アルゴリズムの判断はいつ差別になるのか ── COMPAS 事例を参照して

前田春香（東京大学大学院、理研 AIP センター）

要旨

本論文の目的は、Correctional Offender Management Profiling for Alternative Sanctions（以下COMPAS）事例においてアルゴリズムが人間と似た方法で差別ができると示すことにある。技術発展とともにアルゴリズムによる差別の事例が増加しているが、何を根拠に差別だといえるかは明らかではない。今回使用する COMPAS 事例は、人種間格差が問題になっているにもかかわらず、そのアルゴリズムが公平であるかどうかについて未だ論争的な事例であり、さらには差別の観点からは説明されていない。本論文では、「どのような差異の取り扱いが間違っているのか」を説明する差別の規範理論を使って COMPAS 事例を分析する。より具体的には、差別の規範理論の中から「ふるまい」による不正さを指摘する Hellman 説を適切なものとして選び、アルゴリズムに適用できるよう改良したうえで COMPAS 事例が差別的であるかどうか分析をおこなう。この作業によって、差別的行為を「ふるまい」の問題として独立させ、一見差別の理論が適用できなさそうなアルゴリズムによる差別性の指摘が可能になる。

Abstract

This paper aims to explain how algorithms can discriminate wrongly against humans in a human-like way. With technological developments, discrimination by algorithmic systems has become an issue. However, it is not clear what the wrongness of discrimination by algorithms is. It is also true for a famous case of Correctional Offender Management Profiling for Alternative Sanctions (hereafter COMPAS). The case has been controversial because of the fairness of the algorithm. Besides, the question of whether it is morally wrong discrimination or not remains to be seen. In this paper, I will give an analysis of the case of COMPAS by focusing on the point of whether this algorithm discriminates in a wrong way or not. First, More concretely, I will choose Hellman's account, which can detect the wrongness of discrimination based on the behavior of the discriminator, as an appropriate one for algorithmic discrimination. Secondly, Hellman's account assumes that humans as a discriminator will be improved in terms of power for this account to apply to algorithms. After that, I will offer an explanation of the case of COMPAS from an ethical viewpoint. This paper may provide a way to explain the wrongness of discrimination by algorithms based on their behavior even if it seems to be impossible for a theory of discrimination to apply to algorithms.

はじめに

本論文の目的は、COMPAS 事例を用いて、アルゴリズムが差別を擬似的に ――つまり人に似たやり方で ――行えると示すことにある。

2008 年、Hewlett-Packard（以下 HP 社）のノートパソコンの画像認識プログラムが黒人を認識しないという形で「差別」したとして、投稿者に批判の意図はなかったにもかかわらず、YouTube に投稿された動画[1]が炎上した。HP 社は自社ホームページに掲載した謝罪で、「照明の問題だ」と説明した（Hing 2009）。つまり、これは技術的な問題にすぎず、差別の問題ではないのだというわけだ。

以来、アルゴリズム[2]によって人間が差別されている（とされる）事例は増え続けている。Google の検索エンジンは黒人女性の問題のある表象を提供し（Noble 2016）、Google アドセンス広告は黒人らしい名前には、白人らしい名前より高い頻度で、逮捕の経歴を示唆する広告を返す（Sweeney 2007）。今では雇用やローンといった、より社会的影響が大きい分野においてもアルゴリズムは重大な問題を引き起こし、人々の間の格差を広げているという（O'Neil 2016）。このような偏ったアルゴリズムによる問題全般を指して、アルゴリズムの偏見（algorithmic bias）と呼ばれている（Koene et al. 2018）。

アルゴリズムによってもたらされるこのような現象は、直観のレベルで全く差別的でないと言い切ることは難しいだろう。実際にそのような現象が炎上によって知られたり、その後も話題にされたりしていることは、われわれがこの問題を不正なものと認識していることの証左であるといえよう。しかしさらに一歩踏み込んだ理論的なレベルでは、どのような偏りが倫理的に問題があるのか、つまり道徳的に不正な差別とよべるのかは説明がなされていない。実在の事例は多々あるものの、個別の事例の分析は不十分であることが指摘されている（Sandvig et al. 2016）。本論文では、アルゴリズムによって提示されるどのような偏りが差別的であるといえるのかを検討するために、Correctional Offender Management Profiling for Alternative Sanctions（以下 COMPAS）の事例を分析する。

COMPAS とは、インプットデータに質問紙調査の結果を使い、ブートストラップ法や生存率曲線を用いて再犯リスクを診断する統計的自動判断プログラムである（Brennan et al. 2009）。2016 年 5 月、公益を目的とした独立系報道機関である Propublica が、当該プログラムの判断には人種間格差があると告発し、大きな議論を巻き起こした。当該事例を参照する理由は二つある。第一に、COMPAS が偏ったアルゴリズムの典型と目されていることが挙げられる。実際にアルゴリズムの出力は特定の集団に対して偏っており、さらに当該の偏りは数理的に根拠があるものである。かといって、数理的な根拠があれば差別して良いのだろうか。このように、当該の事例はそもそも公平、ひいては差別とは何か、という問題に関連するものだといえる。次に、COMPAS に関する論争は、調査者側は偽陽性と偽陰性に基づく不公平を主張したのに対して、制作側はキャリブレーションを元に批判に応答した（Dressel and Farid 2018）。つまり制作側は、問題になっているリ

1　https://www.youtube.com/watch?v=t4DT3tQqgRM
2　本論文におけるアルゴリズムとは、一定のタスクを遂行する際の計算式をいう。本論文では、アルゴリズム、プログラム、ソフトウェアは区別せずに扱う。

スクスコアが同一の両人種は、それぞれが同じ割合で再犯しているため、人種差が存在せず、問題ないという認識を示したというわけである。この二つの基準はどちらも数理的な公平性基準であるが、同時に満たすことができないことが数学的に明らかになっている（Kleinberg et al. 2017）。このことは、数理的なこの二つの基準と、道徳的に問題がある差別の基準との関係を探究する必要性を示唆する。

このような問いを扱う理論が、差別の規範理論である。差別の規範理論では、異なる処遇のうちどのようなものが道徳的に不正であるかを論ずる。例えば、性別によってトイレを分けることは男女のどちらも差別していない。その一方で、性別によって雇用の処遇を変えることはおそらく差別の誹りを免れないであろう。これら二つを分かつものは何だろうか。

本論文ではまず、差別の規範理論の中でどのような説がアルゴリズムによる差別に適しているのかを確認した後、COMPAS 事例の検討に移る。先取り的に言ってしまうと、Hellman 説の適切さを示した後に、一見人間にしか満たせないような要件をアルゴリズムも満たすことができると示し、修正した Hellman 説を COMPAS 事例に適用する。

この作業によって、以下のような意義が見込まれる。これまでのアルゴリズムによる差別の研究では、格差の拡大など社会に及ぼす影響をもって悪を判断する説明が主であった（Sandvig et al. 2016）。本論文では、判断そのものを不正なものと考えられる基準を提供するとともに、その適用例を示すことで、より早期にその差別性を察知することを目指す。というのは、現状ではアルゴリズムの判断が引き起こす帰結としての格差が観察されてから対処することになるが、ふるまいによって不正であるかどうかが検討できるならば、より早い対処が可能になるからである。また、COMPAS 事例は有名ではあるが、不正な差別であるかどうかは未だ明確にされていない。不正さの根拠を示すことによって、他の事例への適用への道が開かれる。

