人工知能と公平性

久木田水生

アウトライン

現在の人工知能

ビッグデータに基づく人工知能が引き起こす差別

メディアとしての人工知能

人工知能のメッセージ

テクノロジーは価値中立ではない

Outline

ここで論じる人工知能

「人工知能」という言葉は多種多様なテクノロジーの曖昧な総称である。ここでは現在最も活用され、かつ最も大きな富を生んでいるであろう、**ビッグデータに基づく機械学習システム**について論じる。

以下では「人工知能」は主にこのようなシステムを指すものとする。

機械学習のブレイクスルー

Jeffrey Hinton らが、従来のニューラル・ネットワークよりも大きく複雑な構造を持つ **Deep Neural Network** による**深層学習**の手法を開発。

教師なし学習、畳み込みニューラル・ネットワーク、強化学習、敵対的生成ネットワーク（GAN）などの新しい画期的な手法が生み出される。

2010 年ごろから、画像認識において大幅な精度の向上。

「第三次 AI ブーム」に火をつける。

背景：技術的要因

- ハードウェアの進歩により、膨大な計算が短時間でできるようになった。
- 機械学習の新しい効果的なテクニックが開発された。
- 機械学習には必要な**大量のデータが、インターネットやスマートフォンなどの普及によって容易に入手・利用できるようになった**。

背景：社会的経済的要因

- カスタマイズされたサービスや広告を提供するためにユーザーのデータから属性や行動傾向を予測するというネットビジネスモデルが主流になっている。Cf. アマゾンや YouTube、Netflix などの推薦システム、Google やフェイスブックなどのターゲッティング広告。
- ネットワーク効果による巨大プラットフォームへの資本とデータの集中 → 人工知能とデータ科学への巨額の投資。

人工知能の様々な応用

- 画像認識、音声認識。
- 自然言語による対話システム。
- 機械翻訳。
- 画像や動画、音声や音楽などの生成。
- ターゲッティング広告。
- プロファイリング（人間の属性や行動の予測）。
- 自動運転。

懸念されている問題

- 安全性、信頼性、透明性、答責性、制御可能性。
- プライバシーの侵害。
- 巨大プラットフォーム企業によるデータの独占と濫用。
- 不平等の拡大、差別の助長。
- 自律型兵器。
- テクノジー失業。
- ソーシャル・ロボットによる人間同士の関係の変化。
- 人工知能やロボットの地位。

データ経済と人工知能

現在の人工知能について論じる際、データ経済経済と切り離して考えることはできない。

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、などなどの普及により、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を予測、推測する**。

データが集まれば集まるほど推測できることが増え、またその精度も増していく。

「監視資本主義」

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、どの需要が、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。

それゆえ**データやそれに基づく予測が市場において大きな価値を持つ**ようになる。

Cf. S. Zuboff, *The Age of Surveillance Capitalism*.

行動変容

さらに適切なタイミングで情報を与えることで人々の**行動を変容**させることすらできる。

このような手法はオンライン・マーケティングを超えて、**選挙や紛争にも組織的に利用されている**。

Cf. P・W・シンガー、エマーソン・T・ブルッキング、『「いいね！」戦争――兵器化するソーシャルメディア』；ブリタニー・カイザー、『告発　フェイスブックを揺るがした巨大スキャンダル』；ジェイミー・バートレット、『操られる民主主義――デジタル・テクノロジーはいかにして社会を破壊するか』；一田和樹、『フェイクニュース――新しい戦略的戦争兵器』

Outline

人工知能によるプロファイリング

根拠

- 消費行動
- クレジットカードのスコア
- 健康に関わる習慣
- 住んでいる場所
- 人間関係
- ソーシャルメディアの活動
- などなど

判断

- 採用、人事評価
- 予測警察、再犯の可能性の予測
- ローンの審査
- 保険の料金
- などなど

「人工知能だから公平です」はペテン師の言葉

アルゴリズムには作成者のバイアスが入り込む。

データには社会の持つ偏見が入り込む。

人工知能は公平でも、客観的でもない。

なぜペテンを信じるのか？

人は信じたいものを信じる。

ビッグデータと人工知能が、**複雑な世界の複雑な問題に、簡単な「解決」をもたらしてくれる**、と人々は信じたいのかもしれない。

あるいは人工知能が**自分たちの持っている偏見や先入観を確証してくれる**からこそ、それを信じるのかもしれない。

参考

AI Now, 「「AI」による差別の現状とは？ 事例、原因、世界各地の取り組みを紹介」、2020 年 2 月 17 日、https://ainow.ai/2020/02/17/183256/#6

