ブラックボックスと正義の政治——AIと差別——

五十里翔吾

1　はじめに：我々は中にいるか、外にいるか？

AI（人工知能）の実装に用いられる機械学習の技術が、人間が行う差別を学習するという問題が指摘されている。機械学習とは、与えられたデータからルールや判断基準、知識などを抽出する技術である。機械学習の技術は、我々の日常生活に欠かせない。例えば、検索エンジンや通信販売サイトのサジェストやスパムメールの仕分け、スマートフォンのカメラに備わる顔認識機能などその応用は多岐にわたる。

多くの場合、多様で複雑なデータを扱う機械学習システムはブラックボックス的になる。例えば、その技術の一つであるディープラーニングにおいては、時に数千万以上ものパラメータがデータに適応するように調整されることで、学習が実現される。プロ囲碁棋士を破った初のコンピュータ囲碁プログラムAlphaGoにはディープラーニングが用いられている。そして、AlphaGoがどのような戦略を立て、次の一手を選択するのかという決定プロセスを人間が理解することは不可能である。[1]

機械学習は人が開発した技術である。技術であるから、それを利用することができる。このように考えるならば、我々はこのブラックボックスの「外部」にあって、それを道具として利用しているのだと感じられよう。しかし、通販サイトのサジェストは、我々の購入履歴を学習し、「買いそうな」商品を提示する。また、画像に何が写っているのかを判断するシステムを開発する場合、学習に用いられるデータは、「ある画像」とそれが何の画像かという「正解」のラベルの組からなっており、その「正解」は人間が与えるのだ。そのような作業を仲介し、インターネット上の人々に外注するAmazon Mechanical Turkのようなシステムもある。また、あるサイトのログイン画面で「私は機械ではありません」ということを示すためにユーザが行う文字や画像の識別という簡単な作業が、学習用のデータを整備するのに活用されている。[2] このような見方をすれば、機械学習というブラックボックスを、ある機能を果たす「関数（ファンクション）＝函数」という箱に見立てた場合、我々はその中にいるとも言えるのである。

しかし我々は、単に箱の中にいるだけではない。

例えば、「人がどれだけ信用できるかを予測する」システムを作りたいとしよう。このようなシステムは、クレジットカード会社や銀行も欲しがるだろう。すると、どんな情報を入力して学習させればいいだろうか。

---

1 圧勝「囲碁AI」が露呈した人工知能の弱点　日本経済新聞 web　2016/3/17
https://www.nikkei.com/article/DGXMZO98496540W6A310C1000000/

2 reCAPTCHA
https://www.google.com/recaptcha/intro/v3.html#creation-of-value

学歴だろうか？職種だろうか？このシステムは一体何を学習するのだろうか。学歴、職業、交友関係、住所、人種、ジェンダー、母語、年齢、障害など、個人に関わるあらゆる情報は、再生産される社会経済的基盤、そして文化的構造からは切り離すことはできない。このことを考慮するならば、以下のような事態が発生する可能性に思い至る。——このシステムは人々が行う差別を学習するかもしれない——「〇〇地区の人間は信用できない」「〇〇人は嘘を付く」このブラックボックスは我々の社会が抱える「差別・偏見」が再現された「箱庭」となるかもしれない。すると、我々は庭師だ。

そんな事態は、現に進行している。本稿において、２章では機械学習を用いたＡＩシステムにおける差別の実例を紹介する。３章ではその原因を考察し、既存研究における対策を紹介し、その限界を明らかにする。４章ではこの問題に取り組む上での包括的な指針を得るための予備的な考察を行う。

では、ブラックボックスの中で、外で、その狭間で、何が起こっているのか。まずは、この「箱」を外側から見てみることにしたい。

2　機械学習による差別

マイクロソフトの研究者で、ＡＩの社会的影響を研究する機関 AI Now[3]の協同設立者ケイト・クロフォードは、機械学習における差別的効果（disparate impact）を**配分型**（allocative harm）と**象徴型**（representation harm）の二種類に分類している[11]。配分型の差別とは、（機械学習を用いた）システムが、特定の人達に対してある機会やリソースへのアクセスを不公平に処理することである。象徴型の差別とは、システムが特定の集団のアイデンティティを貶めることに繋がる出力を行うことである。この分類は、後に四章で検討するが、哲学・あるいは政治学の用語として用いられる「再配分（redistribution）」と「承認（recognition）」という二つの概念と対応している。次節において、近年の事例から、それぞれのケースに当たるものを紹介する。