1. 事例の説明とこれまで付されてきた説明

本節では、COMPAS 事例の分析に先立ち、事例の概要を提示し、いかなる道徳的不正さに関係する解釈が提供されてきたかを参照する。当該の解釈には正確性や個人の尊重原理、公平性基準といったものがあるが、COMPAS の場合にはいずれも不十分なものであるため、新たに差別の不正さにかんする道具立てが必要であることを示す。

1）COMPAS 事例の概要

まずは、COMPAS による事例がどんなものであるかを確認しておこう。

2016 年 5 月、Propublica は “Machine Bias” という記事[3] で、COMPAS にまつわる以下の問題を告発した。その報告によれば、被告に対して実際に再犯したかどうかを追跡調査したところ、黒人[4] は実際には再犯していないのに高リスク群としてラベル付けされた頻度が白人の 2 倍であり、

3　当該記事には、再犯リスクが低いと判断された白人男性のインタビューが掲載されている。その白人男性は万引で逮捕され、犯罪履歴には加害暴行や複数回にわたる窃盗、重罪である麻薬密売があったが、再犯リスクが低いと判断された。インタビューで彼は、提示されたリスクが「あまりにも低いので驚いた」と述べている。その後一年以内に、彼は 1000 ドルの万引を含む二つの重罪によって再び告発された（Angwin et al. 2016: pr. 42-44）。

4　ここにおける「黒人」「白人」とは、保安官事務所に保管される書類上に記載されているものを指している。

白人が実際には再犯しているのに低リスク群としてラベル付けされる頻度は黒人よりも高かったという（Angwin et al. 2016）。つまり、全体として黒人が不利に判断されているというのだ。

具体的な調査内容は以下の通りである。COMPAS と同様のアルゴリズムを入手し、フロリダ州ブロワード郡で 2013 年と 2014 年に 7000 人以上に割り当てられた再犯に関するリスクスコアを算出、さらにその再犯予測が現実と合致しているかどうか、その後の 2 年間にわたり追跡調査を行った（Angwin et al. 2016）。一般犯罪の再犯予測が合致している確率は 61%、暴力的な再犯予測が合致している確率はおよそ 20%だった（Larson et al. 2016）。さらにリスクと照合する形で人種や年齢といった要素のロジスティック回帰分析が行われており、当該の分析によって高リスク判定と属性に関連性があることが明らかになった[5]。

この分析結果は、結果を提供した Propublica と制作企業 Northpointe 社の論争を招いた。Angwin ら Propublica 側にとって、格差を評価する基準となっているのは偽陽性と偽陰性の割合が同程度でないことである。つまり、黒人は本来よりも「危険」であると判断される可能性が高く、白人は本来よりも「危険でない」とされる可能性が高い。一方で Northpointe 社は、最も標準的に用いられる公平性の別の基準であるキャリブレーションをもって反論した。反論の根拠としては、第一に再犯予測は高リスク群の黒人にも白人にも同様の正確性を有していること（predictive parity）、第二に黒人でも白人でも再犯者と非再犯者を同様に見分けることができること（accuracy parity）、第三に再犯可能性を算出するために与えられたスコアに人種差がないこと（calibration）が挙げられている（Dieterich et al. 2016）。言い換えれば、Northpointe 社は、どちらの人種に対しても COMPAS は再犯するかどうかを同程度正確に予測できる、したがってキャリブレーションの条件を満たしており公平である、というのだ。実際、Sumpter（2018=2019:82）の検証によれば、ハイリスクを割り当てられた黒人のうち 6 割が再犯し、同様にハイリスクを割り当てられた白人のうち 6 割が再犯しているといい、確かにキャリブレーションの観点からは問題がない。

現在もこの問題は多数の論者を巻き込んで、その予測がどれほど公平であるかどうかなどについて論争が続けられており、未だ決着をみていない。

2）「公平さ」の評価はうまくいっているか

以上では COMPAS という事例の特徴と、その特徴が公平性の基準と関連していることを指摘した。本項では、公平性の基準がうまく差別を指摘できていない上、一般に言及される個人の尊重原理によってもここでの道徳的問題を捉えることができず、したがって別の基準が必要であることを示す。

先述の論争は、どちらかが本当のことを言っていて、どちらかが誤っているというものではない。このように対立する根本的理由は、Propublica が挙げた基準である分類均一性（classification parity）と、Northpointe 社が挙げた基準であるキャリブレーション（calibration）の基準を両立することができないことにある（Kleinberg et al. 2017）。少なくとも今回のケースでは、数学的に差別であるかどうか、公平であるかどうかを判断することはできないのだ。そしてそう考えるならば、Propublica と Northpointe 社の論争は自然なものであったとすら言えるだろう。

5　詳しくは Larson ら（2016）を参照。

もちろん実効性の観点からして、ある公平性基準をみたせないならば差別としてもよいのではないか、という反論がありうる。しかし、どちらの公平性基準も固有の限界を有すると指摘されている（Goel et al. 2018）。分類均一性については、偽陽性率はグループ全体の再犯率に応じて機械的に増加する傾向が指摘されており、当該の傾向は infra-marginality 問題として知られている（Ayres 2002 など）。つまり、再犯率が実際に白人よりも黒人の方が高い場合に、当然の帰結として、黒人の方が再犯率が高いと判定することになるのである。これが差別であるといってよいかは疑わしい。一方でキャリブレーションについては、この基準を満たしたまま差別を行うことが可能である（Corbett-davis and Goel 2018）。例えばローン貸与の文脈で、白人と黒人で同程度のデフォルト率を想定し、住所とデフォルト率に関連性をもたせたとする。しかし住んでいる地域と経済的水準、そして経済的水準と人種に密接な関係性がある場合には、合法的にこの方法でハイリスクな人種を排除することができる[6]（Corbett-davis and Goel 2018: 1）。このことによって、ハイリスクとされる人種はローンを供与されないことになる。結果として、さらなる格差につながりかねない。Binns（2018）によって公平を数理的に定義することが一般に困難だと指摘されているように、差別も数理的に定義することは難しいのである。

ならば次の問いは、どのように差別を特徴づけすることが適切なのかというものになる。どのような集団も異なっていて当然の中、われわれはある差異の抜き出し方が間違っていて、差別的であると感じることがある。では特定の仕方で間違っていると何に基づいて主張できるのだろうか。

一つの仕方は、法律上保障されるべき人権に基づいて説明することだろう。COMPAS の事例では、このような自動判断プログラムを量刑判断に使用することは人権の一つであるデュープロセスの権利を侵害しているとして、実際の被告である Eric Loomis によって裁判がおこされている。被告の反論内容は、1. COMPAS のアルゴリズムの内容が不明であることは、正確な情報に基づいて判断されるという被告の権利の侵害である、2. 個別に判断される権利を侵害している、3. 評価に不適切なジェンダー差がある、の三つであった（State *v.* Loomis）。

これらの論点は正確性・信頼性に関するもの（1 と 3）と個人を尊重しているか否か（2）に分けることができる。州最高裁はこの請求に対して、確かにジェンダーによって異なる尺度が用いられていたことを認めるものの、その異なる尺度はプログラムが非公開であるため[7]検証できず、ただちに不適切だとは判断しかねるとする（State *v.* Loomis）。もっとも、州最高裁もプログラムが不正確である可能性や偏見については指摘しているが、当該プログラムが差別的な影響力を持つかどうかについてまでは明確な判断を行っていない（山本・尾崎 2018）。

仮に正確であったとしても、まだ残っている問題がある。それは、自動判断プログラムを使用して個人の量刑判断を行うことは、本人を尊重しているとはいえない、というものである。