Frederik Zuiderveen Borgesius, “Discrimination, artificial intelligence, and algorithmic decision-making”, The Directorate General of Democracy, Council of Europe, 2018. https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73

アルゴリズムがどのようにして差別に結び付くか

AI に識別させたい対象の性質や、分類のための基準の選択が差別を引き起こしうる。

例えば社員としての能力の評価に遅刻や欠勤の数というファクターを入れると、住んでいる場所、利用できる交通手段、経済的状況、健康状態などによって不利になる人々がいる。

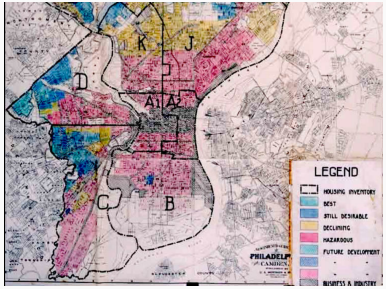
アルゴリズムがどのようにして差別に結び付くか

意思決定に用いる**近似的特徴**が実際には差別になっていることがある。

例えば住所に基づいた判断が実質的には特定のエスニックグループに対する差別になりうる。

個人や組織が意図的に近似を使って特定のグループに対する差別を行なう場合がある。

デジタル・レッドライニング



「レッドライニング」はアメリカの金融業界の、特定の地域の住民に融資を厳しくするといった慣行を指す言葉。

現在、ビッグデータと人工知能によって、この不公平な慣行と実質的に同じようなことが、はっきりと目に見えない形で行われていることを指して、「**デジタル・レッドライニング**」と言われる。

データ収集における偏り

人間の偏見を反映したデータから AI はバイアスを学習する。

たとえば警察が特定の民族集団に対して偏見を持っている場合、その集団により多くの犯罪が見つかる。これに基づいた AI による犯罪予測はさらにデータの偏りを強化する可能性もある。

バッド・プラクティス

- 警察、犯罪予防
  COMPAS というシステムが米国の一部で被告の再犯可能性を見積もるために使われている。このシステムでは人種や肌の色などの情報には基づいていないが、調査によると黒人に不利なバイアスを持っていることが分かった。実際には再犯をしなかったが高リスクと評価された人間は、黒人では白人の倍に上った。逆に白人は低リスクと評価されながら再犯を犯す確率が黒人よりずっと多かった。
- 社員や学生の選別
  社員採用や学生入学試験にも AI が使われるが、そこにも差別が生じ得る。アマゾンは、応募者をスクリーニングをするのに AI システムを使っていたが、女性に不利なバイアスを持っていたためにシステムの使用を中止したと報じられた。

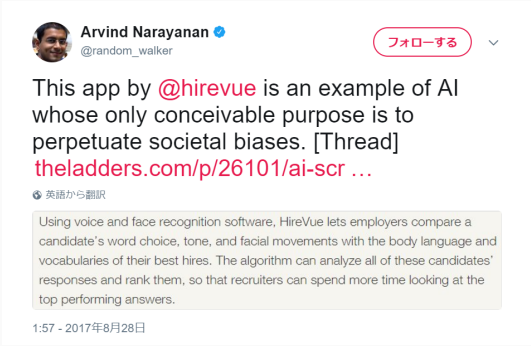
バッド・プラクティス

- 広告
  - Google 検索ではアフリカ系米国人の名前を検索すると、誰それがかくかくの犯罪歴を持つ、ということを示唆する広告を表示する。
  - ユーザー設定で男性とする方が女性とするより、高給の仕事の求人が表示される。
  - Facebook は広告主が、センシティブなデータに基づいて人々に標的広告を行なうことを可能にしている。
  - Facebook は広告主が、特定の民族グループに広告を表示しないようにすることを可能にしている。

バッド・プラクティス

- 価格差別
  オンラインショップはユーザーの過去の購入データに基づき、ユーザーごとに価格を変化させている。これにより他のオプションが少ない地方の人間が高い価格を提示されることがある。
- 画像検索
  Google の検索エンジンで「three black teenagers」と検索すると、mug shots（警察が撮った容疑者の写真）が表示されるが、「three white kids」と検索すると、ほとんどは楽しげな白人の子供が表示される。

