アルゴリズムの公平性を測定する

著者：デボラ・ヘルマン
要約担当者：福家 佑亮[1]

出典

Hellman, Deborah. "Measuring algorithmic fairness." Virginia Law Review 106 (2020): 811-866.

キーワード

- アルゴリズムの公平性(algorithmic fairness)
- 等予測率(equal predictive value)と過誤率の均衡(error rate balance)
- 等過誤率(error ratio parity)
- 不正義の悪化(compounding injustice)
- 異なる取り扱い(disparate treatment)と異なる効果(disparate impact)

Introduction

本論文の目的は、アルゴリズムの公平性(fairness)に関して、概念・規範・法の 3 つの観点から貢献を行うことである。概念に関しては、コンピューター・サイエンス関連文献で支配的な公平性の測定方法の 1 つは、何を行うべきかではなく、何を信じるべきか(what to *believe*)に関わり、後者は公平な取り扱いの測定指標としては不適切であると論じる。規範に関しては、偽陽性と偽陰性の割合が、2 つのグループがアルゴリズムによって公平に扱われているかどうかの重要な指標であると主張する。法に関しては、反差別法が人種等の属性の使用をあらゆる文脈で禁じているという広く共有された想定に対して挑戦を行う。

いかに公平性を測定すべきかに関する論争を理解するうえで有用であるのが COMPAS の例である 。COMPAS は、再犯可能性を示すスコアを算出するツールである。このツールに対して、ProPublica というウェブサイトは、COMPAS がアルゴリズムに人種情報を明示的に使用していないにもかかわらず、黒人の逮捕者や受刑者が、白人のそれよりも、誤って再犯率が高いと判定される確率が遥かに高いことを理由に、COMPAS が黒人と白人を異なる仕方で扱っていると主張した。これに対して、COMPAS を運営する Northpointe は、COMPAS は、白人と黒人の被告人に関しておおよそ同じ割合で間違った判断を下しているとして、両人種を同じように扱っていると主張した。Northpointe は、白人と黒人が同じスコアを付与された場合、両方とも同じ確率で再犯するかどうかという事実に着目していた。

[1] 立命館大学　非常勤講師

一方で、ProPublica は、再犯しなかった黒人と白人が、アルゴリズムから同じ確率で低いスコアを付与されていたかどうかに着目していたのである。

本論文の第 1 部では、Northpointe の主張する公平性は何を信じるべきか(what we ought to *believe*)に関連する一方で、ProPublica 側の公平性は何を行うべきか(what we ought to *do*)に関わり、Northpointe の主張する公平性は異なる集団間の公平な取り扱いの問題には直接関係しないと論じる。第 2 部では、集団間での偽陽性と偽陰性の比率の違い(これは ProPublica が主張する公平性に関係する)が不公平性の存在を示唆すると論じる。偽陽性と偽陰性の比率を等しくしようとするアプローチには欠点があるとはいえ、偽陽性と偽陰性の比率の違いは、アルゴリズムの公平性について重要な点を明らかにする。第 3 部では、人種等の保護された属性を使用してアルゴリズムの正確性を向上させることが、人種に基づいた異なる取り扱いに該当せず、法的に認められ得るケースがあると主張する。

1. Predictive accuracy and Belief

A. The Measures and What They Measure

問題となるアルゴリズムの公平性の測定方法や COMPAS の問題を理解するために、以下の事例を導入する。

疾病検査の事例：特定の病気について、疾病に罹患しているかどうかを調べる医療検査を想定してみよう。以下の表で描写されているように、この検査は誰が罹患しているのかを完璧に検出するものではないが、青色人と緑色人について一定程度信頼性のおける結果を示している。実際の結果は列(Sick, Healthy)に、医療検査による予測は行(＋,－)に示されている。

TRUE OUTCOME			
		Sick	Healthy
TEST	＋	60[a]	20[b]
RESULT	－	6[c]	14[d]
Table1-1(Greens)			

TRUE OUTCOME			
		Sick	Healthy
TEST	＋	16[a]	5[b]
RESULT	－	22[c]	57[d]
Table1-2(Blues)			

たとえば、緑色人の場合は、検査受けた 100 人中の 60 人が真陽性、20 人が偽陽性、6 人が偽陰性、14 人が真陰性である。検査によって陽性と判定された緑色人が、実際に陽性である確率は 75％であり、これを陽性的中率(*positive predictive value*)、検査によって陰性と判定された緑色人が、実際に陽性である確率は 70％であり、これを陰性的中率(*negative predictive value*)と呼ぼう。これを青色人と比較すると、青色人の場合、陽性的中率は 76%、