これは COMPAS のみならず、統計的手法を用いて個人の評価を行うツールに共通する問題である。個人評価にツールを使用することは個人を尊重しているとはいえないのだろうか。そもそも人を尊重して判断するということはどういう行為であり、それは可能なのだろうか。Schauer

6　当該の慣行はレッドライニングとして知られている。
7　ここでいう「内容が不明」とは、プログラムの中身が技術的にどうなっているかわからないといういわゆるブラックボックス性を指すのではなく、当該プログラムが企業の販売品、すなわち営業秘密として保護の対象となることを指している。

(2003) は、個人の尊重をあらゆる不完全な一般化に基づいた取り扱いを排除するものとして理解する場合、そもそも個人を尊重して判断することは不可能であるという。一見個々人に向き合った取り扱いでも、人間はある程度の統計的一般化を行いながら日常生活を遂行しているため、究極的には統計的一般化に基づいているといえるからだ。

確かに、ツールを使用して判断するとき、ツールで計測できない変数は棄却される。例えば被告が酩酊状態だったことや、生活できないほど貧困だったことなどが考慮できないかもしれない。しかし、ツールや感情に流されずに判断するために裁判官がいるのであり、最高裁もツールに全面的に依拠しないよう注意喚起をしている[8]。以上の理由から、プログラムを量刑判断に使用することは合憲であるとされ、この請求は棄却されている。付け加えるならば、COMPAS は、再犯リスクの説明変数として、例えば犯罪心理学で犯罪と関連すると明らかにされている変数を計測する 137 項目をリスク判定の材料としている（Propublica 2016）。これは、少なくとも人間が行う慣行であり人種だけを根拠とする人種プロファイリング[9]よりは、個人をより多角的に判断しているということもできるだろう。実際に Loomis の「ジェンダーが不適切に考慮された」との主張に対して、ジェンダーのみに基づいて量刑判断が行われたわけではなく、量刑判断において多角的な判断がなされていたこと、現在告訴されている犯罪の内容と被告人の犯歴が重視されていたと裁判所は判示している。このことは、裁判所が COMPAS が多角的に当該個人を考慮しており、個人の尊重原理に反するものではないと考えていることを裏付ける（山本・尾崎 2018）。

以上より、COMPAS の道徳的問題の分析のためには、正確性、もしくは個人の尊重原理からのアプローチは不十分である。しかし「自動判断プログラムを量刑判断に使用することはデュープロセスの権利の侵害ではない、したがって米国憲法に定める平等違反ではない」というこの判断には法学の立場から多くの批判がある（Harvard Law Review 2017; Beriain 2018; Freeman 2016; Liu et al. 2019; Washington 2019）。本稿ではこれらとは別の角度から、すなわち道徳的に不正な差別かどうかという観点から批判を加えたい。

第一に、法学的に規定される差別と、道徳的に不正とされる差別が侵害する権利には違いがある。後者のほうがより広い範囲を包含しているのだ。ここでいう法学的に規定される差別は、米国でいう米国市民権法第七編であり、日本での男女雇用機会均等法でいう差別をさす。両者が禁じている間接差別とは、一見中立的な処遇によって属性によって偏った不利な影響を生じさせるような取り扱いのことである。この偏った不利益は明確な比較対象を持って算出する必要があり（Khaitan 2019）、影響が明らかでなく計算できない場合には効果が薄いと考えられる。

第二に、これがより重要な理由であるが、法学的な議論では、当該プログラムの差別性については結局のところ判断されていないということである（山本・尾崎 2018）。裁判所は確かにプログラムを使うことに対して注意喚起を行ったが、これはプログラムを使う人間だけに着目したものであり、プログラムの影響を無視している。とはいえ、判断を効率的もしくは正確にするような影響がないならばプログラムを使用する意味はない。人がプログラムを使用することを原則とし

8　これは GDPR が掲げる、人間は自動判断だけで判断されない権利を有し、非倫理的な判断を防止するために人間を配置する必要があるという方針と一致する。
9　ここでは、人種を根拠にして街頭でのパトロールで聞き取りをするかや捜査をするかどうかを決めることを指す。詳細は後述。

ながら、いかに共同の判断を「よく」すればよいか、すなわちその使い方がここでの問題なのだ。ならば、まずはプログラムがアウトプットするどのような判断が道徳的に許容可能かという問いに答える必要がある。

2. 道徳的側面からの分析

以上では、COMPAS の道徳的問題の有無を解釈するには新たな道具立てが必要であることを述べた。本節では、新たな道具立てとして差別の規範理論が適切であることを示す。次に、差別の規範理論の説の中で適切なものを選び、当該の説を詳細に説明する。

1）規範理論からの検討

どのような判断が道徳的に許容可能であり、どのようなデータの偏りが許容できるのか、その問いに答えるために使用するのが差別の規範理論である。差別の規範理論とは、差別がなぜ不正なのかを探求する哲学的理論である。

差別の不正さの由来には、大きく二つの説がある（堀田 2014）。一つは、被差別者に与える害を基準とする害ベース説、もう一つは基準を尊厳の侵害に求める尊厳ベース説である。この二つは絶対的原理（bedrock principle）として採用しているものによって分けられているため、害ベース説と尊厳ベースが同じ基準を含みうる。例えば、害ベース説でも個人の尊重に言及することはあるし（Knight 2013）、全ての尊厳ベース説が害に無頓着であるというわけでもない（Hellman 2008=2018: 3）。

ではどちらの説がアルゴリズムによる差別に照らしてより適切なのだろうか。尊厳ベース説ではアルゴリズムが社会に及ぼした影響や不利益を不問にしたまま、当該の判断が社会的文脈に照らしてどうか、を判定する。すなわち、仮に実際のサービスに何ら影響を与えなかったとしても、研究室の中で画像認識プログラムが肌色の暗い人を認識しなかった場合には、やはり差別的だと考えることができるだろう。尊厳ベース説ならばこのような含意をくみ取ることができる。

もちろん、害ベース説でもその悪質さを説明・予見することはこの COMPAS 事例においても可能である。人種によって不均等なリスク判定はそのまま、例えば不均等な量刑判断を導きうるし、それはすなわち害に直結するだろう。人種プロファイリングにおいても、害を予見することによってその悪質さを批判することができる（Eidelson 2015）。アルゴリズムが人間の意思決定を真似るものである以上、このような司法だけではないあらゆる場面で、特定の人種が不利に扱われることが予見されるため、予防もできるかもしれない。

しかし、尊厳ベース説にも害ベース説に比して、無意識的な差別や構造的差別を扱えるという利点がある（Lippert-Rasmussen 2006: 21）。この利点はアルゴリズムによる差別においてはより重要になりうる。なぜならアルゴリズムの使用は、われわれの差別意識を顕在化させたり、もしくは差別の件数を増加させたり、その効力を増幅させる可能性があるためである。あるいは、差別であるとわれわれが気付かないケースの可能性もありうる。例えば、われわれ全員が属性 B よりも属性 A が優れていると信じ切っている場合、差別的待遇を不正なものとして受け取らないかもしれない。つまり、害をいかに認識するかについての問題がある。人間がどう考えるかがアル

ゴリズムによって再現され、それに基づく判断が実行されれば、もはや人々がどう思っているか、その評価にたいしてどう思うかは重要でなくなる。アルゴリズムによる差別は、構造的な差別の一部なのである。

2）いかなる説が適切か

尊厳ベース説の中でも、何を根拠にして尊厳を侵害していると判断するかは諸説ある。次は、アルゴリズムによる差別においてどれが最も有効な基準なのかを特定しなければならない。