バッド・プラクティス



「@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」

https://twitter.com/random_walker/status/901851127624458240

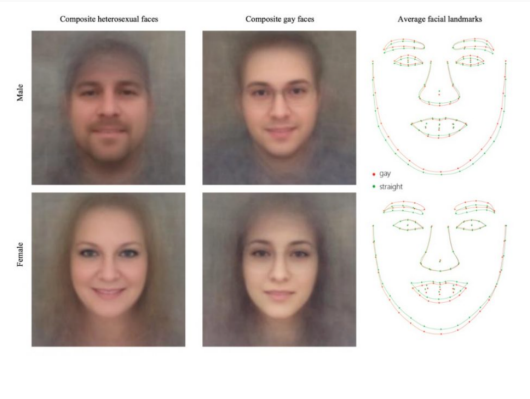
「HireVue は声と顔認識ソフトを利用し、雇用の際に、採用候補者たちの言葉遣い・声のトーン・表情と、最も優秀な社員たちのボディーランゲージ・語彙とを比較することを可能にする。このアルゴリズムは候補者の応答を分析してランク付けすることができる。それによって採用者たちは上位の候補者たちの返答を検討するのにより多くの時間を費やすことができる。」

バッド・プラクティス

http://www.itmedia.co.jp/enterprise/articles/1806/06/news006.html

東京大学では「魅力工学」の一環として、女性の顔の魅力を AI で数値化するという研究が行われている。なぜ女性だけかと言えば、「女性の顔については、古今東西、大体同じ評価尺度が存在することが、心理学の分野で判明している。一方、男性の顔の魅力については、複数の評価尺度があるといわれている。現段階では、数値化の精度を高めるのは難しい」からだという。

バッド・プラクティス



スタンフォード大学では、顔の画像から性的指向（同性愛、異性愛）を判別する AI を開発。

https://shiropen.com/2017/09/08/28032

Outline

人工知能はメディアである

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。

「メディアはメッセージである」

マーシャル・マクルーハンの言葉。

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。

同じ内容のメッセージでも、**どのメディアによって伝えられるかによって、受け取った人間に与える効果は異なる**。

どのメディアを使うかということは、メッセージの内容以上に重要な意味を持つ。

「メディアはメッセージ」

ラジオの時代になって政治家の演説の長さは 1 時間から 10 分程度になった。

テレビの普及は見栄えの良い政治家に有利に働く。

カナダでの国政選挙において立候補者の得票数と外見的な魅力の関係を調査した研究では、魅力的な立候補者はそうでない立候補者よりも多くの票を得ていた（得票率は平均して 32 %と 11 %だった）ことが分かった。

Michael G. Efrain and E. W. J. Patterson, (1974) "Voters vote beautiful: The effect of physical appearance on a national election", *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, 6(4), 1974, 352–356. https://doi.org/10.1037/h0081881

メディアとしての人工知能の特性

伝統的なメディアが基本的に事実を伝えるのに対して、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**。

もちろん新聞やニュースも推測を伝えるが、それは人間が行った推測を伝えているだけである。

それに対して人工知能は機械が行なった推測を伝える。

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**である。

Outline

人工知能の利用例

人工知能の開発者、利用者は**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**。

例えば、労働者としての能力や適性、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。

人工知能を使う動機

人々が人工知能を使う理由は**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や内定を辞退し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。

意思決定とリスク分析

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。

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。

人工知能が伝える情報

要するに人工知能は、**人間を様々なデバイスから取得された機械可読なデータの集積、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。

Remote Risk Assessment

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。

Remote Risk Assessment

このシステムは電話越しの十分程度の会話（決められた質問に対して母国語で答える）をもとに、その内容ではなく声の調子のみによって相手が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、直面している歴史的な難民危機（monumental refugee crisis）に対する究極のソリューション」と宣伝している。

AC Global Risk 社はソフトウェアの詳細に関する The Intercept の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。