陰性的中率は 72%であり、疾病検査はおおよそ正確な予測を行っている。

しかし、病気の緑色人(青色人)が正確な検査結果を得る確率を問題にする場合には、話が違ってくる。

TRUE OUTCOME			
		Sick	Healthy
TEST	+	60^a	20^b
RESULT	−	6^c	14^d

Table1-1(Greens)

TRUE OUTCOME			
		Sick	Healthy
TEST	+	16^a	5^b
RESULT	−	22^c	57^d

Table1-2(Blues)

病気の緑色人が正確な検査結果を得る確率は約 91%であるが、病気の青色人が正確な検査結果を得る確率は約 42%である。また、健康な緑色人が正確な検査結果を得る確率は約 41%である一方で、健康な青色人が正確な検査結果を得る確率は約 91%である。以上の事例から得られる重要な点は、疾病検査は、健康状態の予測について、青色人と緑色人の両方に対して正確であるが、過誤が発生する場合、発生する種類の過誤が異なるということである。緑色人の場合、過誤は偽陽性である確率が高く、青色人の場合は偽陰性である確率が高い。

以下では、疾病検査の事例を COMPAS に置き換えた表を説明に用いる。

TRUE OUTCOME			
SCORE		Will Recidivate	Will Not Recidivate
	High Risk	60^a	20^b
	Low Risk	6^c	14^d

Table3-1(Blacks)

TRUE OUTCOME			
SCORE		Will Recidivate	Will Not Recidivate
	High Risk	16^a	5^b
	Low Risk	22^c	57^d

Table3-2(Whites)

この COMPAS の仮想事例は、白人の扱いに対して黒人を公平に扱っているだろうか。ProPublica に対する最良の応答は、アルゴリズムを調整して、再犯予測率と過誤の種類の両方の側面で両人種を公平に扱うようにすることである。しかし、この仮想事例が示しているように、ある属性に関する基準率(base rate)が 2 つの集団で異なる場合(緑色人の 66%が病気であるのに対して、青色人の有病率は 38%に過ぎない)、予測率と過誤の種類の両方を平等化することは不可能である。注意点として、基準率を定めるデータそのものが信頼性に欠け、バイアスがかかった仕方で不正確でありうるという問題があり得る。この測定誤差(measurement error)と呼ばれる問題については、第 2 部で再度検討する。

以上の 2 つの測定には様々な名前が付けられているが、ここでは、陽性的中率と陰性的中率が同じである場合を等予測率(*equal predictive value*: EPV)、偽陽性と偽陰性の確率がそれぞれの集団で同じ場合を過誤率の均衡(*error rate balance*)と呼ぶ。

B. Predictive Accuracy and Belief

ほとんどの状況において、EPV と過誤率の均衡が同時に成立することができないという事実から、どちらを優先すべきであり、そしてその理由は何かという問題が生じる。だがこの問題を扱う前に、比較が問題とならない個人のケースにおいて、予測の正確性と過誤の種類が持つ認識的・実践的な重要性に着目することは有益だろう

1. Individual Cases

予測率の高い検査やアルゴリズムは情報を与えてくれるが、高予測率自体はどのように行為すべきかについて教えてくれない。この理由を明らかにするために、以下の例を考えてみよう。

レズリーと赤ん坊と蝙蝠(*Leslie, the Baby, and the Bat*): レズリーは、彼女の娘が赤ん坊であったある日、生きた蝙蝠を自宅で見つけた。やがて蝙蝠は家から出て行ったが、それにもかかわらず、かかりつけの小児科医は赤ん坊に狂犬病ワクチンを接種させることを勧めた。なぜか？蝙蝠にかまれた可能性も、蝙蝠が狂犬病を保有していた可能性も低いと医師は考えていたが、狂犬病ウイルスに暴露した後すぐさま治療しないと致命的となるという理由から、それでもなお医師はワクチン接種を勧めた。仮にその赤ん坊が狂犬病に罹患した確率を算出するのならば、その確率は極めて低いものとなるだろう。しかし、偽陰性の判断が抱えるコストが非常に大きいため(狂犬病に罹患した患者に治療しないと死に至る)、医師は治療を勧めた。

この例が示しているように、我々が何を信じるべきか(赤ん坊は狂犬病に罹患していない)と、我々が何を行うべきか(赤ん坊に狂犬病の治療を行う)は、異なる考慮要因に影響を受ける。何を行うべきかについての決定は、過誤によって生じるコストに決定的に依存している。