まず、Alexender（1992）は、行為者の偏見や悪意によって差別行為の不正さを指摘する心的状態説を提唱しているが、これはあまり望ましくない。悪しき心的状態になくても、可能な差別的行為が統計的差別である。統計的差別[10]とは、多くの人々から特定の目的に沿った人を選出しようとするときに、特定の目的を十分に代替すると思われる属性を持つ人々を選ぶことが集団間に異なった影響をもたらす場合に使用される言葉である（Arrow 1971; Phelps 1972 など）。この場合、必ずしも選出者は悪しき心的状態によって特定の属性を選んでいるわけではない。

次に、行為の表現内容によって尊厳を侵害しているかどうかを指摘できる表現説がある（Hellman 2008=2018; Scanlon 2008）。例えば、ジェスチャーによって侮辱を表現することには問題があると解される。両者の説には意図を考慮するか、考慮しないかで重要な違いがあり、その点でもっともラディカルになりうる説といえる。

最後に、心的状態を考慮に入れるものの、差別者の行為や態度を総合的に斟酌するものとして熟慮説がある（Eidelson 2015）。

アルゴリズムによる差別には、心的状態によってその不正さを問うことは不可能である（Binns 2019）。このことに照らせば、もっともアルゴリズムによる差別の不正さを問題化するためにふさわしいのは表現説、とりわけ差別者の心的状態を考慮しない Hellman 説であるということになる。

3）Hellman 説の詳細

Hellman 説における不正な差別が、個人の尊厳を尊重しないようなものであることは先述した。Hellman は当該の不正な差別を貶価と呼んで、その成立条件を二つに分けている。一つは被差別者の尊厳を毀損するような表現内容がなされるという表現条件、もう一つは当該表現を押し付ける権力性を意味するヒエラルキー条件である。ただ個人を侮辱するだけでなく、侮辱によって表現された劣等性を押し付けることによってはじめて貶価になるのだ。

これでは人前で上司が部下を侮辱することと差別行為の区別がつかないので、もう少し説明が必要である。上記の二条件を満たすような上司が部下を侮辱する場合であったとしても、その侮辱の理由が国籍や人種、性別といった攻撃を受けている人が所属しているカテゴリーに関連したものであれば、われわれには差別的だと感じられるであろう。さらに、その上司は当該カテゴリーに所属していない人に対しては侮辱を行わず、対等に接するものとする。この場合、侮辱の理由は当該カテゴリーに由来しているものとみることができ、そのカテゴリーに属しているかいない

10　本来 Arrow や Phelps がさす統計的差別には、雇用者と被雇用者の選択が合致してしまうことにより状況が固定化されることを問題視するが、ここでは本文中にあるような意味で取り扱う。

かによって、上司は差異をつけた処遇をしているとみられる。このような差異処遇は道徳的に不正だといえるだろう。

両者は何が違うのだろうか。Hellman によれば、当該の属性の背負ってきた歴史を含む社会的状況によって、当該の属性を基準に分類することそのものが不当になる可能性をはらむ（Hellman 2008=2018:35-42）。ここで着目したいのは、「分類することそのもの」という言葉である。上の例でいえば、上司が部下を属性によって侮辱する場合に、当該属性によって不利な状況が生じるからではなく、当該の差異処遇 ── ここでは侮辱すること自体が、道徳的に不正になりうるというのだ。

とはいえ、侮辱自体が道徳的に問題であることに疑問の余地はない。上司が人前で部下を侮辱しているとき、その侮辱が人種に基づいたものであろうとなかろうと、その行為は同様に不正である。では誰に対してなされても一般的に行為自体が不正であるとは考えられないような、次のような事例を考えよう。校長がバスの前方座席に白人生徒を座らせ、後部座席に黒人生徒を座らせることには道徳的に問題があるだろうか（Hellman 2008=2018:37）。Hellman によれば、この差異処遇には問題がある。なぜなら、私たちの[11]文化的理解では、公共交通機関の座席を人種ごとに区分することは劣等性を意味すると考えられるからである（Hellman 2008=2018:37）。しかも、人種によって人々を分離することは特定の人種の劣等性を表現するために、これまでも行われてきたことである。また、校長の命令は一般的な生徒にとって逆らえるようなものではない。当該の命令は、仮に被差別者がその意図に気づかず、差別されたことによる害を感じなかったとしても道徳的に問題があると考えることができる。

したがって、当該の表現行為が不正であるかどうかは表現内容を検証し、強制力を持つかどうかで判定すればよい。その基準になるのは、当該行為をおこなうもの、当該行為がおこなわれた文脈、当該行為において用いられる言葉である（Hellman 2008=2018:90）。これらはそれぞれ、話者がヒエラルキー条件、文脈と用いられる言葉が表現条件に大まかに対応しているとみることができるだろう。表現の内容の検証には、当該の表現がどれほど特定の属性の劣等性を示す典型的な手段と類似しているか検討することも含まれる（Hellman 2008=2018:61）。これを以下では慣習条件と呼ぶ。

アルゴリズムもこの条件を人間と同様にみたせるのであろうか。文脈と用いられる言葉が問題になりうることはアルゴリズムでも同様といえよう。実際、われわれは多くのアルゴリズムによる差別的な表現を道徳的に問題があるものとみなしてきた。Google Photo のアプリが肌の色が暗い人にゴリラとタグ付けしたことで炎上し（Hing 2009）、Tay によるヘイトスピーチもわれわれは確かに問題であると感じるのはその典型的な例だ。しかし、行為者についてはどうであろうか。アルゴリズムは一般に道徳的行為者たりえず、その立場を人間と同一視することはできない。それにもかかわらず、Hellman 説の図式において、アルゴリズムは人間同様に道徳的問題をはらんでいると言えるのだろうか。

11　ここでいう私たちの文化とはどこの文化なのかは疑問の余地がある。もっとも狭くとらえるならば Hellman のいるアメリカの文化ということになるだろう。ただし、日本においても歴史の教科書でこの状況がよく知られていると考えるとき、同じように人種による座席の分離が劣等性を意味すると考えられても不思議はない。

3. 理論および COMPAS 事例の批判的検討

2 節では、本論文で使用する説について説明した上で、依然として課題が残っていることを示した。その課題は道徳的行為者性に関係するものである。本節では、Hellman 説の図式においてアルゴリズムが行っていることが差別であると言えるかについて、理論的側面の検討、および COMPAS 事例の分析を通して検討をおこなう。

1）Hellman 説の理論的検討 ── ヒエラルキー条件

2 節の最後では、アルゴリズムが道徳的行為者でないために、差別行為を行うことができず、したがって Hellman 説の問題にならない可能性に言及した。この問題を追及するために、以下では差異処遇に話を戻して、まず Hellman 説の 2 条件のうち、当該の問題はヒエラルキー条件の問題であることを指摘する。次に、アルゴリズムがヒエラルキー条件をみたすかどうか理論的検討をおこなう。

差異処遇は、必ずしも人間が行うわけではない。バスの例でいえば、特定の人物が分かれて座るよう人間が直接言語をもって命令している場合ばかりではないだろう。他にも、間接的な命令にあたる事例として、法制度や政策などが考えられる。これらの共通点は、人間に指示内容を受け入れさせるような強制力を持ち、人々を一定の方角へ方向付けることである。この 2 つは Hellman の説の条件のうち、ヒエラルキー条件が取り扱うものである。