Remote Risk Assessment

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。

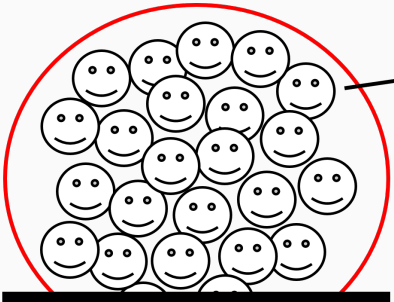
人工知能の長期的な意図しない影響

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。

アンフェアな推論



このグループのうち、8割の人が
暴力的傾向を持っている

この人物は80%の確率で
暴力的傾向を持っている

この人物は危険である

リスク分析の本質

政策決定者や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。政府や企業は全体的な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。

特に企業においては、重要なのは金銭的な収支であり、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。

人工知能のメッセージ

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**。

他者をリスクとして扱うことの弊害

家族、友人、パートナー、メンターと弟子のような、特別で個人的で親密な関係は、成立する前に「この相手は自分にとってどれだけ利益・損害をもたらすか」というような態度で臨んでいては構築できない。

信頼に基づく協力関係は、まず理由もなく、信頼することから始まる。人間は信頼されればそれに応えようと努力するものである。つまり信頼がきっかけになって、相手が本当に信頼に足る人間になるということがある。

他者をリスクとして扱うことの弊害

		プレイヤー2の選択	
		協力	裏切り
プレイヤー1の選択	協力	(4, 4)	(0, 6)
	裏切り	(6, 0)	(2, 2)

括弧内の数値は、左がプレイヤー 1 の得点、右がプレイヤー 2 の得点を表す。

Outline

ある人工知能研究者のツイート

2019 年 12 月にある人工知能研究者が、特定の国を名指し、自分の会社ではその国の人間を雇うことはしない、とツイート。批判を浴びると、AI による自動選考では人種や国籍、年齢や性別もパラメータとして入力しており、それらの属性とパフォーマンスの間に因果関係がある、と主張。

その後、特定の国籍の人々の能力に関する自分の会社の判断は AI が限られたデータに適合しすぎた結果の「過学習」によるものと釈明し謝罪した。

学会のリアクション

2019 年 12 月 10 日、人工知能学会倫理委員会、日本ソフトウェア科学会 機械学習工学研究会、電子情報通信学会情報論的学習理論と機械学習研究会が共同で「機械学習と公平性に関する声明」を発表。

そこで「機械学習は道具」、「不正を引き起こすのは人間」ということが強調されていた。http://ai-elsi.org/archives/888

学会のリアクション

「機械学習はあくまでも道具にすぎず、その使い方を定めるのは人間です。機械学習は人類社会の繁栄に大きく貢献できる可能性を秘めているとともに、不適切な利用をすれば人類社会の利益に反する可能性もあります。機械学習は過去の事例に基づいて未来を予測しますから、偏りのある過去に基づいて予測する未来は、やはり偏りのあるものになりかねません。もし、過去と異なる「あるべき未来」を求めるのであれば、機械学習による予測や判断が公平性を欠くことがないように人間が機械学習に注意深く介入する必要があります。」

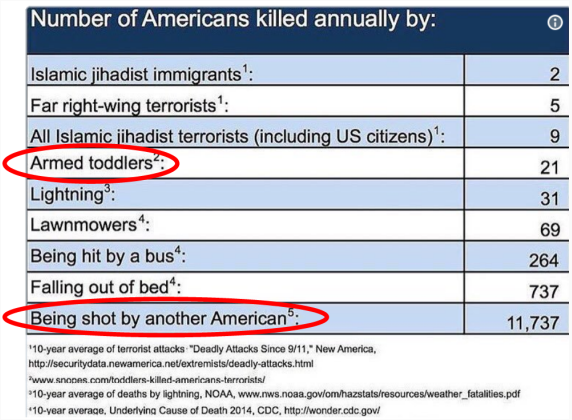
学会のリアクション

「同時に、「何が公平か」については、科学技術や工学だけの問題ではなく、現在の人類社会が何を求めているか、という価値観の問題抜きには語れません。機械学習という「道具」を正しく使うためには、それが「公平性」という私たち人類社会の価値観に対して、どのような影響を与えるかを正しく理解し、そのリスクを評価し、方策について合意しなければならないのです。この点は、私たちだけではなく、機械学習に携わる技術者や利用者、経営者、そして組織や社会の全体が把握し向き合っていく必要があります。」

