜならできるから

以上で私たちは人工知能の判断にどのようにしてバイアスが入り込むか、そしてそれがどのような影響をもつかということを考察した。本節では、以上で述べたのとは異なる意味での「バイアス」について論じたいと思う。

テクノロジーについてはしばしば「テクノロジーはあくまでも人間が何らかの目的のために利用する道具に過ぎない。テクノロジー自体は価値中立的であり、それを善用するのも悪用するのも、それを使う人間次第である」ということが主張される。これをテクノロジーについての中立論と呼ぶことにしよう。

中立論は非常に自明な意味では正しい。あるテクノロジーを利用する人間がいなければ、それは人間に何の影響も与えない。しかしながら、個々のテクノロジーは悪用・濫用されやすさ、あるいは悪用されたときの影響の程度において異なっている。例えばマッチやナイフも悪用されうるが、その悪用されやすさは比較的軽微である。それに対して銃や麻薬はより悪用・濫用されやすい。地雷や水爆となると、善用されるユースケースを考えるのは難しい（図 2）。このような個々のテクノロジーが持つ悪用・濫用のされやすさを、テクノロジーの持つ「バイアス」と呼ぶことにする。本節ではこのような意味での人工知能の持つバイアスについて考えたい。

上述したように、人々が人工知能を使う動機は主に、人工知能の与えてくれる情報が彼らの利益と損害に直結しているからである。しかしそうとばかりも言えないようなケースもある。例えば、最近、ドイツの匿名のプログラマーが顔認証技術を用いて、ソーシャルメディアのプロフィール画像とポルノ動画サイトにアップロードされている動画を比較して、過去にポルノに出演したこのある女性を特定したということで大騒ぎになった（Chen [2]）。その後、このプログラマーはこのプロジェクトとすべてのデータを破棄した、と言っている。このようなデータの利用はヨーロッパでは違法である。EU 一般データ保護規則（GDPR）は、EU に居住する人間のデータを、本人の同意なしに他人が収集することを禁じている（そしてこれはデータの利用者が EU 圏外の人間であっても適用される）。

そもそも、なぜこのプログラマーはこのようなことをやろうと考えたのだろう。インターネットには、過去にポルノに出演したことのある女性を特定して、その個人情報を曝す人々がいる。彼らは単に面白半分の嫌がらせで、他人の人生を踏みにじる。インターネットにはそのような事例が溢れている。Cocking and van den Hoven [3] はオンラインでの人々の様々な非倫理的な行為を報告している。その多くが、明確な動機や利益関心に基づいたものではなく、単に「できるから」という理由で遂行される、と彼らは指摘する。オンラインで

非倫理的な行動をとる人間は、特別に「邪悪な」人間、あるいは他者への共感を全く欠いたサイコパスというわけではない。Cocking and van den Hoven [3] は、オンラインの環境では、通常ならば機能する人間の道徳性が発揮されないと主張する。彼らはこれを「道徳の霧 moral fog」と表現する。人間の道徳性は人間が進化してきた環境に適応して発達してきた。しかしそれは現代の私たちが置かれている情報環境において人々を倫理的な行動に導くようにはできていない。

人工知能が道徳の霧をさらに厚くする蓋然性はかなり高いと思う。前節で述べたように、人工知能は他者とウェブ上で収集されるデータの集積と見なすように私たちを促す。このことの倫理的な含意は重要である。なぜならば他者を単なるデータと見なすことは一般に、他者と自分の間の心理的距離を広げ、そしてそのことは他者に対する共感や道徳的配慮を減少させる効果を持つからである。ポルノ出演者特定アプリはもちろん、顔から人の性的志向を推測する人工知能や、女性の顔の魅力を数値化する人工知能についても、人々がそれを善用することはほとんど期待できない。女性の写真を入力すると、その女性の裸の画像を作成するアプリ、「DeepNude」の作成者は、公開して数日後にこのアプリをシャットダウンした。作成者によれば、彼らが予想していた以上にダウンロードされ、「50 万人もの人がそれを使うとなれば、人々がそれを悪用する可能性があまりにも高い」と判断したということである。「世界はまだ DeepNude への準備ができていない」と作成者は述べている[*6]。賢明な判断であるが、すでにこのアプリと類似したアプリが次々につくられている。

