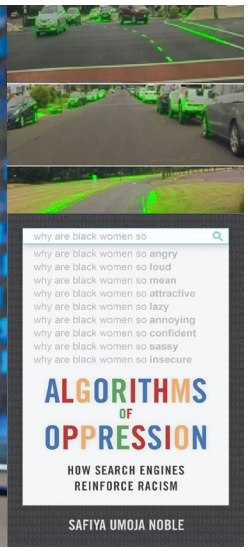
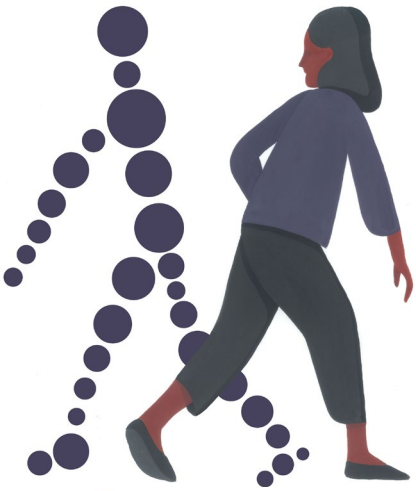
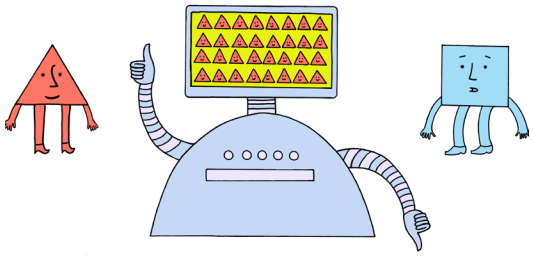
MIT Technology Review
Published by KADOKAWA / ASCII

# AI and bias

## 人工知能は公平か？





人工知能（AI）が社会のさまざまな場面で使われるようになり、AI が導き出す判断や結果が人の人生を左右したり、生命のカギを握ったりすることも多くなってきた。だが、機械学習によって訓練された AI はどのようにして結果を導き出しているのかはわからず、データに潜む意図しないバイアスの影響を心配する声も上がっている。

AI からバイアスを排除し、説明可能な AI を実現することは可能なのだろうか。MIT テクノロジーレビューが 2017 年 4 月～ 2018 年 4 月にかけて掲載した記事から、AI と倫理、特にバイアス問題に関する取り組みを中心に紹介する。

# CONTENTS

# 人類に残された、AIを信用しない、使わない、という選択肢

by Will Knight

Photo by Adam Ferriss; Frederic Lewis | Getty

医療や裁判、軍事作戦など、取り返しのつかない場面でAIを使う可能性が現実化している。
しかし今なら、なぜそう判断するのか本質的に説明できないAIを、
使わない、と判断する選択肢が人類には残されている。

昨年、ニュージャージー州マンモスカントリー市の閑静な通りに奇妙な自動運転車が現れた。半導体チップメーカー、エヌビディア（Nvidia）の研究者が開発した実験車両は、他の自律自動車と外見は同じだが、グーグルやテスラ、ゼネラルモーターズがデモで見せた車とは別の形で、人工知能（AI）の興隆の興隆を示していた。エンジニアやプログラマーからの指示は一切受けず、自動車は人間の運転を観察して自分で覚えたアルゴリズムだけで運転したのだ。

自動車が自分で運転方法を学んで運転できるようになったのは、称賛されて当然の業績だ。ただし、どうにも腑に落ちないところがあるとすれば、自動運転車がどう運転時に判断しているのか完全には分からないことだ。自動車のセンサーが検知した情報が人工ニューロンの巨大ネットワークに送られ、データが処理され、ハンドルやブレーキなど、システムを操作するのに必要な命令が車に伝えられる。その結果、人間が運転しているかのように動作する。しかしある日、もし木に衝突したり、青信号なのに停車したりするなど、予期できない事態を自動車が引き起こしたらどうなるだろうか。現時点で、事故の理由はまず解明できない。自動運転システムは非常に複雑であり、何か1つの動作の原因を突き止めるのは、開発したエンジニアでさえ苦労するはずだ。自動車に突き止めろと聞いても無駄だ。そもそも、ある動作の理