Hellman は、最もわかりやすいところでは、話者が権力を持っていることは強制性を構成する大きな内容になりうると記述している（Hellman 2008=2018: 91-94）。例えば、先述の上司が部下を侮辱する例がそれにあたるだろう。侮辱を通じて表現された部下の劣等性は、第三者にそれを受け入れさせる契機となりうる。次に、命令者が厳密には人間ではない事例であるが、法律や政策もその強制性が大きいと考えることができる。アパルトヘイトは、たとえ完全には法律が守られなかったとしても、人種によってエリアを分けるよう人々に命令する [12]。

こう考えるならば、人間でなくても強制力を持ちうることになる。このことがここでは重要である。われわれは、直接明示的に命令を下されなくとも一定の方向に向かって動くことがある。チェスやスポーツをするとき、われわれは反則をせず、なおかつ勝つように方向付けられる。このような仕組みを強制することができるならば、スポーツなどの自発的に何かをするという局面から脱して、われわれが否応なく営む日常生活のような局面でもその効力を発揮し得る。

では、アルゴリズムはそのような強制性をもつといえるだろうか。物質的な空間でも、例えば入場・出場がアルゴリズムによって管理されている場合にはあてはまるだろう。駅に設置されている自動改札機は、電子マネーの残高が十分であるかどうかによってふるまいを変え、残高が十分でない場合にはそのゲートを閉じる。この場合、一定の方向付けがプログラムによってなされているのだ。もし仮に、この基準が「十分に電子マネーの残高が残っているか」ではなくて人種であったとしたら、われわれはそれを差別的だと感じるはずであるし、当該の場合でも Hellman

12 次に問題になるのは、第三者が存在しない場合にも Hellman 説が成立するかということである（石田 2019: 4-5）。人間が行為している場合には成立しうるが、アルゴリズムが判断している場合に成立するか否かは議論の余地がある。

説は効力を発揮するだろう。われわれはそうと意識しない場合でも、日常生活を営む上で強制性にしたがっているのだ。ローンを貸すかどうかアルゴリズムだけが判断するなどの場合にも、同様の理由で強制性が発揮されているといえる。

無論、これだけでは（特に今回の事例において）アルゴリズムが強制性を持っていると述べるには十分ではない。自動改札機はアルゴリズムがわれわれを物質世界において従わせるもっともシンプルな例であるし、まして現実における多くの事例ではアルゴリズムだけが判断を下しているわけではない。より重要なのは、アルゴリズムが「真実」を産出し、われわれに当該の「真実」を信じさせることで、一定の方向付けをおこなうという機能の方である。

アルゴリズムは、いかに当該人物が扱われるかを決定するだけでなく、いかに扱われうるかという範囲を狭める役割を担っている（Beer 2017: 8-9）。一定の方向付けがどれほど、そしてどのように決定的となっているのかが強制力との関連では重要である。Beer は、真実の産出を行う方法について二種類を挙げている。

第一に、人間に選択肢を直接提示することによるものである（Beer 2017: 12-13）。アルゴリズムが人間の取れる行動範囲を提案するとしよう。結果、アルゴリズムのアウトプットは、人間を通して真実になっていくのだ。Google が私たちのウェブ上での行動をトラッキングし、次に買うもの、見るものの範囲をコントロールすることで、アルゴリズムが挙げた選択肢が後に現実に存在するようになってゆくのである。もちろん、人間にはレコメンデーションに従わない自由もあるが、本論で述べる強制力を有することがここで確認できるだろう。レコメンデーションとならぶ同様の方法として、ナッジを挙げることも可能である（cf. Yeung 2016）。

第二のものは、アルゴリズムが提示する規範に従わせる方法である（Beer 2017: 13）。Wii スポーツがわれわれの運動データを参照し、健康年齢を算出するとき、当該の健康年齢は無論実際の年齢ではなく、運動データと関連付けられたいわば架空の数値である（Chenny-Lippold 2017=2018: 150）。しかし、おそらく実際の年齢より若ければ喜ぶ人も多いだろう。この場合直接行動をサジェストされるわけではないが、それにもかかわらずわれわれの行動は、望ましい方へ、つまりここでは健康へ向かうように、一定の方向付けを得る。このはたらきは規範に自らを添わせようとする権力の働きの一種なのだ（Chenny-Lippold 2017=2018: 151-152）。

この場合、アルゴリズムが規範を提示するといっても、何もないところからそれが生み出されているわけではない。あくまで、現実に存在する価値と結びついて、アルゴリズムがその権力を行使しているのだ。そのような事情から、第二のアルゴリズムの方向付けの説明には、Foucault の知と権力の関係が参照されることもある（Rouvroy and Berns 2013; Introna 2016; Beer 2017 など）。Foucault は知と権力の関係について、権力は知識や、知識を運用する機構なくしては機能しないものと述べている（Foucault 2004: 34）。アルゴリズムは知を運用する機構となっているのであり、この場合の知とは、すなわちわれわれについての知識である（Chenny-Lippold 2017=2018: 55）。アルゴリズムは観念を具体化し、人間は割り当てられた観念、すなわち自分に対するアルゴリズムの解釈を認識することによって新たな行動を生み出すのだ[13]。こうしてアルゴリズムそのも

13　この観念を割り当てるはたらきがどう結果として表れるかは当然、アルゴリズムのアウトプットによって異なる。本人の認識を通じて本人の行動を変えることもあるが、より重要かつ潜在的な効果として、資源の配分を管理することがあることが指摘されている（Chenny-Lippold 2017=2018: 200）。

のが価値観を指し示すものとなるのである（Beer 2017: 14）。

Ziewitz（2016）が指摘するように、このようなアルゴリズムの働きは、決定的な影響力をもつものとして捉えられている。コンピュータの言語は、ただ世界に存在するにとどまらず、「世界を創る」ものでもあるのだ（Introna 2016: 12）。これらの記述は第一、第二の方法とも、アルゴリズムの強制力と、その強制力に従わざるを得ない構造下におかれる人間の状況を反映したものである[14]。アルゴリズムにおいては双方が結びついて、人間に命令を下さないまでも、やはりその影響下においているということができる。

かくして、こうして狭められた範囲・もしくは決定、そしてその狭められた範囲に反映されたアルゴリズムが提示する規範にも人間は従っているといえる。必ずしも決定や規範が実現されずとも、ここで見る限り十分に「一定の方向付け」をしており、したがって強制力をもっているといえるだろう。

次項では、COMPAS 事例に Hellman 説を適用し、その不正さを分析する。順に慣習条件、表現条件、ヒエラルキー条件を検討する。

2）COMPAS 事例の Hellman 説による批判的検討

2 節 3 項では、差別かどうかの条件が表現条件とヒエラルキー条件の二つから成り立っていることを説明し、なおかつ前項ではヒエラルキー条件をアルゴリズムがみたせる見込みがあることを説明した。したがって、Hellman 説においては、アルゴリズムの差別の不正さをその判断そのものに着目して指摘することができる可能性がある。

それでもなお、次のような反論があるかもしれない。それは、ヒエラルキー条件をアルゴリズムが一般的にみたせるとしても、アルゴリズムはやはり道徳的に不正な差別をおこなうことができないというものだ。なぜなら一つには、差別も一つの行為である以上、アルゴリズムが人間と同様に差別という行為をおこなえるわけではないから、という根拠が考えられる。確かにアルゴリズムは人間と全く同様の行為をおこなうことはできないが、この反論が成立するのは、アルゴリズムのみを「行為者」と考えた場合である。実際にはアルゴリズムと人間は多くの場合協同して判断しており、その場合には人間とアルゴリズムのエージェンシーが重なっている（mesh）場合の問題として捉えることができよう（Beer 2017: 7）。本論ではアルゴリズムに行為者性を認めず、アルゴリズムの影響力を主に考慮することを目的としている。そのため、この反論は、本論文の立場からして不十分な反論である。