偽陽性と偽陰性によって生じるコストは我々が何を行うべきかについて影響を与える。もう 1 つの例を考えてみよう。

異なる法律上の基準(*Different Legal Standards*): ジョンはビルの顔面を殴ったかどで逮捕・裁判にかけられた。法廷で提示された証拠は、ジョンがビルを殴ったという主張を支持している。スーは、証拠について傍聴していた陪審員のメンバーである。スーは、ジョンがビルを殴ったと信じているが、確証は持てない。ジョンがビルを殴ったかどうかについてのスーの

信憑度の度合いは 75%である。

75%の信憑度がジョンに有罪票を投じるのに充分であるかどうかは、問題となるのが民事裁判か刑事裁判かによって異なる。そして、刑事裁判と民事裁判の違いは、過誤によって生じるコストに存在する。刑事裁判での立証責任が非常に重いのは、偽陽性（無実の人を有罪にする）のコストが非常に高く、また偽陰性（有罪の人を無罪にする）のコストよりもはるかに重いからである。これに対して、民事裁判では、偽陽性（無実の人に責任を帰する）のコストは、偽陰性（有罪の人に責任を負わせない）のコストとほぼ同じである。我々が信じるべきことは証拠に応じて変化する。我々が行うべきことは、我々が信じていることと、我々の信念が間違いであると判明した場合に、当該の信念に基づいて行動することによって生じる危険性に応じて変化する。

以上の 2 つの例は、予測上の正確性と行為の間の関係性について 2 つの点を例証している。1 つは、蝙蝠の例が示しているように、正確な信念は、いかに行為すべきかを決定するために必要ではないという点。もう 1 つは、異なる法律上の基準の例が示しているように、正確な信念は、いかに行為すべきかにとっても十分でないということである。

2. Collectives Cases

それでは、検査があるグループよりも他のグループに対して正確であることは不公平を意味するのだろうか。この問題を検討するために、以下の事例を考えてみよう。

教育上の選択(*Pedagogical Choice*)：教授は、どの種類のテストを学生に課すのかを決めなければならない。全て小論文形式か、それとも全て多項選択式か、あるいは両者の組み合わせかを選べるとしよう。更に、全てが論述形式の場合には、テストは女子学生よりも男子学生の実際の知識をよりうまく反映でき、全てが多項選択式の場合には、その逆であると想定しよう。教授は、75%は多項選択式形式、25%は小論文形式を選択した。

この場合、試験は男性よりも女性の実際の知識に関して信頼のおける指標となるが、男性と女性どちらが不公平な取り扱いを受けたのだろうか。男性の方が実際よりも良い点数をだしたり、逆に悪い点数をだす可能性もあるため、この問いに答えるためにはより多くの情報が必要となる。言い換えれば、公平性にとって重要であるのは、テストが男女に対して等しく正確であるかどうかというより、不正確さがどのように機能するかということなのである。

しかし、精度の低い尺度で判断されることには、不公平があるのではないか。ここで、テストが、男子学生に対して実際よりもよい点数を出すものと想定しよう。ある意味では、男性はこの予測の正確性の低下によって恩恵を受けているが、別の意味では、男性は危害を被っている。準備万端の男子学生は、準備不足の男子学生と得点差をつけることができないか

らである。ここで、受験者の性別を知ることができないと仮定すると、準備万端の女子学生も準備不足の男子学生と一括りにされてしまい、得点差をつけることができなくなっている。これが正しいとすれば、男性受験者は女性受験者に比べて不公平な扱いを受けているわけではない。むしろ、準備万端の受験者は、準備不足の受験者との差別化が困難なテストを受験させられる点で、不公平な取り扱いを受けていると言える。

この不公平性の主張は、性別に基づいた不公平性の主張ではなく、誰もが利用可能な最も正確なテストを受験する権利があるという主張である。これは、あるグループの受験者が別のグループの受験者に対して公平に扱われているかどうかのような比較的な主張ではなく、利用可能な最善の意思決定ツールに対する権利の主張である。つまり、この主張は、集団間の不公平性に関する主張ではないのである。

Ⅱ. Error rates and fairness: The normative claim

第 1 部では、予測の同等性が信念に影響を与えることを確認したが、予測の同等性の欠如は行為に間接的に影響を与えるのみであり、多くの人々の関心は、より実践的な意味での公平性にあるように思われる。第 2 部では、集団間の偽陽性と偽陰性の比率の違いが、このような実践的な意味での不公正の存在を示唆していると主張する。この議論を提示する前に、より実践的な意味で「公平性(fairness)」という言葉がどのように使われるのか、いくつかの異なる用法を明らかにしておくことが有益であろう。