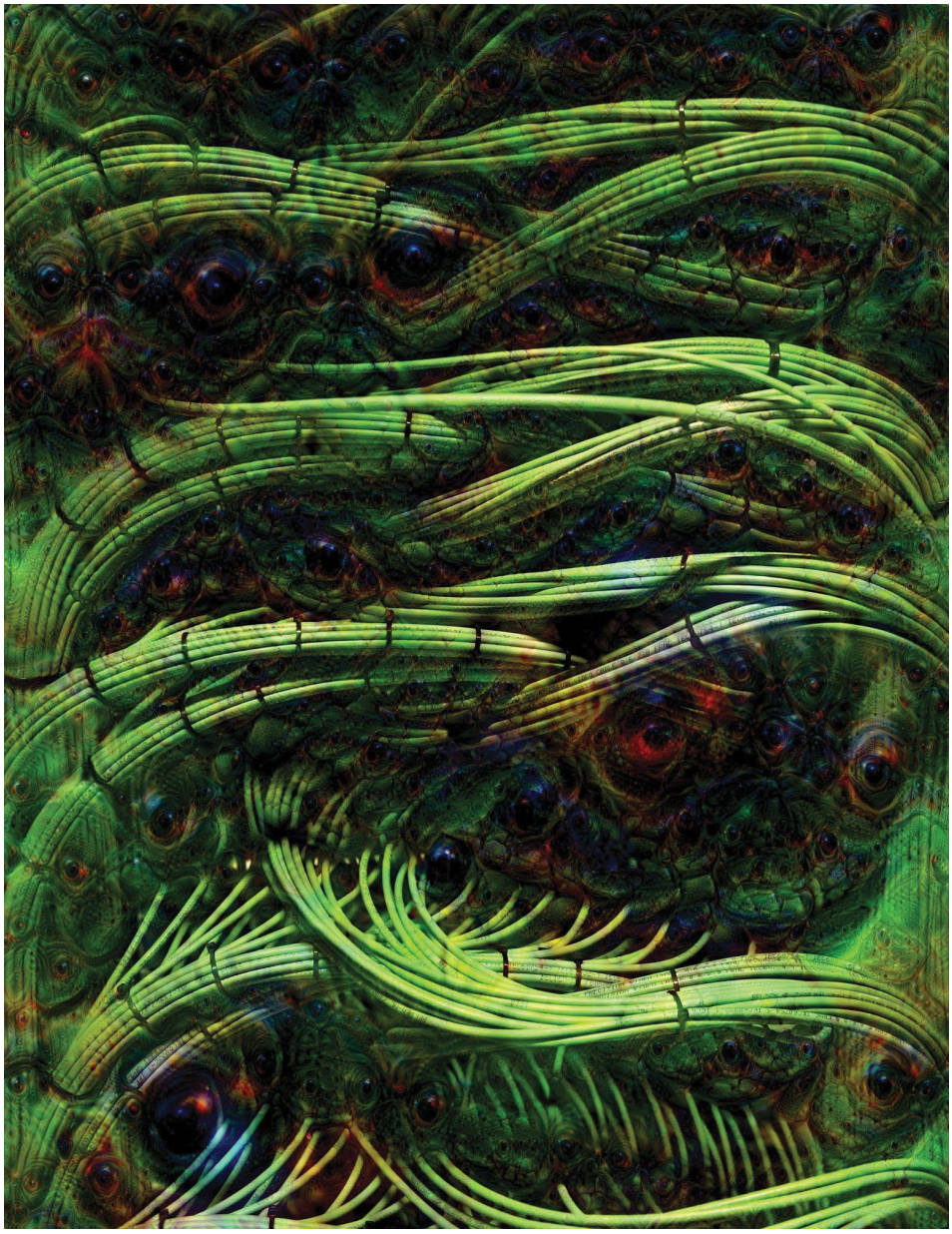
由を常に説明できるような自動運転車を設計する方法がないのだ。

自動運転車の不可解さを知ると、AI の問題点が浮き彫りになってくる。自動運転車に使われているAI テクノロジー「深層学習」は近年、問題解決の非常に強力な手法だと判明し、画像の説明文を自動化、音声認識、言語翻訳などの分野で幅広く使われている。現在、深層学習は、命に関わる病気の診断、数億円単位の売買取引の意思決定など、数多くの業務で使われており、産業界全体を変えると期待されている。