以下では、この不十分さを具体的に示すために、いかにアルゴリズムは道徳的に不正な表現をおこない、その内容に人間を従わせることができるか、COMPAS 事例を用いて検討する。具体的には、Hellman が提示する条件のうち、慣習条件、表現条件、そしてヒエラルキー条件をいかにみたしているかを順に検討する。これらの条件を検討することで、アルゴリズムは道徳的に不正な差別をおこなえないという反論に対する再反論とする。そのためにはまず、これまで人種、とりわけ黒人という属性が犯罪の文脈でどのような意味を持ってきたか、そして人種プロファイリ

14　ただし強制力は実際に命令の結果が実現するかどうかとは区別される（Hellman 2008=2018: 51）。ここで重要なのはあくまで、アルゴリズムがその影響下に人間をおくことができるかどうかである。

ングと比較して、COMPAS は何がどう異なるのかを述べる必要があるだろう。最終的には本項は、COMPAS が表現しているのはどのような意味なのか、それは平等な人格的価値の毀損にあたるのかを確認する。

慣習条件

慣習条件とは、問題となっている行為が、歴史上において劣っていると烙印を押すような行為とどれほど類似しているかというものであった。COMPAS の事例がそのような行為と類似しているかを検討するために、以下では、犯罪という文脈で黒人という属性がいかなる意味を持っているのかを確認する。

黒人というステレオタイプは、長らく悪いイメージを持たれてきたのは周知の通りである。例えばアメリカでは、アフリカ系アメリカ人や有色民族と思われる人を、その民族的ルーツを根拠に捜査したり、身体検査や職務質問（stop and frisk）をかけたりする人種プロファイリング（racial profiling）が行われているといわれている。他の特徴に基づいたプロファイリングと比較して、人種が特有の表現内容を持つとして問題になる可能性があるのは、当該人種カテゴリーが、他の性別や年齢といった特徴ではありえないような重要性を持つという意味で、特有のネガティブさを伴っているからである（Hellman 2005: 236）。例えば、ここでいう黒人というカテゴリーは、単に黒人とされる人々を指し示すだけでなく、アメリカで黒人であるとはどういうことなのか、どういう人だと考えられるのかというアメリカの歴史に根ざした一種の文化的理解をも指し示すものだといえるのだ。

これは黒人といった人種だけでなく、9.11 テロ以後のアラブ人や、2021 年現在におけるアジア人も同様である。9.11 テロによってアラブ人という人種にはテロリストとしてのイメージが付き纏い、それ以降重点的に取り締まられるようになった（Burkeman 2002）。また、現在はコロナウイルスの流行により、アジア人も「危険」の烙印を押されている。この二つは、社会的状況との関連で「他の状況の場合にはない」「社会的重要性」が付与されてしまった例である。

この「社会的重要性」が平等な人格的価値を否定するかどうかによって、統計的一般化に基づいたプロファイリングが道徳的に問題になるかどうかが決まる。犯罪予測の文脈で言えば、犯人としての黒人は一つのステレオタイプであり、しかもそのことは人種プロファイリングによって侮辱的に行使され、黒人を貶価する（Hellman 2005: 237）。黒人が犯人である可能性が高い、もしくは今から犯罪を行う可能性が高いとするのは軽蔑的な意味を当人に付与し、平等な人格の尊重を行っていないといえる。

COMPAS と人種プロファイリングは、人種が問題になっているという点で共通点があるものの以下の点で異なる。第一に用途の違いが挙げられる。人種プロファイリングは主に警察が行うものであるため、それが活用されるのは捜査や逮捕、身体検査などのプロセスに限られる。一方で COMPAS は、量刑判断、リハビリ、仮釈放判断など、主に逮捕以後のプロセスにおいて使用される。COMPAS が量刑判断に使用されることが問題化されることが多いが、リハビリテーションに使用することはほとんど言及されない。このことはリハビリテーションに使用されることは問題がないが、量刑判断に使用することは問題含みであると示唆するものである（Barabas et al. 2018）。つまり、使用が許容される目的と、使用が許容されない目的があると考えられているのだ。

第二に、差別の量・質が変わる可能性があることである。

確かに、COMPAS と人種プロファイリングを比較するとき、考慮するのが人種だけでない COMPAS の方が人種プロファイリングよりも多角的に当該人物を評価しているといえるかもしれない。しかしここで着目すべき点は、当該の統計的一般化が侮辱的であるかどうかであり、多角的に判断しているかどうかには関係がない。多角的に判断していたとしても、差別は差別であるというわけである。COMPAS はもともと人種を変数として考慮するソフトウェアではなく、その質問紙の内容からして本人の経歴や社会的環境に目を向けている（Propublica 2016）。人種だけによって犯罪者かどうかを判断されるよりはまだ良いという反論もありうるが、考慮する項目を増やしたとしても、その項目に侮辱的な内容が含まれているならば、それもやはり問題だといえるのではないだろうか。

念を押しておきたいのは、ここでの「侮辱的」とは、意図とは関係なく、貶価するようなものをさすということである。侮辱する意図なしに他者を侮辱することは可能である。特にこの場合、「科学的な隠れ蓑」を使っていて、意図説で処理しようとするのは賢明でない。例えば女性の勤続年数が短い傾向にあるから女性の雇控えをする場合、差別者の意図が侮辱的であるから不正であるわけではない。質問紙を作った人も、おそらくは COMPAS が最善の働きをするように作っていることが想定され、必ずしも特定の人種を侮辱する意図はないだろう。問題は、人種プロファイリングと同じように、貶価する点にあるといえるのだ。

しかしながら相違点としては、COMPAS のようなアルゴリズムの導入によって、差別が増幅されたり、差別に新たな、もしくはより強い「根拠」が付加されることが考えられる。しかもその「根拠」は、これまで使用されていたものよりも強固かつ説得的でありうる。貶価する点で人種プロファイリングと大まかな構造は同一かもしれないが、その力は大いに異なっているだろう。その力というのは、例えば裁判官などといった、しばしば実在の人間を通して行使される。

第三に、人種変数を考慮しているか否かの違いが挙げられる。COMPAS では検証の結果人種との関連性が認められており、しかも当該人種パラメータのうち、統計的に優位であったのは通常犯罪における再犯では黒人、ヒスパニック、ネイティブアメリカンの三種類の「人種」であったが、暴力犯罪の再犯では黒人だけであった（Larson et al. 2016）。

これらの違いは、COMPAS が直接表現している内容そのものに影響を与えるものではない。COMPAS は直接人種を考慮しておらず、したがって人種ゆえに誰かが再犯をするだろう、と「告げて」いるわけではないが、むしろ別の形で侮辱的に働いているといえるだろう。暗黙裡に人種が考慮されているために、当該の人種と犯罪がいかに結びついてきたかという慣習条件も引き継がれているといえる。

表現条件

これまで、犯罪予測において人種がどのような意味を持っているかを記述してきた。これは Hellman 説における慣習条件にあたるものである。COMPAS はそのような慣習を引き継いだ上で、いかなる表現を行っているといえるのだろうか。アルゴリズムによる表現は、基本的にはいかにわれわれの目に当該の属性が意味づけされて見えるかという観点から表すことができる。

特定の人物の再犯リスク判定を提示するということそのものの意味をまず考えよう。当該リスク判断は、リスクレベルを従属変数として、質問紙で尋ねられる以下の説明変数が設定されている。本人の犯罪歴、法律不遵守、被調査者の家族の犯罪、友人関係、住居、社会環境、教育、職業、