2・1　機械学習による差別の事例

配分型の差別

Amazon の人材採用ツール

2014 年から、Amazon はより有能な労働者を採用するために、候補者の能力を予測するシステムを開発してきた。しかし、開発中の実験で、ソフトウェアエンジニアや一部の技術職において、ジェンダー間に不公平な見積もりを行っていたことが明らかになった。[4] そのため、Amazon はこのシステムを実際の採用プロセスには使用しなかった。具体的には、履歴書中の「women」や「womens' chess club」といった単語や、一部の女子大学名を低評価と結びつけていた。さらに、男性の履歴書に多く見られた「executed」や「captured」といった単語を高評価と結びつけていた。このシステムにおいては、過去十年間の採用候補者のデータを学習に用いており、候補者には男性が圧倒的に多かった。そのため、このような

[3] https://ainowinstit ute.org/

[4] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-AI-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

ジェンダー中立でない学習が行われたのである。(Amazon は二〇一七年以降技術職員のジェンダー比率を公表していないが、Google、Facebook、Microsoft において、技術職にある労働者の 80 %前後が男性である)

再犯可能性予測システム (COMPAS)

Northpointe 社が開発した COMPAS(Correctional Offender Management Profiling for Alternative Sanctions) は、犯罪者の再犯リスクや、犯罪的性格、薬物乱用のリスクなど 20 以上の「犯罪兆候」を予測するシステムである。ニューヨークの裁判所では、二〇〇一年から導入されており、全米の多くの州で量刑に利用されている。二〇一二年に行われた保護観察下にある犯罪者のうち、71 %において再犯リスクを正確に予測できた。しかし、二〇一六年にフロリダ州の七〇〇〇人の犯罪者を対象に行われた実証実験 [8] で、システムが人種に基づいて差別的な判定を下している可能性が明らかになった。「再犯リスク高」と予測され、**再犯をしなかった**白人の犯罪者は約 24 %であった一方、同様の予測をされたアフリカ系アメリカ人のうち、約 45 %が再犯をしなかった。逆に、「再犯リスク高」と予測され、**再犯した**白人の犯罪者は 48 %ほどであった一方、同様の予測をされたアフリカ系アメリカ人のうち、28 %が再犯を犯した。

Northpointe 社は、COMPAS の入力変数に「人種」を採用していないと主張している。もしこれが真実であるとすれば、機械学習を用いたシステムが、入力変数以外から差別的な構造を学習している可能性がある。同様の結果が、個人の収入を予測するための学習を検討した研究によっても報告されている。国勢調査のデータセットを用いた学習を行った結果、「ジェンダー」の項目を入力変数に用いた場合よりも、用いなかった場合のほうが、男性を高収入、女性を定収入であると判別する割合が高くなった [2]。

象徴型の差別

瞬目検知システム

Nikon S630 デジタルカメラは、レンズに写った人の顔を認識し、瞬きを検知して使用者に知らせるというシステムを搭載している。二〇〇九年、このシステムがアジア人女性が常に「瞬きしている」と判別したというケースが報告されている。[5] 同様の問題が、HP 社のウェブカメラ HP Pavillion においても報告されている。Time.com はこの問題の原因は、学習アルゴリズムに内在すると指摘している。このシステムは入力画像から解像度を落として画像を処理しており、細い目と閉じた目を区別できなかったというのだ。[6]

オンライン広告提示システム

google.com のような検索エンジンでは、入力した単語に対してページ上部に広告が表示されることがある。どのような広告が表示されるかは、機械学習を用いたシステムによって決定されている。instantcheckmate.com は、検索エンジンに広告を表示しているウェブサイトの一つだ。このサイトでは、アメリカ国民の様々な個人情報を名前と紐付けて記録しており、例えば「Shogo IKARI」というように人名を入力して検索した場合「We Found: Shogo IKARI」や「Shogo IKARI, Arrested」といった広告

[5] https://thesocietypages.org/socimages/2009/05/29/nikon-camera-says-asians-are-always-blinking/

[6] https://petapixel.com/2010/01/22/racist-camera-phenomenon-explained-almost/

が表示される。ハーバード大学のラタニア・スウィーニーは二〇一二年に行った研究[12]で、google.com と reuters.com において、表示される広告を「Found」「Located」といった中立的な広告と「Arrested」広告に分け、名前と広告の文面がどのように結び付けられているのかを分析した。その結果、アフリカ系に多く見られる名前を検索した場合のほうがヨーロッパ系に多く見られる名前を検索した場合よりも「Arrested」という内容が表示される確率が高く、表示される広告の内容が人種という変数に従属していることが統計的に有意であることが示された。

