アルゴリズムの公平性についての統計的な基準について

ブライアン・ヘッデン[1]

出典

Hedden, Brian. "On statistical criteria of algorithmic fairness." *Philosophy and Public Affairs* 49.2 (2021).

キーワード

・予測アルゴリズム（predictive algorithm）
・公平性（fairness）
・公平性についての統計的な基準（statistical criteria of fairness）
・基準率（base rates）
・「境界に達していないこと（infra-marginality）」

1. はじめに

予測アルゴリズム（predictive algorithm）が、我々の生活の中でますます大きな役割を果たすようになるに伴い、予測アルゴリズムに不公平やバイアスがあるのではないかという懸念が高まっている。アルゴリズムの公平性に関する最も有名な事例は，再犯の予測に使用される COMPAS（Correctional Offender Management Profiling for Alternative Sanctions）である。COMPAS は、過去の逮捕歴等を基に、再犯のリスク・スコアを付与するツールである。COMPAS は、人種情報を使用していないにもかかわらず、白人よりも黒人の非再犯者に、誤って高い再犯リスクを付与する点から、「黒人に対してバイアスがかかっている（biased against blacks）」と批判を受けた。この批判に対して、COMPAS を提供するノース・ポイント社が、アルゴリズムの予測は白人と黒人に対して同じように正確であると反論し、論争が起こった[2]。

近年、公平性の様々な基準に関して、予測アルゴリズムがそうした基準を満たす仕方で予測を行うことができるかについて、急速に研究が進んでいる。そして、多くの不可能性定理（impossible theorems）が、ごく例外的な場合を除いて、直観的には非常に説得力があるように見える公平性の基準のいくつかに関して、それらを同時に満たすことが不可能である

[1] ブライアン・ヘッデンは、現在オーストラリア国立大学で教鞭をとる哲学者で、合理性や意志決定理論を専門としている。

[2] COMPAS については、本ホームページ記載のデボラ・ヘルマン「アルゴリズムの公平性を測定する」の中にさらに詳しい説明があるので、そちらを参照されたい。

と示している。多くの学者はこの結果を悲観的に解釈して、公平な予測は不可能でありモラル・ジレンマは避けられないと考えている。

私は、これまで議論されてきた公平性についての統計的な基準のうち、較正（calibration）以外の基準は、予測アルゴリズムが公平性を満たしているかどうかの必要条件ではないと主張する。その理由は、較正以外の他のすべての基準は、公平であることが疑い得ないような予測アルゴリズムでも満たすことができないからだ。さらに、驚くべきことに、関連する集団間で基準率が等しい場合であっても（*even when base rates are equal*）、公平なアルゴリズムはそうした基準を満たすことができない可能性がある。ある予測アルゴリズムが（較正以外の）基準のいずれか（あるいはすべてを）を満たすことができない場合、こうした事実は、当該のアルゴリズムが不公平である、あるいは、バイアスがかかっていることを示す一見自明な（*prima facie*）証拠になるという主張と、私の議論の結論は両立する。しかし、実際には、こうした基準を満たしていないことは、アルゴリズムが不公平であることを意味する、あるいは、不公平を含意するわけではないのだ。

2. 公平性の基準と不可能性定理

本論で検討していく予測アルゴリズムは、既知の特徴に基づいて、ある個人がポジティブかネガティブ、2 つのクラスのいずれかに分類されるのかを予測することを目的とする。ここでは、議論を簡単にするために、これらのリスク・スコアが区間[0, 1]に収まると仮定し、またリスク・スコアとは、個人がポジティブ・クラスに分類される確率と解釈する。加えて、[訳者注：連続的な値をとる]リスク・スコアとバイナリーな予測を行うアルゴリズムについて検討していく。