A. Fairness Three ways

1 つ目の概念上の区別は、公平性の比較的な概念と、非比較的なそれの区別である。公平性の比較的な概念は、Y がいかに取り扱われているかと比較して、X が公平に扱われているかどうかを検討する。他方で、公平性の非比較的な概念は、他の個人や集団がどう扱われているかにかかわらず、Xが取り扱われるべき仕方で取り扱われているかどうかを問う。

焦点がアルゴリズムの公平性にある場合、公平性の比較的な概念は、更に 2 つに分類することができる。1 つは、アルゴリズムによって評価されている個人や集団が、同じアルゴリズムに評価されているほかの個人や集団と比較して、公平に扱われているかどうかを問う。もう 1 つは、アルゴリズムによって評価を受ける人々と、この評価に影響を受ける人々両方に着目する概念である。たとえば、アルゴリズムが、白人と比較して、黒人を公平に扱っているかを問うとき、アルゴリズムによって評価される黒人と白人だけでなく、直接評価されないものの、この評価に影響を受ける黒人と白人も考慮に入れるのがこの公平性概念である。

以下では、公平性の比較的な概念、特にアルゴリズムで評価される集団間での比較に焦点を当てる。

B. Error ratio parity

本節では、偽陽性率と偽陰性率の割合がアルゴリズムの評価対象となる集団間で同じであるかどうかに焦点を当てるべきであると主張する。この測定を等過誤率(*Error Ratio Parity*: ERP)と呼ぼう。アルゴリズムが公平かそうでないかは、等過誤率だけで決まらないものの、ERP は、問題となっている集団が過去不正に取り扱われてきた場合には不公正の存在を示唆するものである。

アルゴリズムも完璧ではありえないため、アルゴリズムによって生じる偽陽性と偽陰性のコストについて重み付けを行う必要がある。そして、適切な重み付けは状況によって異なるものとなる。空港でテロリストを探知する場合、テロリストを取り逃がした場合のコストが非常に高い。そのため、テロリストを探知するためのツールの偽陽性率は、高く設定される可能性が高い。反対に、刑事裁判の場合は、「十人の真犯人を逃すとも一人の無辜を罰する勿れ」という「ブラックストーンの比率(Blackstone's ratio)」が示すように、偽陰性のコストが最大の関心事となる。

問題となる人物が関連する属性を備えていることに対してどれだけ確信していなければならないかについて、2 つの異なるルールが存在する。ルール A は、関連する事実が真理であることについてある一定程度の確信を求めるのに対して、ルール B は、偽陽性と偽陰性の比率が一定であることを求めるものである。ルール A と B を人種間で異なる仕方で適用することは人種に基づく異なる取り扱いとなる。

たとえば、異なる法律上の基準(*Different Legal Standards*)を例に考察してみよう。ジョンが黒人である場合、ジョンが無罪である確率の方が高いにもかかわらず、有罪票を投じることは人種に基づく異なる取り扱いである。私の主張は、異なる取り扱いは様々な仕方で理解することができるというものである。つまり、こうした異なる取り扱いは、黒人に有罪を下すための確信度が白人のそれよりも低い(ルール A の異なる適用)、また、被告人に有利なブラックストーンの 10 対 1 の過誤率が白人には適用されるが黒人には適用されない(ルール B の異なる適用)といった仕方で生じうる。

アルゴリズムによって評価された集団間の公平性は、各集団について同じ仕方で偽陽性と偽陰性の比率を適用することを求めているのである。

C. The limitations of Error Ratio Parity

しかし、刑事裁判を例に考えると、等過誤率は、逮捕された人々の中の特定の集団に固有のものではないという問題を抱えている。この問題について理解するために、路上で目に付いた 100 人を逮捕したと想像してみよう。この場合、有罪率は 10 対 1 の比率とはかなり異なる比率となるだろう。なぜなら、これらのランダムに逮捕された 100 人が罪を犯していたと考える理由は存在しないからである。つまり、有罪率に関する 10 対 1 の比率は、有罪の可能性について同じ程度の証拠がある黒人と白人について、同じでなければならない。しかし、過誤率は、有罪の可能性について同じ程度の証拠がある個人だけでなく、(証拠の程度