本節では、機械学習を用いたシステムが差別的な情報を出力した例を挙げた。これ以外にも、機械翻訳において特定の単語が特定のジェンダーと結び付けられて訳出されたケース、[7] チャットボットが差別的発言を学習したケースなど[8]（どちらも象徴型）、様々な事態が報告されている。次章では、このような差別が発生する原因を分類し、既存の研究を紹介する。

3　箱庭の中の差別

本稿の冒頭に述べたとおり、ディープラーニングに代表されるような、大量のデータを用いた機械学習システムの中身は基本的にはブラックボックスであり、詳細な情報の流れを追うことは難しい。しかし、差別が学習される原因を定性的な二つの過程に分けることは可能である。これを手がかりに、「暗箱」の内側で何が起こっているのかを明らかにしていきたい。

3・1　差別が学習される原因

（1）学習アルゴリズム

瞬目検知システムにおいては、画像を処理するプロセスに問題が指摘された。しかし、実際にはすべてのケースにおいて、学習アルゴリズムが差別の学習を直接的にであれ間接的にであれ誘導している可能性は否定できない。システムを構築するプロセスには、設計者による入出力変数の設計やモデルの選択など、様々な恣意性が入り込む余地がある。前章の国勢調査の例は、そのような選択が差別的な学習を誘導する可能性を示している。さらに、多くの学習モデルは、学習に用いるデータから多くのサンプルに共通する特徴を抽出するように開発されている。それゆえ、少数のグループにのみ共有される特徴はノイズとして無視するような性質を持っている。このような性質は、少数グループに対する差別が学習されてしまう原因となる。

（2）学習に用いるデータ

前章の Amazon の採用システムがその例である。瞬目検知システムにおいても、偏ったデータセットが用いられていた可能性がある。偏ったデータが学習に用いられる場合、その原因は二通りある。

（a）サンプリングバイアス

ジェームズ・ズーらが二〇一八年に Nature 誌に寄せた記事[14]による

---

7 三人称表現に性差のないトルコ語において、「人」は「職業」であるという形式の文章が、例えば「He is a doctor」「She is a nurse」といった形で翻訳された。https://www.facebook.com/photo.php?fbid=10154851496086949&set=a

8 Microsoft のチャットボット"Tay"が、人種差別を支持する発言を学習した。https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

と、スタンフォード大学のデータセットであるImageNetに含まれる画像のうち、45％がアメリカから投稿されたものであった。アメリカの人口は世界の4％を占めるに過ぎないが、人口の36％を占めるインド、中国から投稿された画像はわずか3％であった。このようなデータの偏りが、あるアルゴリズムにおいて生じた、アメリカの結婚式の写真に「bride」「dress」「wedding」といったラベルを与える一方で北インドの結婚式の写真には「performance art」あるいは「costume」というラベル付けをしてしまう[10]、という事態を部分的に説明するとズーらは指摘する。また、自然言語のデータセットによく用いられるWikipediaの記述は、82％が男性によって行われている[1]。

（b）ラベル付けにおける問題

さらに、ズーらは学習データセットにおける「正解」のラベル付けが、大学院生やクラウドソーシングを通じて集められる人々によって行われることが多いという点を指摘する。この過程で、ジェンダー、文化、人種に関する偏りが生じている可能性がある。例えば、Amazon Mechanical Turkに仲介されたワーカーの75％がアメリカ人であることが分かっている[3]。

本節では、機械学習というブラックボックスの中に差別を再現した「箱庭」が生まれる原因を、内側から整理した。研究者達は、少しづつではあるがこの問題に取り組み始めている。次節では、そのような試みを簡単に紹介したい。

3・2　公平性を保証する学習

データに駆動された自律的な意思決定に差別が組み込まれてしまうことに対する懸念は、米政権らによる報告書でも繰り返し言及されてきた[5] [6]。

それに呼応する形で、機械学習の理論家たちは近年、公平な学習アルゴリズムを目指して、様々な理論を定式化している。[9] 以下に紹介するケースは、学習結果にある確率的な制約を持たせることを条件に加えている。以下に、「人材採用システム」を想定して、それぞれの定式化を整理する。

Demographic parity [7]