私がここで関心を持っているのは主に予測（*predictive*）アルゴリズムであり、その予測に基づいて行われる可能性のある決定（*decisions*）については副次的な関心しかもたないと強調しておきたい。予測アルゴリズムが完璧に公平であっても、その予測が明らかに悪意のある用途に使用される場合もある。この予測と決定の区別は重要であり、3 節で再び取り上げる。

本論の焦点は公平性に置かれているが、私が注目したいのは、あるアルゴリズムが個人に対して不公平であるか、あるいは、集団に対して不公平であるかではない。そうではなく、あるアルゴリズムが、特定の集団に所属していることを理由として（*in virtue of their membership in a certain group*）、個人に対して不公平であるかに注目したい。

この公平性についての考え方は、他の公平性についての考え方とどう違うのか。特定の集団への所属という観点からは不公平でなくとも、個人に対して不公平であることはあり得る。たとえば、面接中に目を合わせないといった、集団への所属に関係しない理由によって不利な扱いを受ける場合がそうである。また、特定の集団に所属していることを理由として、

個人に対して不公平であったとしても、当該の集団に対して不公平ではないこともあり得る。たとえば、ある個人に対して人種等を理由として不利な扱いをしつつも、同時に当該の人種集団に全体としては便益となる施策を実施する場合がそうである。

どのようにすれば予測アルゴリズムが不公平であるか判断できるか。ここでは、「公平性についての統計的な基準（statistical criteria of fairness）」と私が呼ぶものについて検討していきたい。アルゴリズムの公平性の必要条件として提案されてきた候補には、以下の 11 の統計的な基準が存在する。まず、すべての基準について述べた後で、こうした基準を採用する動機づけについて説明したい。

連続的な値をとるリスク・スコア関する公平性についての統計的な基準

(1)集団内での較正（Calibration Within Group）：あらゆる可能なリスク・スコアに関して、当該のリスク・スコアを割り当てられた、実際にポジティブである個人の（期待される）割合は、関連するそれぞれの集団で同じであり、またその割合はリスク・スコアに等しい。

(2)ポジティブ・クラスのバランス（Balance for the Positive Class）：実際にポジティブである個人に割り当てられた（期待される）平均リスク・スコアは、関連するそれぞれの集団で同じである。

(3)ネガティブ・クラスのバランス（Balance for the Negative Class）：実際にネガティブである個人に割り当てられた（期待される）平均リスク・スコアは、関連するそれぞれの集団で同じである。

バイナリーな予測に関する公平性についての統計的な基準

(4)均等な偽陽性率（Equal False Positive Rate）：誤ってポジティブと予測されたが実際にはネガティブである個人の割合が、関連するそれぞれの集団で同じである。

（5）均等な偽陰性率（Equal False Negative Rate）：誤ってネガティブと予測されたが実際にはポジティブである個人の割合が、関連するそれぞれの集団で同じである。

（6）均等な陽性予測値（Equal Positive Predictive Value）：実際にポジティブである個人のうち、ポジティブであると予測された個人の（期待される）割合が、関連するそれぞれの集団で同じである。

（7）均等な陰性予測値（Equal Negative Predictive Value）：実際にネガティブである個人のう

ち、ネガティブであると予測された個人の（期待される）割合が、関連するそれぞれの集団で同じである。

（8）偽陽性率と偽陰性率の比率の均等性（Equal Ratios of False Positive Rate to False Negative Rate）：偽陽性率と偽陰性率の（期待される）比率が、関連するそれぞれの集団で同じである。

（9）全体の過誤率の均等性（Equal Overall Error Rates）：偽陽性と偽陰性の数を母数で割った（期待される）数が、関連するそれぞれの集団で同じである。

（10）統計的な同等性（Statistical Parity）：ポジティブと予測される個人の（期待される）割合が、関連するそれぞれの集団で同じである。

（11）予測される陽性者と実際の陽性者の比率が等しいこと（Equal Ratios of Predicted Positives to Actual Positives）：ポジティブと予測される個人の数を、実際にポジティブとなった個人の数で割った（期待される）数が、関連するそれぞれの集団で同じである。

