5ZJ-03

顔認証システムの倫理的課題について
-人工知能とバイアス-

冨島　悠介†　渡邉　聖†　齋藤　太一†　西川　真央†　北村　智花†　山本　将一朗†　今　諒平†
中央大学　国際情報学部‡

1.はじめに

人工知能（以下、AI）への注目が高まる中、顔認証(facial recognition)技術は最も代表的な研究の一つとして挙げられる。わが国でも JR 東日本が顔認証システムを用いて不審者の検知を行う事例など、顔認証システムが社会に普及し始めていた[1]。現在わが国だけに留まらず、世界中で使用されている顔認証システムであるが、その倫理的側面に未だ課題が残る。それは人種によって認証精度に偏りがある点だ。2020 年 1 月にアメリカのデトロイト市警察が顔認証システムの誤認識によって罪のないアフリカ系アメリカ人の男性を逮捕した事件によって、この問題が浮き彫りとなった。さらに 2020 年 5 月の BLM（ブラック・ライブズ・マター）運動を受けて、IBM が顔認証や分析のためのソフトウェアの停止を表明した[2]。これに続くように Amazon や Microsoft も同様の決断を下したほか、政府や国際機関も AI の倫理的なコンセンサスの形成を進めている。2021 年 11 月に採択された UNESCO の AI 倫理勧告では、「人種や肌の色、家系、性別、年齢、言語、宗教、政治的見解、国籍、民族、社会的起源、出自の経済的又は社会的状態、障害、その他の立場に関係なく」AI が利活用されることに繰り返し言及している[3]。AI 又は AI を用いた顔認証システムに対する倫理的課題が浮き彫りになっている現在、顔認証システムが安全に社会で利活用されるために何が必要かを検討する必要がある。

本研究では、AI を用いた顔認証システムのデータバイアス問題に触れつつ、AI の利活用時に人間が無意識に持つ潜在バイアスの減少の必要性を主張する。また人間の持つ潜在バイアスの減少のために、行動経済学のナッジを活用した解決モデルを作成する。

The ethical issues of facial recognition system -Artificial intelligence and its bias
† Yusuke Tomishima, Hijiri Watanabe, Taichi Saito, Mao Nishikawa, Tomoka Kitamura, Shoichiro Yamamoto, Ryohei Kon
‡ Faculty of global informatics, Chuo university

2. 顔認証システムが抱えるデータバイアスの現状と人間の持つ潜在バイアス

前述のとおり、顔認証システムは人種、つまり肌の色によって認証の精度に偏りが生じる。米国国立標準技術研究所（NIST）による顔認証技術における人口統計学的調査についての言及では、アフリカ系アメリカ人女性に対する誤認識率が高く、冤罪を招く危険性が高いことが述べられている[4]。

現在、このような誤認識率を限りなく低くするべく研究が進められているが、たとえそれが実現した社会でも顔認証システムには未だ課題が残るだろう。なぜなら人間が意識しないうちに抱いてしまう偏見（以下、潜在バイアス）が安全な利活用を阻害する恐れがあるからだ。それを最も示唆している事例が、AI 再犯予測システム「COMPAS」である。再犯の高リスク判定が黒人に偏るという問題の理由はデータセットが人間の潜在バイアスを反映させている点にある。すなわち、「犯罪者としての黒人」という過去長い期間を通じてアメリカ社会で存在し続けたバイアスにより、COMPAS 自身も「人種と犯罪」という関係性におけるバイアスを強化しているのである[5]。AI を用いた顔認証システムの安全な利活用のためには、人間社会に根付く差別を助長する潜在バイアスを減らす必要がある。我々はこの人間の持つ潜在バイアスを減少させていくために、行動経済学のナッジ理論の活用を提案する。

3.ナッジを利用した潜在バイアスの減少モデルの提案

ナッジ（nudge）とはその言葉通り「肘で軽く突く」という意味をもち、行動経済学上では「選択の自由を維持しながら人々を望ましい方向に導く」仕組みを指す[6]。ここでは、サスティーン氏が提示する 2 つのナッジの区別ごとにそれぞれモデルを提案する。1 つ目は人々が自分自身の行為主体性の力を高めることを目的とす