趣味、社会的孤立、犯罪的人格、怒り、犯罪への態度の合計 13 セクション、合計質問数は 137 問である（Propublica 2016）。

この判断は生まれもった環境を考慮に入れている。これをリスク判定に持ち込むことは裁判所のルールに反する。裁判で重要なのは、「その人が誰であるか」ではなく、「その人が何をしたか」であるからだ。当該リスク判断は、量刑判断が行われる法廷で使用する場合には、「その人が誰であるか」を不当に裁判に持ち込むものであるといえる。もちろんリスク判断ではその人が何をしたかも考慮されるが、これが未来の判断にも適用されることによって、「その人が何をしたか」よりも「その人が誰であるか」の比重が大きくなっているといえるのだ。彼らはこれまで危険だった人間と共通の属性を持つがゆえにこれからも危険なのである。したがって、仮に人種との関連性がなくとも、この表現は道徳的に問題になりうるといえる。

もちろん、当該表現は人種との関連でより一層問題となっている。これが問題となるのは、量刑判断の文脈で言えば裁判所が、社会的カテゴリーとしての人種ごとの差異を認めるような印象を与えるからだ。認められた差異が量刑判断やリハビリテーションなどの個々の文脈に持ち込まれるため、裁判所以外でも「その人が誰であるか」に比重が移るのである。

白人であれば再犯する確率がより低く、黒人であれば再犯する確率が高いとみなすことは、「何をしたか」ではなく「誰であるか」に考慮の比重が移ることによって、白人と黒人の評価基準が結果的に変化してしまう。誰であるかを非対称に評価することは、その非対称性はここでみてきたように、平等な人格的価値を毀損するような仕方で、である。よって、COMPAS は表現条件をみたすと考えられる。そして、もちろん結果として、黒人のステレオタイプを強化すると考えることができる。

ヒエラルキー条件

以上、本節では Hellman 説における道徳的に不正な差別の条件である慣習条件と表現条件について検討してきた。最後に COMPAS の事例におけるヒエラルキー条件について検討する。判断を下すアルゴリズムは誰のどのような立場を代理しているのだろうか。

量刑判断において裁判官の判断を補助するために COMPAS が使用されるということは、当該ソフトウェアがいわば「国のお墨付きを得た」ソフトであることを意味する。この判断が法廷に持ち込まれる場合と、リハビリテーションプログラムの決定のために持ち込まれる場合では、当該判断の役割は異なっているだろう。法廷は COMPAS が、個人を尊重した形で量刑判断を行い因果的責任を持つことを保証している（Rubel et al. 2019: 16）。したがって COMPAS は、部分的には裁判官、ひいては国家を代理して判断するといえる。

当該リスク判断は当然、裁判官の判断に影響を及ぼすことが望まれているが、問題は裁判官の判断に悪影響を与える場合である。ここでいう悪影響というのは、差別的な判断を行うようアルゴリズムが仕向ける場合を指す。

Hellman 説の先述の検討からすれば、代理していても Hellman 説は満たせるといえるだろう。その上、裁判官はプログラムを精査する技量や能力に欠けており、算出されたリスク判断を目にしてなお、公正に判断することは不可能である（Harvard Law Review 2016; Liu et al. 2019）。この点からして、アルゴリズムによる影響は悪影響である可能性が高く、しかも修正が不可能であ

る[15]。以上より、COMPAS はヒエラルキー条件をみたしている。

むすび

本論文では、以下の作業を行った。2 節で COMPAS の特殊性とこれまでなされてきた説明が不十分であり、差別の問題として依然検討に値するものであることを示した。3 節では、分析の道具として Hellman 説が適切であることを論じ、Hellman 説の詳細について説明した。4 節では、Hellman 説の理論的検討を行い、そのうちの条件を補った。次に、当初の想定よりもその適用範囲を拡張した Hellman 説を用いて、COMPAS の道徳的不正さを説明した。

この作業によって得られた結論は以下のとおりである。COMPAS がおこなっているのは道徳的に不正な差別である。その不正さの根拠は、Hellman 説が基盤とする個人の尊厳の不尊重による。また、アルゴリズムによる不尊重は個人の尊厳に反するような表現を提示することによって行われており、したがって Hellman 説が規定する道徳的に不正な差別であることを示した。

一方で課題も残されている。今回立証したのはアルゴリズムによる差別的な表現の不正さであり、人間がどのように判断を下しているかを考慮していない。アルゴリズムだけが判断を下し、実際の対応につながる場合には[16]、このような議論が成り立つかもしれない。しかしながら、現状の一般的なアルゴリズムを取り入れた判断が完全に自動化されているとは考えにくい。そのうえ、自動化されたプロセスだけに基づいた決定に人間が従わない権利を有する旨が記述されている GDPR22 条の文言を考えても、今後中心的な問題となる可能性は低い。

むしろ必要なのは、人間とアルゴリズム双方が協力して判断を下す場合にいかに意思決定が行われるかという視点であるといえよう。とりわけ機械学習の分野では、誤った意思決定を防ぎ、意思決定の結果をより公平にするために、アルゴリズムだけに判断させるのではなく、判断の過程に人間も入れる人間参加型（human-in-the-loop）機械学習という取り組みが着目されている。今回用いた事例である COMPAS は機械学習を使用していない統計的プログラムであるが、自動的に判断を下すプログラムであるという共通点から、どのように人間がアルゴリズムに補助されて意思決定を行っているかに今後着目することは有益であると考えられる。

謝　辞

この論文の執筆にあたり、東京大学大学院教授佐倉統先生、同大学院の水上拓哉氏にはさまざまな助言をいただいた。また、この論文の初期のバージョンは修士論文の形で公開されている。修士論文執筆にあたっては上記のお二人に加え、東京大学大学院の戸田聡一郎助教、東京大学大学院所属の石田柊氏に貴重なコメントをいただいた。感謝の意を表する。

15　本稿の射程を超えるが、この点はさらなる論争を喚起するだろう。というのは、これまでの法廷での判決がすべて公正だったとは言えない可能性が高く、比較すればやはりアルゴリズムの判断のほうが公正なのだという結論もありうるからである。また、バイアスの度合いに仮にそれほど違いがなかったとしても、修正可能性は人間よりもアルゴリズムの方が高いという可能性も否定できない。

16　例えばアルゴリズムがローンの審査を行い、貸与の可否までアルゴリズムが一貫して決定を行う場合などが想定される。

引用文献

Alexender, Larry, 1992, “What Makes Wrongful Discrimination Wrong?: Biases, Preferences, Stereotypes, and Proxies”, *University of Pensilvania Review*, 141: 149-219.

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner, 2016, “*Machine Bias*”, (Retrieved December 18, 2020, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

Arrow, Kenneth J., 1971, *The Theory of Discrimination*, Prinston University.

Ayres, Ian, 2002, “Outcome Tests of Racial Disparities in Police Practices”, *Justice Research and Policy*, 4: 131-42.

Barabas, Chelsea, Karthik Dinakar, Joichi Ito, Madars Virza and Jonathan Zittrain, 2018, “Intervention over Predictions: Reframing the Ethical Debate for Actual Risk Assessment”, *Proceedings of Machine Learning Research*, 81: 1-15.

Beer, David, 2017, “The social power of algorithms”, *Information, Communication & Society*, 20 (1): 1-13.

Beriain, Iñigo De Miguel, 2018, “Does the use of risk assessments in sentences respect the right to due process? A critical analysis of the Wisconsin v. Loomis ruling”, *Law, Probability and Risk*, 17 (1): 45-53.