しかし深層学習が社会にうまく浸透するには、深層学習などの機械学習テクノロジーが、まず製品開発者に理解できるようになり、ユーザーの問い合わせに答えられるようにならなければ実現しないし、実現すべきでもない。いつ事故が起きるか予想できないし、誤動作の発生も必然だ。このせいもあって、エヌビディア製自動車は実験車両でしかないのだ。

数学に基づくモデルはすでに米国で、誰を仮釈放するか、誰のローンを承認するか、誰をある仕事に採用すべきかの判断に使われている。数学モデルの中味がわかれば、どんな条件に基づいて判断されているのか理解できる。しかも現在、銀行や軍隊、雇用者は、さらに複雑な機械学習の手法に注目しており、利用範囲の拡大により、自動意思決定のメカニズムは今以上に不可解になる可能性がある。機械学習の手法のうち最もよく使われる深層学習は、プログラムによって動作を定義するコンピューターとは、考え方が根本的に異なる。機械学習の応用を研究しているマサチューセッツ工科大学（MIT）のトミー・ヤコラ教授は「（人間にはAI がなぜそう判断したのか理解できない問題）は既に大問題であり、しかも今後さらにずっと大きな問題になるでしょう。投資や医学の判断、ひょっとすると軍事上の判断で中味の分からない『ブラックボックス』な手法だけには頼りたくありません」という。

すでに、AI システムがなぜそう判断したのかの説明を受けることは、基本的な法的権利ではないかとの議論がある。2018 年夏から、欧州連合（EU）は企業に、ユーザーに自動化システムが



アーティストのアダム・フェリスが制作した画像。7 ページの画像も、フェリスが深層ニューラル・ネットワークのパターン認識機能を刺激するように画像を調整するグーグルのプログラム「ディープ・ドリーム」で制作した。どちらの画像もニューラル・ネットワークの中間レベル層で作られた。

どのように意思決定したかを説明する義務を負わせる可能性がある。とはいえ、深層学習で広告をユーザー別に表示したり、次に聞くべき歌を推薦したりしているアプリや Web サイトなど、表面的には比較的簡単そうなシステムでも、説明は不可能かもしれない。この種のサービスを実現するコンピューターは、機械学習により、コンピューター自身がプログラムを作るのと同様に動作しており、コンピューターがどのように自身のプログラムを調整したのか人類には理解できない。アプ

リを開発したエンジニアでさえ、アプリの動作を完全には説明きないのだ。

AI の動作を誰も説明できない事実を知ると、とんでもない疑問が沸いてこざるを得ない。AI のテクノロジーは進歩しており、今すぐにも、失敗の可能性があると分かりながら、保証のないまま、ある一線を越えて AI を使い続けざるを得なくなる可能性がある。もちろん、人間も自分の思考過程を常に偽りなく説明できるとは限らないが、本能的に人を信じており、人間同士なのだから、あとで失敗（成功）にいたった過程を評価する方法を見つけられる。だが、人間とは別の方法で考え、判断する機械に、この方法は適用できるのだろうか。人類はいままで、作った本人が理解できない方法で動作する機械を作ったことはない。知能のある機械がどう判断するのか予測もできず、理解もできない判断をする可能性があるとき、どうすれば知能のある機械と意思疎通し、共存できるだろうか。この疑問が著者の頭の中に浮かび、実用化されているテクノロジーだけではなく、試験段階の AI アルゴリズムまで調べることにした。グーグルからアップルまで、さまざまな開発段階にある AI を調査し、現代のもっとも偉大な哲学者とも対話することにした。

## えたいの知れない高度な知性

2015 年、ニューヨークのマウント・シナイ病院の研究グループは、患者の記録を収めた巨大データベースの分析に深層学習を適用するアイデアに取り憑かれていた。データセットの特徴は、患者の試験結果、医師の訪問など、数百の変数を設定できることだ。研究グループは完成した分析プログラムを「ディープ・ペイシャント（Deep Patient）」と名付け、70 万人のデータで症例を訓練した。訓練後、新しい記録を分析させると、驚くほど高い精度で、病気を予測できた。ディープ・ペイシャントは、専門家に指導されることなく、肝臓がんなどさまざま病気について、発症途中にあることを示す隠れたパターンを病院のデータから発見したのだ。マウント・シナイ病院の研究チームを率いるジョエル・ダッドリー准教授（次

世代ヘルスケア研究所所長）は「病気の予測について『かなりよい』結果を出す方法はたくさんあります。しかし、ディープ・ペイシャントは『とにかく、ずば抜けている』のです」という。