る「教育的ナッジ」、もう 1 つは人間の脳に直感的に働きかける「非教育的ナッジ」である[7]。

3.1 教育的ナッジ：IAT の活用と潜在バイアスの自認

教育的ナッジの一案として我々は IAT(Implicit Association Test: 潜在連合テスト）を活用していく。このテストでは、「白人-黒人」「良い-悪い」という組み合わせを用意し、評価対象を画面の左上又は右上どちらかに分類するテストを行うことで潜在バイアスを測定することができる。無意識下で強く連合する言葉の分類作業を通じて、固定観念や偏見、差別を見極めることが出来る。

図 1：IAT のテスト画面例[8]

我々は顔認証システム使用者、特に AI の出したアウトプットに基づく意思決定をする者に対し、IAT の受講を義務づけることを 1 つ目の潜在バイアス減少モデルとする。IAT を受講することで自身の潜在バイアスを自認し、それに基づいた行動を減少させることが期待できる。

3.2 非教育的ナッジ：ポスター掲示による視覚的効果

視覚情報によって与えられる影響は大きいため、我々は視覚効果を用いることで人々がもつ潜在バイアスに直感的に訴えることが可能になると考える。実際に視覚的ナッジの成功事例として、京都府宇治市役所の取り組みが挙げられる。施設の出入り口に設置した消毒液の使用を推奨するため、人々を誘導する黄色い矢印を床に貼り付けた。その結果、消毒率が 9.7%向上したという。矢印という視覚的ナッジによって人々が無意識的に誘導されていることがわかる[9]。

上記の例のような視覚的効果は人々へのナッジとして有効である。そこで我々は非教育的ナッジに基づき、肌の色が異なる人間が一緒に写っているポスターの掲示を 2 つ目のバイアス減少モデルとする。図 2 では視覚的ナッジのためのポスター例を掲示している。以下のように、非有色人種と有色人種のクローズアップした顔写真を並べることで、二者の間には肌の色以外には何ら差がないことを閲覧者に示唆する。これによって直感的な人種差別を助長する潜在バイアスの減少が期待できる。

図 2：視覚的効果のためのポスター例[10]

4. 終わりに

本研究では、AI を用いた顔認証システムの安全な利活用のために、人間による差別を助長する潜在バイアスに着目し、減少させるためにナッジを用いて 2 つの潜在バイアス減少モデルを提案した。今後は提案した 2 つのモデルによる実施効果をより具体的に観察し、議論を深めていく予定である。

参考文献

1. 東日本旅客鉄道株式会社(2021).「東京 2020 オリンピック・パラリンピック競技大会に向けた鉄道セキュリティ向上の取り組みについて」.
https://www.jreast.co.jp/press/2021/20210706_ho02.pdf
2. IBM(2020).「IBM CEO's Letter to Congress on Racial Justice Reform」.
https://www.ibm.com/blogs/policy/facial-recognition-sunset-racial-justice-reforms/
3. UNESCO(2021).「DRAFT RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE」.General Conference 41st session. 筆者訳.
4. NIST(2019).「NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software」.
https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software
5. 前田春香（2020）.「人工知能は道徳的に悪質な差別ができるか？―COMPAS 問題を事例に」.第 34 回人工知能学会全国大会論文集. セッション ID：2Q4-OS-13a-01.
6. リチャード・セイラー, キャス・サンスティーン, 遠藤真美（訳）(2009).『実践行動経済学―健康、富、幸福への聡明な選択』.日経 BP 社.
7. キャス・サスティーン, ルチア・ライシュ, 遠藤真美（訳）(2020)『データで見る行動経済学：全世界大規模調査で見えてきた「ナッジ（NUDGES）の真実」』.日経 BP 社.
8. IAT Corp.「IAT テスト ホームページ」.
https://implicit.harvard.edu/implicit/japan/
9. 柴田浩(2020).「行動経済学のナッジが消毒・手洗い行動に変容を及ぼす効果の検証について」.
http://www.env.go.jp/earth/ondanka/nudge/renrakukai16/mat_02.pdf
10. GAHAG「著作権フリー写真・イラスト素材集」
http://gahag.net/006109-women-portrait/
（いずれも 2022 年 1 月 5 日最終閲覧.）