Binns, Rauben, 2018, “Fareness in Machine Learning: Lessons from Political Philosophy”, in *Journal of Machine Learning Research*, 81: 1-11.

Brennan, Tim, William Dieterich and Beate Ehret, 2009, “Evaluating the predictive validity of the COMPAS risk and needs assessment system”, *Criminal Justice and Behavior*, 36 (1): 21-40.

Burkeman, Oliver, 2002, “Visa detainees allege beatings”, The Guardian, (Retrieved January 1, 2020, https://www. theguardian. com/world/2002/may/24/usa. afghanistan).

Chenny-Lippold, John, 2017, *We Are Data*：*Algorithms and the Making of Our Digital Selves*, New York: New York University Press.（= 2018, 高取芳彦訳,『We are Data――アルゴリズムが「私」を決める』日経 BP 社.）

Corbett-Davies, Sam and Sharad Goel, 2018, “The measure and mismeasure of fairness: A critical review of fair machine learning”, *ArXiv*, 1808.00023v2[cs.CY], (Retrieved December 18, 2020, https://arxiv.org/abs/1808.00023).

Dieterich, William, Christina Mendoza and Tim Brennan, 2016, “*COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*”, (Retrieved December 18, 2020, http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf).

Dressel, Julia and Hany Farid, 2018, “The Accuracy, Fairness, and Limits of Predicting Recidivism”, *Science Advances*, 4 (1): 1-5.

Eidelson, Benjamin, 2015, *Discrimination and Disrespect*, Oxford University Press.

Foucault, Michele, 2004, *Society must be defended: Lectures at the Collège de France, 1975-76*, London: Penguin.

Freeman, Katherine, 2016, “Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis”, *North Calorina Journal of Law& Technology*, 18 (5): 75-106.

Goel, Sharad, Ravi Shroff, Jennifer L. Skeem and Christopher Slobogin, 2018, “The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment”, *SSRN Electronic Journal*, 1-21.

Harvard Law Review, 2017, “Criminal Law — Sentencing Guideline — Wisconsin Supreme Court Requires Warning before Use of Algorithmic Risk Assessments in Sentencing. — State v. Loomis, 881 N.W.2d 749 (Wis. 2016).”, *Harvard Law Review*, 130: 1530-7.

Hellman, Deborah, 2005, “Racial Profiling and the Meaning of Racial Categories”, Andrew I. Cohen and Christpher H. Wellman eds., *Contemporary Debates in Applied Ethics*, Hoboken: Wiley-Blackwell, 232-44.

――― , 2008, *When is Discriminarion Wrong?*, Cambridge: Harvard University Press.

（= 2018, 池田喬・堀田義太郎訳, 『差別はいつ悪質になるのか』法政大学出版局. ）

Hing, Julianne, 2009, *HP Face-Tracker Software Can't See Black People*, (Retrieved December 18, 2020, https://www.colorlines.com/articles/hp-face-tracker-software-cant-see-black-people).

Introna, Lucas D., 2016, “Algorithms, Governance, and Governmentality: On Governing Academic Writing”, *Science, Technology & Human Values*, 41 (1): 17-49.

石田柊, 2019,「差別と危害 : 帰結主義的差別論の擁護」『社会と倫理』34: 1-12.

Khaitan, Tarunabh, 2018, “Indirect Discrimination”, in *The Routledge Handbook of the Ethics of Discrimination*, London: Routledge, 30-41.

Kleinberg, John, Sendhil Mullainahan and Manish Raghavan, 2017, “Inherent Trade-offs in the Fair Determination of Risk Scores”, *Computer Science, Mathematics*, 67:1-23.

Knight, Cirl, 2013, “The Injustice of Discrimination”, *South African Journal of Philosophy*, 32 (1): 47-59.

Koene, Ansgar, Liz Dorthwaite and Suchana Seth, 2018, “IEEE P7003™ standard for algorithmic bias considerations”, *IEEE*, (Retrieved December 18, 2020, http://fairware.cs.umass.edu/papers/Koene.pdf).

Larson, Jeff, Surya Mattu, Lauren Kirchner and Julia Angwin, 2016, *How We Analysed the COMPAS Redicisism Algorithm*, (Retrieved December 18, 2020, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm).

Lippert-Rasmussen, Kasper, 2006, “The Badness of Discrimination”, *Ethical Theory and Moral Practice*, 9 (2): 167-185.

Liu, Han-Wei, Ching-Fu Lin and Yu-Jie Chen, 2019, “Beyond State v Loomis: artificial intelligence, government algorithmization and accountability”, *International Journal of Law and Information Technology*, 27 (2): 122-41.

Noble, Safiya U., 2018, *Algorithms of Oppression*, New York: New York University Press.

O'Neil, Cathy, 2016, *Weapons of Math Destruction: How Big Data Increases Inequality and Threaten Democracy*, Phoenix: Crown.

Phelps, Edmund S., 1972, “The Statistical Theory of Racism and Sexism”, *American Economic Review*, 62 (4): 659-61.

Propublica, 2016, “Sample-COMPAS-Risk-Assessment-COMPAS-“CORE”,” documentcloud, (Retrieved December 18, 2020, https://www. documentcloud. org/documents/2702103-Sample-Risk-Assessment-COMPAS-CORE. html).

Rouvroy, Antoinette and Thomas Berns, 2013, *Algorithmic Governmentality and Prospects of Emancipation: Disparateness as a Precognition for Individuation through Relationships?*, 177 (1): 163-96.

Rubel, Alan, Clinton Castro and Adam Pham, 2018, “Algorithm, Bias, and the Importance of Agency”, Jo

Bates, Paul D. Clough, Robert Jäschke, and Jahna Otterbacher eds., *Proceedings of the International Workshop on Bias in Information, Algorithms, and Systems*, 9-13.

Sandvig, Christian, Kevin Hamiton, Karrie Karahalios and Cedric Langbort, 2016, "When the Algorithm Itself is a Racist: Diagnosing Ethical Harm in the Basic Components of Software", *International Journal of Communication*, 10: 4972-90.

Scanlon, Thomas M., 2008, *Moral Dimensions: Permissibility, Meaning, Blame*, Cambridge: Belknap Press.

Schauer, Frederick, 2003, *Profiles, Probabilities, and Stereptypes*, Cambridge: Harvard University Press.

State *v.* Loomis, 881 N.W.2d 749, 767（Wis. 2016）.

Sumpter, David, 2018, *Outnumbered: from Facebook and Google to Fake News and Filter-Bubbles: The Algorithms That Control Our Lives*, London: Bloomsbury.（= 2019, 千葉敏生・橋本篤史訳,『アルゴリズムはどれほど人を支配しているのか? ── あなたを分析し、操作するブラックボックスの真実』光文社.）

Sweeney, Latanya, 2013, "Discrimination in Online Ad Delivery", *ArXiv:1301.6822 [Cs],*（Retrieved December 18, 2020, http://arxiv.org/abs/1301.6822）.

Washington, Anne L, 2019, "How to Argue with an Algorithm: Lessons FROM THE COMPAS-PROPUBLICA DEBATE", *The Colorado Technology Law Journal*, 17 (1): 1-37.

山本龍彦・尾崎愛美, 2018,「アルゴリズムと公正 ── State v. Loomis 判決を素材に」『科学技術社会論研究』16: 96-107.

Yeung, Karen, 2017, " 'Hypernudge': Big Data as a Mode of Regulation by Design", *Information, Communication& Society*, 20 (1): 1-19.

Ziewitz, Malte, 2016, "Governing Algorhthms: Myth, Mess, and Methods", *Science, Technology, & Human Values*, 41 (1): 3-16.