それぞれの基準について述べたところで、その背景にある動機について簡単な概要を与えたい。まずは、(1)（6）（7）から。基準（1）の背景にあるのは、公平性は、与えられたリスク・スコアが関連するそれぞれの集団で「同じことを意味する（mean the same thing）」という考えだ。(6）と（7）は、(1）をバイナリーな予測に一般化したものだ。

次は、(2）(3）(4）(5）である。基準（4）と（5）には、次のような考え方が含まれている。つまり、異なる集団に属していながらも同じ行動をとる個人に対して、ポジティブと予測されるか、あるいは、ネガティブと予測されるかの点で、アルゴリズムは彼らを平均して同じように扱うべきだという考え方である。たとえば、ある集団の実際にネガティブである個人が、他の集団の実際にネガティブなメンバーと比較して、高い確率でポジティブになると予測される傾向があるとしたら、不公平だろう。基準（2）と（3）は、基準（4）と（5）をバイナリーな予測の場合からリスク・スコアの場合に一般化したものと見ることができる。

基準（8）の背後にあるのは、公平性は、それぞれの集団に対して、偽陽性率と偽陰性という 2 つの主要な過誤に等しい相対的な重みを割り当てる必要があるという考えである。基準（9）は、他の集団と比較した時、ある集団に対して、単純にアルゴリズムの精度が低いことは不公平であるという考えを取り入れている。

ここで、ポジティブという予測が、融資を受ける等の有益な結果に対応していると考えると、基準（10）を動機づけるのは以下のような考え方である。つまり、公平性は、各集団からの応募者の同じ割合がそうした有益な結果を享受することを要請するという考え方であ

る。しかし基準（10）は、集団間での基準率の違いを無視することを理由として、退けられることが多い。実際、基準率が異なる時、完璧な予測が可能なアルゴリズムでも、基準（10）を満たすことができないのである。そして、基準率が違うからといって、完璧な予測が可能なアルゴリズムが不公平なものになるわけではないだろう。この点で、基準（11）の方がすぐれている。基準率が等しい場合には、基準（11）は基準（10）を満たすことを含意している。しかし、基準率が異なる場合、基準（11）は基準率の違いに対応して、個人がポジティブと予測される比率が異なることを求める。

これらの基準は、アルゴリズムの予測と実際の結果が、それぞれの集団で同じであることを要請している。これらの基準はそれぞれ動機が異なり、ある基準は他の基準よりも魅力的である。しかし、これらの基準はいずれも、予測アルゴリズムが公平である、あるいは、バイアスがかかっていないための必要条件の候補として考えられてきた。

しかし、一連の不可能性定理によって示されているように、ごく例外的な場合を除いて、これらの基準をすべて満たすことは不可能である。たとえば、Kleinberg et al. (2016)は、（A）関連する集団間で基準率が等しい、（B）アルゴリズムが完璧な予測を行う（すべての実際にポジティブな人にリスク・スコア 1 を割り当て、すべての実際にネガティブな人にリスク・スコア 0 を割り当てる）といういずれかの条件が満たされない限り、（1）、（2）、（3）を満たすアルゴリズムが存在しないことを証明した[3]。（A）（B）に関する同じ条件の下で、（4）、（5）（6）を満たすアルゴリズムが存在しないことを、Chouldechova (2017)が示している[4]。さらに、Miconi (2017)は、（A）（B）に関する同じ条件の下で、（α）（4）と（5）、（β）（6）と（7）、（γ）（11）の中から、（α）（β）（γ）を同時に一つ以上満たすアルゴリズムが存在しないことを示した[5]。（8）、（9）、（10）を対象とした、これまで公表された不可能性定理は存在しない。だが、基準率が等しくない場合には（9）と（10）を同時に満たすことが不可能であること、同様に、各集団に偽陽性・偽陰性が存在しない場合を除いて、（8）と（10）を同時に満たすことが不可能であるとは容易に理解できる。