が低いないしは高い)すべての個人に適用される。その結果、ある集団が他の集団よりも再犯率が高い場合、再犯率が高いとされた集団はそうでない集団と比べて、より多くの人々が間違って再犯の可能性が高いハイリスク群と評価され、偽陽性の割合が高くなるだろう。したがって、過誤率に関する情報だけでは、2 つの集団に関して、偽陽性と偽陰性の負担を平準化しているかどうかを判断することができない。ランダムに逮捕された人なのか、それとも疑わしき犯罪者であるのかといった、対象の集団の背後にある属性を理解して初めて、(不)公平性に関する主張が可能となる。

しかし、この結論は、等過誤率の欠如が無意味であることを意味するものではない。等過誤率の欠如は、基準率の分布の違いから生じる。過誤率の不平等は、基準率に違いがあることを強調することで、データにバイアスがかかっていたり、データが過去の不公正さの産物である可能性を検証する理由となるのである。

D. Why Error Ratio Parity is relevant to fair treatment

等過誤率の欠如が重要であるのは、基準率の違いが現実の世界でどのように姿を現すかを明確にし、基準率の違いについて精査する義務を生み出すからである。

まず第 1 に、基準率の差が等過誤率における著しい違いを生み出しているという事実は、アルゴリズムが依拠するデータが正確なものであるかどうかをチェックする理由となる。過去に差別を受けてきた集団の基準率は不正確である危険性がある。アルゴリズムが依拠するデータが、対象となる属性の代替指標にすぎず、しかも代替指標が偏向していることがある。また、偽陰性と偽陽性の割合が著しく異なる場合、保護されるべき集団にバイアスがかかった形でデータが不正確ではないかどうかを検証する理由となる。

第 2 に等過誤率の欠如が、アルゴリズムが既に存在している不正義を悪化させていることを示している可能性がある。たとえば、低い学歴が再犯率の予測に使えるとしよう。また、学校の質が良くないため、黒人の方が中退する確率が高いとも想定してみよう。アルゴリズムが学歴を再犯率の予測に使うのならば、黒人が過去不公平に扱われてきた事実を、今日彼らを不当に扱うことの正当化に使うことにつながるかもしれない。これが「不正義の悪化(*compounding injustice*)」と私が呼ぶ問題である。

アメリカの歴史を考えれば、黒人と白人の間での犯罪に関する基準率の違いは、測定誤差という事実問題と不正義の悪化という道徳的問題が原因であると推測するに足る理由が存在する。したがって、等過誤率の欠如は、他の場合にもましてデータの正確性について検証し、既に存在する不正義を悪化させる恐れのあるデータの使用に慎重になる道徳的理由を生じさせる。

E. Rebuttal and Reply

一部の学者は、アルゴリズムによって人種的マイノリティが被る害は、評価によって影響を受ける同じマイノリティ集団の他のメンバーへの利益によって、ある程度補うことがで

きると示唆している。たとえば、Aziz Huq は、刑事司法の文脈でアルゴリズムの使用が許されるかどうかは、アルゴリズムが集団としての人種的マイノリティに危害よりも便益を与えるかどうかを通じて評価すべきであると主張している[2]。

この議論には確かに大きな魅力がある。しかし、この議論は、問題とするべき公平性が何かについて暗黙の想定に依拠したものとなっている。私の議論は、アルゴリズムによって直接評価を受ける集団間の比較に焦点を当てているが、Huq はアルゴリズムによって直接評価を受けない集団への影響も考慮している。私の見解では、以下の議論が示すように、アルゴリズムによって直接評価を受ける集団間での比較の方が道徳的に重要な問題である。

その理由について検討するために、教育上の選択(*Pedagogical Choice*)の事例に戻ろう。

TRUE OUTCOME			
GRADE		Prepared	Unprepared
	A	60[a]	20[b]
	C	6[c]	14[d]

Table4-1(Women)

TRUE OUTCOME			
GRADE		Prepared	Unprepared
	A	16[a]	5[b]
	C	22[c]	57[d]

Table4-2(Men)

この表によれば、テストの評価(A or C)に関して等予測率が成立している。つまり、A 評価を得た女子生徒が実際に A 評価に値する確率は 75%であり、男子学生の場合は 76%である。C 評価を得た女子生徒が実際に C 評価に値する確率は 70%であり、男子学生の場合は 72%である。しかし、過誤率の同等性は成立しておらず、女子学生の場合、偽陰性よりもはるかに多くの偽陽性が生じており、男子学生の場合はその逆である。