http://ai-elsi.org/archives/888

「機械は中立」か？

- アメリカでは年間 1 万人以上が銃で殺されている。
- 幼児による銃の誤射で平均 20 人ほどが犠牲になっている。



Number of Americans killed annually by:	
Islamic jihadist immigrants[1]:	2
Far right-wing terrorists[1]:	5
All Islamic jihadist terrorists (including US citizens)[1]:	9
Armed toddlers[2]:	21
Lightning[3]:	31
Lawnmowers[4]:	69
Being hit by a bus[4]:	264
Falling out of bed[4]:	737
Being shot by another American[5]:	11,737

[1]10-year average of terrorist attacks "Deadly Attacks Since 9/11," New America, http://securitydata.newamerica.net/extremists/deadly-attacks.html
[2]www.snopes.com/toddlers-killed-americans-terrorists/
[3]10-year average of deaths by lightning, NOAA, www.nws.noaa.gov/om/hazstats/resources/weather_fatalities.pdf
[4]10-year average, Underlying Cause of Death 2014, CDC, http://wonder.cdc.gov/

http://www.euronews.com/2017/01/31/ armed-toddlers-kill-twice-as-many-americans-each-year-than-terrorists

「機械は中立」か？

THE FAMOUS ANTI-GUN CONTROL SLOGAN:

"Guns don't kill people. People kill people."
The unofficial slogan of the National Rifle Association
Contrary to popular belief, this is not an official NRA slogan, though it has been used by NRA members and other gun rights advocates since at least the 1950s. Ironically, it has also been used as a gun safety slogan. For example, in a May 31, 1959 *Associated Press* story I found in NewspaperArchive.com, Fred A. Roff Jr., president of the Colt Patent Fire Arms Co., was quoted as saying *"Our big concern is to make sure that guns get into the hands of only those who know how to use them. Guns don't kill people. People kill people."*

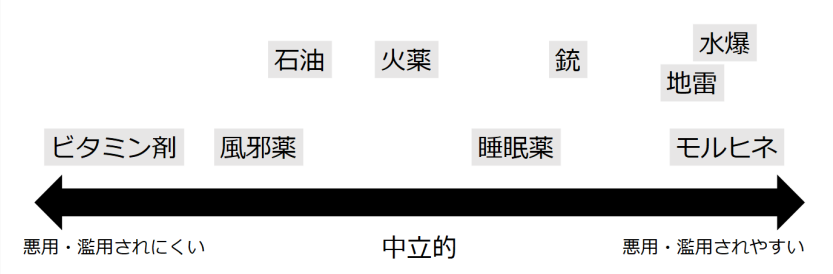
全米ライフル協会はよく「銃は人を殺さない、人が人を殺す」という言葉を使う。

これは使う人がいなければ誰も被害にあわないという意味では自明に正しい。

しかし銃による犠牲者の多くが、銃がなければ命を失わなかったということもまた事実である。

http://www.quotecounterquote.com/2011/03/guns-dont-kill-people-people-kill.html

道具の中立性には度合いがある



道具には悪用・濫用されやすいバイアスを持つものとそうでないものがある。

すべての道具を中立と考えることは、道具の適切なコントロールを難しくする。

DeepNude という、着衣の助成の画像から裸の画像を生成するアプリが開発、公開された。しかし想像した以上に多くのユーザーがダウンロードしたために、開発者アプリをダウンロードできなくして、有料バージョンの購入者には返金を行った。

開発者は次のように述べる。「安全策（ウォーターマーク）が講じられているとはいえ、50 万人もの人が利用していれば、それが悪用される可能性はあまりにも高い。私たちはこんなやりかたでに金儲けをしたくはない。……世界はまだ DeepNude を使う準備ができていない。」

https://twitter.com/deepnudeapp/status/1144307316231200768

まとめ

- 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」は詐欺師か詐欺師に騙されている人間の言うセリフ。

まとめ

- 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、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可能性がある。
- 機械に責任を転嫁することは誤りだが、「機械は道具であり、良いも悪いも使う人間次第」という考えも無条件に正しいわけではない。
- 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。
- 機械そのものが持つバイアスを認識することが、適切なコントロールにとって重要。
- 世界はまだ人工知能を使う準備ができていないのかもしれない。