この概念が表すのは、差別を生む可能性がある属性を「センシティブ属性」として特定し、その「センシティブ属性」のすべての属性値と、システムの「出力ラベル」（採用か非採用か）が無関係（独立）であるように学習せよ、という要求である。数式で表すと、出力ラベルを$\hat{Y}$、値1と0をそれぞれ「採用」「非採用」とし、センシティブ属性Sの属性値$s, s' \in \Sigma$(Σは属性値の取りうる集合)としたとき、

$$\mathbb{P}(\hat{Y}=1|S=s)=\mathbb{P}(\hat{Y}=1|S=s') \quad (1)$$

が成立するように収束せよということになる。この制約は、アファーマティブ・アクションの実行とみなすことができる[9]。Demographic parity「データにバイアスが存在する」場合を想定し、不平等を是正するものだ。しかし、データに偏りがなく、センシティブ属性が採用に影響を与える

[9] 本節をまとめるにあたって第21回情報論的学習理論ワークショップ（IBIS 2018）における発表　福地一斗「公平性に配慮した学習とその理論的課題」を参考にした。

必然性がある場合には、ある属性値を持つ者を、別の属性値を持つ者よりも優遇してしまう可能性がある。

Equalized odds [7]

Equalized odds は、データにバイアスが存在しないことを仮定したモデルである。このモデルが与える制約は、ある属性の元に「多数派」と「少数派」が存在した時に、「多数派」のみに適合した学習が起きないようにするものである。データのラベル（採用されたか否か）を Y とし、Y は値 0 または 1 を取るとすると、

$$\mathbb{P}(\hat{Y}=1|Y=y,S=s)=\mathbb{P}(\hat{Y}=1|Y=y,S=s') \tag{2}$$

と表すことができる。

Lipschitz 連続性による制約 [4]

以上のモデルは、差別が生じる要因をある「属性」に特定して、確率的に対処するものであった。一方で、Lipschitz 連続性による制約は「似た個人は似た結果を得られるべきだ」という主張を定式化する。すなわち、「ある属性値の違いのみによっては、出力ラベルに違いが生じない」ように制約をかける。数学的には、入力データ上での距離を d、出力ラベル上での距離を D、入力データを出力ラベルに対応させる写像を M とした時に、

$$\forall x,y\in \text{入力データ},\quad D(Mx,My)\leq d(x,y) \tag{3}$$

と定式化される。現時点では、一般的な場合において数学的な最適性を満たす解を必ず発見する学習方法は見つかっていない。また、このモデルはデータの偏りによって生じる差別を助長する可能性もある。

以上に、機械学習の理論家によって提案されている、学習時に公平性を保証するためのモデルを挙げた。しかし、これらのモデルは、すべて二章の冒頭で示した分類における「配分型」の差別に対処するためのものである。すなわち、機械学習アルゴリズムが一般的に持つ、「多くのサンプルに共通する特徴を抽出する」あるいは「少数のサンプルにのみ現れる特徴は無視する」といった差別を助長しかねない性質を緩和することで、少数グループを軽視した「配分」が行われないようにするものである。

それでは、「象徴型」の差別にはどう向き合えばよいのだろうか。一見すると、「象徴型」の個々のケースにおいて、問題の解消のために設けるべきアルゴリズムやデータの制約は異なっているように思える。よって、この論点に対して意義のある提案を行うためには、「機械学習における差別」を、既存の「正義の政治・哲学」の理論を踏まえて包括的に議論する必要があると考えられる。「機械学習における差別」に固有の特徴は存在するのだろうか。また、この問題について考察することが、既存の理論に何か示唆を与えはしないだろうか。そこで次章は、「再配分」と「承認」を相互に還元できない概念として対置した理論家であるナンシー・フレイザーの論を手がかりに、「箱庭の中の差別」に対する是正策のための考察を行いたい。

4　暗箱の中の政治学

4・1　改善策の組み換え

機械学習システムにおいては、差別はその出力変数とシステムの使用環境に依存して、「再配分」、あるいは象徴的な差別、すなわち「誤承認」の

問題として我々のもとに現れる。そして、それらシステムはその使用目的の差異により他のシステムに対して閉じており、あるシステム内での変化はそのシステム内にしか影響を与えない。ゆえに状況の改善は、個々のシステムにおけるデータやアルゴリズムといった工学的構造、あるいはその背後に存在する社会経済的構造を問いに付して議論し続けることでしか達成されない。しかしながら、社会全体で問題に取り組むためには、このような試みを統合するための指針が必要である。