**「私たちは、病気を予測するモデルを作れます。しかし、モデルがどう動作するのかは分からないのです」**

ディープ・ペイシャントはずば抜けているが、不可解な部分もある。統合失調症など、精神疾患の発症の予測は驚くほど精度が高いが、医師にとって統合失調症は予測が難しいことで有名であり、一体どうすれば予測できるのか、とダッドリー准教授は思った。しかし、ダッドリー准教授はまだ答えが分からない。ディープ・ペイシャントは、どう予測しているかを示す手がかりをまったく与えてくれないのだ。もしディープ・ペイシャントのようなプログラムが実際に医師を支援するなら、予測の論理的根拠を示し、医師に根拠が正確だと安心させ、たとえば患者に現在処方されている薬を変更するのが正しいと示せるのが理想的だ。だがダッドリー准教授は悲しそうに「私たちは、病気を予測するモデルを作れます。しかし、モデルがどう動作するのかは分からないのです」という。

AI が常に理解不能だったわけではない。当初から、AI の動作はどの程度理解できるべきか、あるいは説明できるべきかについて、2 つの考え方があった。多くの研究者はルールと論理によって判断する機械を作り、プログラムを調べれば、誰でもどう動作するのか明白にするのが当然だと考えた。しかし、機械が生物学からインスピレーションを受け、観察と経験によって学習したほうが知能を実現しやすいと感じた研究者もいた。つまり、コンピューター・プログラムの運用を、プログラム自身にさせると考えたのだ。プログラマーが処理内容を書いて問題を解決するのではなく、コンピューター自身が、以前のデータと望ましい結果から逆算し、自分でアルゴリズムを生成するのだ。深層学習や強化学習など、現在もっとも強力な AI システムにつながる機械学習の手法は、本質的に機械が自分自身をプログラムする道から発

展した。

この手法は初め研究も実用化も進展せず、1960 年代、1970 年代には、ほとんど AI 研究の隅に追いやられていた、といっていい。しかし多くの産業でコンピューター化が進み、訓練に使える大規模なデータセットが入手できる環境が整い、新たな関心が生まれた。新しい状況に対応し、さらに強力な機械学習手法、特に「人工ニューラル・ネットワーク」手法の改良版が開発されたのだ。1990 年代には人工ニューラル・ネットワー

クは手書き文字を自動的にデジタル化できるようになっていた。

ただし、小規模の変更や改良があったとはいえ、人工の大規模（深層）ニューラル・ネットワークがコンピューターによる知覚処理を劇的に改善させたのは、2010 年代に入ってからだ。現在、AI が爆発的に普及しているのは深層学習のおかげである。まるで人間のように会話を認識し、複雑すぎて手作業ではプログラム化できない熟練作業など、コンピューターが従来なら考えられない能力を身につけたのは深層学習のおかげだ。深層学習は、コンピューターに対する世間の評価を変え、機械翻訳の品質を劇的に改善した。深層学習は現在、医学や金融、製造など、あらゆる分野で重要な意思決定の支えになっている。

## AI を理解するための試み

すべての機械学習の仕組みは、コンピューター科学者にとっても、手作業でプログラミングされたシステムよりも本質的に不透明だ。ただし、将来のすべての AI 手法が、同様に理解できないとは限らない。深層学習は、その性質上、特に中の見えないブラックボックスなのだ。

深層ニューラル・ネットワークの処理を追いかけても、どう動作しているのかはよく分からないだろう。ニューラル・ネットワークの意思決定は、シミュレートされた数千もの神経同士が繊細に相互接続された数十〜数百の層の振る舞いとして実現されている。第 1 層にあるそれぞれの神経は、たとえば画像の 1 ピクセルごとの明度などの入力を受け取り、何かを計算して明るい暗い、黄色い、青い、といった信号を出力する。複雑なニューラル・ネットワーク内で、こうした出力が次の層の神経に渡され、同様の処理が次々に繰り返され、ニューラル・ネットワーク全体として、これはネコ、「あ」と言った、などの出力が生成されるのだ。さらに「誤差逆伝搬」プロセスにより、各神経の計算結果は答えから微調整され、ニューラル・ネットワーク全体の出力がなるべく正しくなるように学習する。

深層ネットワークには多くの層があるため、物