GDPR のようなデータ規制は人工知能によって引き起こされる問題に対する有力な対抗手段である。反差別法や憲法も重要な武器になるだろう。しかし個別的な領域については、それぞれに独自の規制が必要だろう。冒頭で述べたように、近年は人工知能の倫理についての議論が盛んであるが、正直言って「倫理指針」といったものでは実効的な対策にはならないと思う。ひょっとすると世界はまだ人工知能に対する準備ができていないのかもしれない。

6 終わりに

以下のことを繰り返して強調しておきたい。

- 人工知能は、人間を対象にした確率論的リスク分析のための革新的なツールである。
- 人工知能は現在、大金持ちを超大金持ちにすることに最も有効に活用されている。
- 一方、人工知能はしばしば社会的弱者をさらに不利な状況に陥れる。
- アルゴリズムには作成者の偏見や先入観や意見が、データには社会の持つそれらが反映される。
- 「人工知能だから公平」はペテン師の言うこと。
- 人工知能は、他者を機械可読なデータと、そこから推論される属性の束とみなし、リスク分析の対象として扱う態度を助長する。
- おそらく多くの人間は人工知能を悪用する誘惑に勝てない。
- 「倫理」だけでなく、しっかりした規制も必要。
- 世界はまだ人工知能に対する準備ができていないのかもしれない。

[*6] https://twitter.com/deepnudeapp/status/1144307316231200768

参考文献

[1] Frederik Zuiderveen Borgesius, “Discrimination, artificial intelligence, and algorithmic decision-making”, The Directorate General of Democracy, Council of Europe, 2019. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73

[2] Angela Chen, “The guy who made a tool to track women in porn videos is sorry”, *MIT Technology Review*, May 31, 2019. https://www.technologyreview.com/s/613607/facial-recognition-porn-database-privacy-gdpr-data-collection-policy/

[3] Dean Cocking and Jeroem van den Hoven, *Evil Online*, Wiley-Blackwell, 2018.

[4] J. Dastin, “Amazon scraps secret AI recruiting tool that showed bias against women”, *Reuters*, October 10, 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

[5] 江間有沙、『AI 社会の作り方——人工知能とどう付き合うか 』、化学同人、2019 年。

[6] Jacob Kastrenakes, “Controversial deepfake app DeepNude shuts down hours after being exposed”, *The Verge*, Jan. 27, 2019. https://www.theverge.com/2019/6/27/18761496/deepnude-shuts-down-deepfake-nude-ai-app-women

[7] A. Kofman, “Dangerous junk science of vocal risk assessment”, *The Intercept*, November 25, 2018. https://theintercept.com/2018/11/25/voice-risk-analysis-ac-global/

[8] マーシャル・マクルーハン、『メディア論——人間の拡張の諸相』、栗原裕、河本仲聖訳、みすず書房、1987 年。

[9] 村上祐子、「人工知能の倫理の現在——研究開発における技術哲学・倫理の意義」、*IEICE Fundamentals Review*, 11(3), pp. 155-163, 2018 年。

[10] 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。

[11] 平和博、『悪の AI 論』、朝日新聞社、2019 年。

[12] 平和博、「見えないアルゴリズム :「再犯予測プログラム」が判決を左右する」、2017 年 8 月 8 日。https://www.huffingtonpost.jp/kazuhiro-taira/clime_b_11381184.html

[13] 弥永真生・宍戸常寿編、『ロボット・AI と法』、有斐閣、2018 年。

[14] Ed Young, “A popular algorithm is no better at predicting crimes than random peopl”, *The Atlantic*, January 17, 2018. https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/