ナンシー・フレイザーは、「承認」と「再配分」を相互に還元することのできないカテゴリーであるとした上で、実践的な関心に基づきそれぞれに対する不正義を是正するための方策の統合を試みている。

それにも関わらず、事実上あらゆる事例で問題となっている毀損は、誤承認と不公正な配分との両方を含んでおり、それらは双方がどちらかの是正を介した間接的な策のみによっては完全に是正され得ないため、それぞれを独立して実践的な観点から考察しなければならない。したがって実践的には、事実上あらゆる事例の不正義を克服するためには再配分と承認の両方が必要である。[10]

フレイザーは、「承認」と「再配分」を統合しようと試みる際に、それらが相互に緊張を伴うものであることを主張する。

先に論じたように、不公正配分に対する完全に納得のゆく改善策は、それ単独で考えると、誤承認を悪化させうるし、逆に、誤承認に対する完全に納得のゆく改善策も、それ単独で考えた場合には、不公正配分を悪化させうる。そして、それぞれの次元では不正義を正すことに成功した個別の改革も、それらが合わせて追求されると、相互に傷つけ合う可能性がある。それゆえに、必要なのは、不公正配分と誤承認を同時に是正しうるような統合されたアプローチなのである。（[15]『再配分か承認か?: 政治・哲学論争』 p.100）

この緊張に対する注目は、機械学習システムにおける不正義を考える上で不可欠である。なぜなら、差別が問題となり得るシステムは、公の場において運用されるものであり、その上、是正を行うためには差別を引き起こしている属性的な要因を特定する必要がある。よって、何らかの是正策が検討された場合には、その仕様が公に知られることにより「誤承認」の拡大が生じる可能性が否定できない。

以上のような問題意識のもと、フレイザーは**統合の基本姿勢**として、「改善策の組み換え」を提案する。

その一つを私は改善策の組み換えと名付ける。このことが意味しているのは、正義のある次元に関わる改善策を正義の他の次元に係る不正義を正すために用いるということ、したがって、配分に関わる改善策を誤承認を是正するために使用し、不公正配分を是正するために承認に関わる改善策を使用することである。（同書 p.101）

フレイザーは、「改善策の組み換え」という戦略と同時に、さまざまな改革が集団の「境界」に与えうる影響に対して自覚的である必要性（境界戦

---

10 タナー講義（https://tannerlectures.utah.edu/_documents/a-to-z/f/Fraser98.pdf）を和文 [15]『再配分か承認か?: 政治・哲学論争』 p.30 を参考に訳したもの。

略）を訴える。これらの概念は、実質的な戦略を編みだす媒体となるものだとフレイザーは述べている。

次節において、これらの提案を踏まえて、「機械学習における差別」に取り組むための包括的な指針に向けた検討を行う。

4・2　暗箱のリバースエンジニアリング

先に述べたとおり、機械学習における差別の特異性は、個々のケースの「外面上」の独立性である。包括的な対処を行うためには、それぞれのケースにおける対策を統合するための指針が必要である。すでに見たように、差別の表れには不公平な配分「配分型」、承認に対する攻撃「象徴型」の二極があり、それぞれのケースにおいて、一見すると、それらの原因も取りうる対策も異なっているように見える。このような事態において問題に対する対策の指針を統合しようとするならば、フレイザーの着眼に倣うのは意義深い。すなわち、改善策を組み換える視点で、包括的な対策の指針を見出そうとする試みが有効であると考えられる。

「組み換える視点」とは、機械学習において何らかの「不公正な配分」が出力における「象徴的な差別」を生み、何らかの「象徴的な差別」が出力における「不公正な配分」を生み出す構造を特定しようとする視点のことであると定義する。この視点に従った指摘とは、例えば、二章の「瞬目検知」のケースにおいて、学習やテストに用いた顔のデータセットの偏りが製品において「象徴的な差別」を引き起こしているという可能性を指摘することである。また、「採用システム」において、学習されたネットワークの内部に特定の属性に対する「象徴的な差別」がエンコードされていると指摘することである。