まとめると、これらの不可能性定理の結果は、直観的には魅力的に映る公平性についての統計的な基準の多くが、ごく例外的な場合を除いて、同時に満たすことができないことを示している点で、印象的である。これらの結果は、公平性のジレンマが避けられないことを示していると解釈することもできる。あるいは、これらの統計的な基準はすべて、アルゴリズムが公平である、あるいは、バイアスがかかっていないための必要条件ではないと解釈する

---

[3] Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores." *arXiv preprint arXiv:1609.05807* (2016).

[4] Chouldechova, Alexandra. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big data* 5.2 (2017): 153-163.

[5] Miconi, Thomas. "The impossibility of" fairness": a generalized impossibility result for decisions." *arXiv preprint arXiv:1707.01195* (2017).

こともできる。では、どの基準が公平性の真の条件なのか。

3. 人々、コイン、部屋

ある基準がアルゴリズムの公平性の真の必要条件であるかどうかを判断する一つの方法は、完璧に公平なアルゴリズムを見つけ、このアルゴリズムが当該の基準に反することが可能かどうかを確認することである。もし反するのならば、その基準はアルゴリズムの公平性の真の必要条件ではない。この方法は現実には実施困難だが、コイントスを用いて考察を進めることができる。

様々な偏りを持つコインの束があると想定してみよう。各人にはランダムにコインが割り当てられる。そして、彼らはA、Bという2つの部屋のどちらかにランダムに割り当てられる。我々の目的は、各人について、その人の持つコインが表になるか、それとも、裏かどちらかを予測することである。それぞれのコインには偏りがラベル付けされており、区間[0, 1]内の実数がその偏り、つまりコインが表となる客観的な確率を示している。

これは、完全に公平で偏りのない予測アルゴリズムである。各人について、ラベルに「x」と書かれていれば、その人にはリスク・スコアxが割り当てられる。$x > 0.5$ の場合は、その人はポジティブであるという、バイナリーな予測が行われる(ここでは、「0.5」と書かれたコインはないと仮定する)。

このアルゴリズムは完全に公平でバイアスがかかっていない。とりわけ、特定の部屋に所属していることを理由として、個人に対して不公平であるわけではない。加えて、単に特定の部屋への所属を理由として不公平でないことにとどまらず、このアルゴリズムにはどこにも不公平なところはなく、これ以上ない最適な仕方で個人の持つコインの表裏の予測を行っている。

ここで、アルゴリズムの予測が行為（レジュメ制作者注：第2節で言及した、予測と区別された決定等に関わる側面のこと）と切り離されているのならば、当該のアルゴリズムが不公平でないのは、トリビアルな意味に過ぎないと批判されるかもしれない。第一の反論として、この批判は、私が提案したコインの表裏を予測するアルゴリズムが公平であること、そして、当該のアルゴリズムが満たすことのできない統計的な基準が公平性にとって必要ではないことを認めている。第二に、近年の認識的正当化における道徳の侵入（moral encroachment）や認識的不正義（epistemic injustice）の議論が示唆しているように、認識論的に特有の危害や不正義が存在する可能性がある。たとえば、人種等を理由として目撃者の証言を無視することには、それが問題のある行動に結びつかなくても、道徳的な問題が存在しているように見える。第三に、リスク・スコア等に応じて、個人に様々な便益が与えられると、私の例に変更を加えることも可能だろう。しかし、たとえこのような変更を加えたとしても、集団への所属を理由に、個人を異なる仕方で扱っているという意味での不公平性は、依然として存在していないのである。