私がここで検証したいのは、テストが男性を不公平に取り扱っているとして、この不公平性がアルゴリズムによって評価されていない他の男性への便益によって補填され得るという議論である。この主張を検討するために、以下の仮想事例を検討してみよう。

怠け者の昇進(*The Slacker Bump*): 現在の労働市場において、よりスキルが必要とされる職に対して、男性が女性よりも準備が不足していると想定しよう。もし、十分に準備した男性が、誤ってテストによって準備不足と評価された場合、当該の男性の競争力はテストを受けていない男性よりも低くなる。そして、もし多くの職がジェンダーによって分離されているとすると、アルゴリズムによって誤った評価を受けた男性への危害は、当該男性と競争した可能性の高い、スキルに乏しい男性に便益を与えるはずである。この時、テストは男性たちを公平に扱っていると結論付けることができるか？

[2] Huq, Aziz Z. "Racial equity in algorithmic criminal justice." Duke Law Journal 68 (2018): 1043.

私の答えはこの議論は成功していないというものだ。怠け者の昇進(*The Slacker Bump*)の事例は、他の人への便益も道徳的な重要性を持つが、アルゴリズムによって評価を受ける 2 つの集団間の不公平性を改善するものではないということを示している。

III. RACIAL CLASSIFICATION WITHOUT DISPARATE TREATMENT: THE LEGAL CLAIM

等過誤率の欠如が示唆する不公平性をどうやって緩和するか。2 つの可能性があり、1 つは過誤による負担を減少させること、もう 1 つはアルゴリズムの正確性を向上させ過誤の頻度を減らすことである。前者のアプローチには、政治的・実践的な障壁があり、後者のアプローチには、正確性を向上させるために人種等の保護された特性を使用することは違法であるという認識が障壁となっている。第 3 部では、後者の違法性が誇張されたものであり、アルゴリズムにおいて人種に基づいた分類を使用することは、人種に基づく異なる取り扱いには必ずしも該当しないと主張する。

A. Reduce the Burden of Errors

偽陽性と偽陰性の比率が 2 つの集団間で異なるとき、過誤率の不均衡が存在する。一方の過誤が他の過誤よりも多くの負担を生み出すものであれば、過誤率の不均衡は問題である。異なる負担を減少させる 1 つの戦略は、過誤によって生じる結果を変更することである。

たとえば、Sandra Mayson は、ハイリスク群という分類が、投獄ではなく支援や機会へのより大きなアクセスという結果をもたらすのならば、黒人の被告の間でより高い偽陽性が生じることへの憂慮は減少すると論じている[3]。しかし、このアプローチには、要求値が高いことに加えて、過誤によって生じる負担を減らす方法を、刑事裁判や雇用等の様々な文脈に応じて適用することが難しいという実践上の問題点が存在する。

B. Improve Accuracy Overall by Using Protected Traits

正確性を向上させる 1 つの方法は、アルゴリズムが人種や性別などの保護された特性を組み込むことである。しかし、法律家だけでなく、アルゴリズム設計者の間でも、人種等をアルゴリズムに使用することは違法であると認識されているため、アルゴリズムは、「人種に対して盲目(*race blind*)」であるように設計されている。以下では、この認識が誇張されたものであると論じ、保護された特性をアルゴリズムに使用することが法的に許容される場合もあると主張する。結論を先取りすれば、人種間で異なる閾値を設定することは許されないが、人種以外の属性をアルゴリズムに導入する際に、人種を参照することは許されると結論づける。この結論には、アルゴリズムに保護された特性が組み込まれることで、正確性と公平性が改善されるかもしれないという実践的な意義と、現状の法学上の定説が認める以

[3] Mayson, Sandra G. "Bias in, bias out." Yale Law Journal 128 (2019): 2218.

上に、「異なる取り扱い(*disparate treatment*)」と「異なる効果(*disparate impact*)」の区別が曖昧であることを示すという概念上の意義が存在する。

1. Different Thresholds Versus Different Tracks

本節では、保護された特性である「人種(*race*)」情報をアルゴリズムが使用する 2 つの方法に焦点を当て、一方は法的に問題含みであるが、他方はそうではないと論じる。

a. Legal Background

はじめに、米国の反差別法に関する簡単な導入が有益である。多くの場合、年齢等に基づいて区別を行うことは法的に禁じられていない。しかし、ある特定の属性に基づく分類や区別は法的に問題含みであり、このような属性は「保護された特性(*protected traits*)」と呼ばれ、人種や性別などが含まれる。憲法と連邦法では、保護された特性の範囲が異なるが、以下では取り扱われる事例が憲法にカバーされる範囲のため、憲法に焦点を当てる。