このような分析は、公平性を確保するという観点において、出力を直接制御することのない指標を定義することを可能にするかもしれない。また、入力におけるある属性がどのような仕方で他の情報と結び付けられているのかを公平性の視点から分析することで、各属性が適切な承認を受けているのかを判別するための指標を得られる可能性がある。例えば、他の情報と組み合わせられることなく出力に影響を与えている属性は、なんらかの誤承認と結びついている可能性が高い。その偏りは、エントロピーなどの指標を用いることで数値化できると考えられる。このような統合されたアプローチは、その仕様が公開されたときの影響を考慮すると、差異を特定して是正するという単純な方策よりも優れている。そして、このような検討を重ねることで、個々のシステムが運用される社会経済的な背景が自ずと主題化されるだろう。すなわち、データセットが構築される過程はどのようなものであるのか、あるいは個々のシステムが資本主義の力場においてどのような影響力を持っているのかについての議論が促進される。

さらに、この「組み換える視点」を階層的に行使することで、例えば「不公正な配分」の原因となった「象徴的な差別」を引き起こしている「不公正な配分」を特定することができる。このような研究は、現に差別を学習したシステムに対して「リバース・エンジニアリング」を行うことということを意味する。[11]「箱庭の中」の秩序は、我々外部の社会のなかで生み出されたデータから学習される。よって、このような研究は「箱庭の外」[13]

[11] このようなリバースエンジニアリングの試みはすでに行われ始めている。https://ascii.jp/elem/000/001/584/1584663/ 等を参照

の社会に対しても何らかの示唆を与える可能性がある。

本節では、フレイザーの「改善策の組み換え」の視点に基づいて、機械学習における差別を是正するための包括的な戦略を立てる上で必要と考えられる現状の差別の分析方法の提案を試みた。ここでの検討は、あくまで抽象的な提案を行うにとどまっている。より具体的な考察を行うことは今後の課題である。

5 おわりに：「暗―箱庭」と共に

本稿を通じて、機械学習という「ブラックボックス」の中で生じる差別についてその内外から考察を行ってきた。本稿で検討した論点の他にも、議論すべきことはたくさんある。例えば、ある経済的あるいは政治的背景において、ある差別を自然化するために「AIによる意思決定」という標語が利用されることもあり得るだろう。しかし、本稿で行った提案に基づく議論は、そのような状況にも抗うことが可能になるような枠組みを整備していくためにも不可欠であると考えられる。哲学の対象としての社会はある種のブラックボックスであり、その内外に働く政治的な力場を整理するために、様々な議論が行われてきた。AIに代表されるような、機械学習を応用したシステムは、ある場合には我々の社会の何らかの構造を写し取る「箱庭」になる。それゆえ、AIの社会実装を考える上では、政治的な運動についてのあらゆる哲学が考慮される必要がある。本稿の考察はあくまで予備的なものに留まっており、具体的な指針やその有効性をはっきりと示すには至っていない。しかし、我々は現代の社会においてAIという「暗―箱庭」と共に生きていくしかないのであり、これは必要な議論であるのだ。

参考文献

[1] Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. Gender differences in wikipedia editing. pp. 11–14, 10 2011.

[2] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, Vol. 21, No. 2, pp. 277–292, Sep 2010.

[3] Djellel Eddine Difallah, Elena Filatova, and Panagiotis G. Ipeirotis. Demographics and dynamics of mechanical turk workers. In *WSDM*, 2018.

[4] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. Fairness through awareness, 2011.

[5] Executive Office of the President May 2014. BIG DATA:SEIZING OPPORTUNITIES, PRESERVING VALUES. https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, May 2016.

[6] Executive Office of the President May 2016. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. https://obamawhitehouse.archives.gov/sites/default/files/

microsites/ostp/2016_0504_data_discrimination.pdf, May 2016.

[7] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[8] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. Machine Bias, May 23 2016.

[9] Hussein Mouzannar, Mesrob I. Ohannessian, and Nathan Srebro. From fair decision making to social equality, 2018.

[10] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world, 2017.

[11] Sidney Fussell. AI Professor Details Real-World Dangers of Algorithm Bias [Corrected] , 08/17 2012.

[12] Latanya Sweeney. Discrimination in online ad delivery. *CoRR*, Vol. abs/1301.6822, , 2013.

[13] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society - AIES ゛18, 2018.

[14] James Zou and Londa Schiebinger. Ai can be sexist and racist ― it゛s time to make it fair. *Nature*, Vol. 559, pp. 324–326, 07 2018.

[15] Nancy Fraser Axel Honneth 加藤泰史 高畑祐人 菊地夏野 舟場保之 中村修一 遠藤寿一 直江清隆『再配分か承認か?: 政治・哲学論争』法政大学出版局 叢書・ウニベルシタス 二〇一二.