悪意ある動機や明示的に人種に基づく分類は、人種に基づく異なる取り扱いに該当し、厳格審査に服する可能性がある。しかし、等予測率を目指すことや、等予測率を犠牲にして過誤率の割合を等しくすることは、人種に基づく異なる取り扱いに該当しない。なぜなら、前者は悪意ある動機に基づいておらず、後者も予測率を不平等にして異なる効果を与えることを目的とはしていないからである。

それでは、アルゴリズムの設計者が、アルゴリズムに人種情報を使用することで、過誤率の不均等を縮減し正確性を向上させることは、法的に許容されるであろうか。

b. Different Threshold

過誤率の不均等を抑制するためにアルゴリズムが人種に基づく分類を使用する 1 つの方法は、対象となる特性に関して、人種間で異なる閾値を設定することである。白人の場合では、点数が 6 点以上だと高リスク群に分類されるが、黒人の場合では 8 点以上で高リスク群に分類される場合、アルゴリズムは人種間で異なる閾値を使用している。このようなアプローチは、法的に禁止されていると広く考えられており、私も閾値を用いるアプローチは厳格審査に服し、それを通過する見込みも低いと考えるため、この手法についてはこれ以上論じない。

c. Different Tracks Within Algorithms

人種に基づく分類を用いるもう 1 つの方法は、ある予測を行う際にどの特性を使用するかについて、人種を参照するものである。再犯率を予測するある特性について、ある人種よりも他の人種に対して予測の精度が高いと想定してみよう。たとえば、ある研究では、居住の安定性(housing stability)がマイノリティの再犯率の予測に関して、白人の場合よりも信頼

性に劣る可能性が指摘されている[4]。この時、再犯率を予測するアルゴリズムに、白人の場合は居住の安定性を考慮するが、黒人の場合はそうしないといったことが考えられ得る。

こうした仕方で人種カテゴリーを使用することを法が禁じているかどうかは、アルゴリズムに人種情報を使用することが人種に基づく異なる取り扱いに該当するかに依るが、興味深いことにその答えは明確ではない。以上のような事例は、異なる取り扱いという概念によって法が意味するところは何かを明らかにする必要性を示している。

2. Racial Classification Without Disparate Treatment

以下では、人種に基づく分類が異なる取り扱いに該当しない 2 つの状況の考察から始め、これらの事例から 2 つの原理を抽出する。これらの原理を用いながら、アルゴリズムにおける人種に基づく分類の使用について検討を加え、これが人種に基づく異なる取り扱いに該当せず、高められた司法審査(heightened judicial review)に服さない可能性があると結論付ける。

a. Information Not Use

法や政策が人種に基づく分類を使用しても、必ずしも異なる取り扱いに該当するわけではない。たとえば、国勢調査等といった情報収集は人種カテゴリーを活用しているが、人種に基づく異なる取り扱いに該当しないように思われる。こうした実践において人種に基づく分類が広範に使用されているという事実は、人種に基づく分類が許容されているということを示唆している。

Morales v. Daley 判決では、人種情報を収集する国勢調査は厳格審査に服すべきと原告側が主張した。しかし、裁判所は原告の立場は、政府が統治活動に必要な人口統計上の情報を収集することと、やむにやまれぬ利益 (a compelling interest)なく疑わしき分類(suspect classification)を政府が使用することの区別に関する誤解に依拠するものとして、原告の主張を退けた。情報の収集は使用と異なるため、情報の収集は異なる取り扱いに該当せず、したがって厳格審査に服さないのである。

以上の事例は、人種に基づく分類が実際の世界に及ぼす影響によって、異なる取り扱いに該当する人種に基づく分類とそうでないものが区別されていることを例証している。加えて、国勢調査の事例は、異なる取り扱いに該当する人種に基づく分類であるためには、その影響が直接的でなければならないことを示唆している。

b. No Racial Generalization

---

[4] Corbett-Davies, Sam, et al. "Algorithmic decision making and the cost of fairness." Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017.

人種に基づく分類が直接的な影響を及ぼすとしても、必ずしも厳格審査に服するわけではない。そうした分類が使用される仕方も重要である。人種に基づく分類の使用が人種集団に関する一般化に依拠する際には厳格審査が適用され、そうでない場合には必ずしも厳格審査は適用されない。

たとえば、特定の人種が犯人であるという目撃証言に従って、当該の人種を集中的に捜査することは、捜査対象の決定に人種情報が使用されているにもかかわらず、人種に基づく異なる取り扱いとはみなされていない。

容疑者の特徴として人種情報を活用することが人種に基づく異なる取り扱いに該当しないのは、こうした事例では人種に関する一般化が行われていないからである。確かに、ある種の一般化が行われているが、ここでの一般化とは、目撃者の証言は信頼に値する可能性が高いというものである。こうした一般化は、黒人であることを理由として犯罪者である可能性が高いと判断する人種プロファイリングに基づく一般化とは異なるものである。

c. Principles and Application

以上の事例を参照することで、人種に基づく分類の使用が異なる取り扱いに該当しない場合を確定するための指針となる原理が浮かび上がる。1 つ目として、国勢調査の事例が示唆しているように、人種に基づく分類の使用が異なる取り扱いに該当するためには、そうした分類は直接的な効果(a proximate effect)を生み出さなければならない。2 つ目として、容疑者の人種情報の活用事例が示唆しているように、人種が一般化において使用される場合には、人種集団に関する一般化のみが人種に基づく異なる取り扱いに該当する。

前述の居住の安定性のような、人種以外の要因にどれほどの重みを与えるかを決定する際にアルゴリズム内で人種情報を使用する場合、以上の 2 つの特徴は存在していない。まず、犯罪率等の予測をする際にどの要因を考慮するべきかに関して人種情報が使用されているため、人種に基づいた分類は直接的な効果を生み出していない。また、アルゴリズムが用いている一般化も居住の安定性と再犯率の間のものであるため、人種に関する一般化ではない。以上のような構造的な類似性を考慮すると、アルゴリズムにおいて人種情報を使用することは許容されているし、許容すべきであると考える十分な理由が存在する。

3. Ricci's Irrelevance

学者の中には、人種間で異なる効果を及ぼすことを避けるためにアルゴリズムを変更することは、Ricci v. DeStefano 判決によって禁じられていると考えている人々もいる。もしこの認識が正しければ、人種カテゴリーをアルゴリズムに使用することも認められないだろう。しかし、私の考えでは、こうした学者たちは Ricci 判決に過剰な解釈を行っている。その理由について理解するために、当該事件の内容について検討しよう。

コネチカット州ニューヘイブン市の消防局で実施された昇進試験で、試験に合格したマイノリティはごく少数であった。この結果を受け、市はこうした試験が公民権法第七編で禁

じられている十分な理由なく異なる効果を生み出す選抜に該当するとして、試験結果を認証しなかった。しかし、最高裁は、試験結果を認証しないという市の判断自体が、試験に合格した消防士に対する人種に基づく異なる取り扱いに該当するとして、試験結果を認証しないという市の判断を棄却した。一部の学者たちは、この Ricci 判決を、人種間で異なる効果を及ぼすことを避ける意図や、人種を意識すること自体を禁止するものだとして解釈している。

しかし、こうした解釈は Ricci 判決を誤って解釈している。まず、Ricci 判決では特定可能な個人に対する影響が問題となっていたのに対して、アルゴリズムの場合、事前に特定可能な個人に対する影響は問題となっていない。したがって、アルゴリズムの問題を評価するにあたって、Ricci 判決は限定的な価値しか持たない。また、Ricci 判決の解釈に関する論争では、人種間で異なる効果をもたらすという意識自体が厳格審査を惹起するのに十分かどうかが争われたが、こうした意識自体は厳格審査を招くものではない。最後に、人種を意識することは Ricci 判決では禁じられていない。それどころか、判決は、異なる取り扱いかどうかの判定に関して、直接的な効果の重要性を支持するものとなっている。Ricci 判決では、市の判断が特定可能な人々に直接的な効果を与えたという事実が、本判決を他と分ける特徴となっているのである。

Conclusion

本論文では、アルゴリズムの公平性に関する論争に対して 3 つの貢献を行った。1 つ目は概念的な貢献である。等予測率と過誤率の比率に関して、前者は何を信じるべきかという信念の問題に関わるものであり、人々をどのように扱うべきかにかかわる公平性の問題には不適当であると主張した。2 つ目は規範的な貢献である。アルゴリズムによって評価を受ける集団間で公平性を保つためには、偽陽性と偽陰性の比率がそれぞれの集団間で同じである必要がある。3 つ目は法的な貢献である。保護された特性以外の考慮要因を決定する際に、保護された特性をアルゴリズムが活用することで、アルゴリズムの正確性を改善し、等過誤率の欠如が示す不公平性を抑えることができる。こうした仕方で人種情報を活用することは憲法で禁じられていない。