ここで得られる重要な洞察は以下の通りである。この設定でも、(1)集団内の較正は満たされている一方で、(2)から(11)の他の基準はすべて、私が考案した公平なアルゴリズムでは満たすことができない場合がある。さらに、(2)から(11)の基準は、同時に、また、たとえ2つの部屋の基準率が同じであっても（*simultaneously* and *even when base rates are equal between the two rooms*）満たすことができない場合がある。アルゴリズムの公平性に関する議論の多くは、不公平の原因を基準率が等しくない事実に求めてきたため、この洞察は重要である。

以上を確認するために、以下の事例を考えてみよう。A室には「0.75」とラベルが張られたコインを持った人が 12 人、「0.125」とラベルが張られたコインを持った人が 8 人いる。前者は全員リスク・スコア 0.75 が割り当てられ、コインが表の人と予測されるが、実際にはコインが表の人はそのうち 9 人である。後者にはリスク・スコア 0.125 が割り当てられ、コインが裏の人になると予測され、実際にはコインが表の人はそのうち 1 人である。B室には「0.6」とラベルが張られたコインを持った人が 10 人、「0.4」ラベルが張られたコインを持った人が 10 人いる。 前者には全員リスク・スコア 0.6 が割り当てられ、コインが表の人と予測されるが、実際にはコインが表の人はそのうち 6 人である。後者にはリスク・スコア 0.4 が割り当てられ、コインが裏の人になると予測されるが、実際にはコインが表の人はそのうち 4 人である。ここで基準率が同じであることに注意されたい。各部屋ともに、20 人のうちコインが表の人は 10 人である。

(2)から(11)の統計的な基準について検討していこう。A室で実際にコインが表である人（ポジティブ）に割り当てられたリスク・スコアの平均は（9×0.75＋1×0.125）/10＝0.6875である。一方で、B室で実際にコインが表である人に割り当てられたリスク・スコアの平均は（6×0.6＋4×0.4）/10＝0.52 である。これは(2)ポジティブ・クラスのバランスに反している。(3) ネガティブ・クラスのバランスについて同様の計算を行うと、A室は 0.3125、B室は 0.48 である。よって、(3) ネガティブ・クラスのバランスにも反している。

A室の偽陽性率は 3/10 で、B室の偽陽性率は 4/10 なので、(4)均等な偽陽性率に反している。A室の偽陰性率は 1/10、B室の偽陰性率は 4/10 なので、(5)均等な偽陰性率にも反している。A室の陽性予測値は 3/4 であるのに対し、B室の陽性予測値は 3/5 であるため、(6)均等な陽性予測値に違反している。A 室の陰性予測値は 7/8、B 室の陰性予測値は 3/5 なので、(7)均等な陰性予測値に違反している。A室の偽陽性率と偽陰性率の比は 3 であるのに対して、部屋 B のそれは 1 であるため、(8)偽陽性率と偽陰性率の比率の均等性に反する。A室の全体の過誤率は 4/20 で、B室は 8/20 なので、(9)全体の過誤率の均等性に反する。コインが表だと予測される割合は、A 室の場合は 12/20 で、B 室の場合は 10/20 であるため、(10) 統計的な同等性に反する。そして、A室のコインが表であると予測される人数と実際にコインが表である人数の比率が 12/10 であるのに対し、B室は 10/10 であり、(11) 予測される陽性者と実際の陽性者の比率が等しいことにも反している。

予測アルゴリズムが、ある部屋への所属を理由として、個人に対して不公平であった、あ

るいは、バイアスがかかっていたということをこれらの事実が示していない点は明らかなはずだ。アルゴリズムは、完全に公平なのである。問題があるのは、集団内での較正を除いた統計的な基準の方である。

以上の議論は、コイントスの性質に依存したものではない。再犯率等のお好みの行動について取り上げてみよう。そして、対象が持つある特徴に基づいてアルゴリズムが予測を行うとしよう。さらに、このアルゴリズムは、アルゴリズムがリスク・スコアxを付与する個人について、個人が行動を行う確率はxであるという意味で、完全に較正されていると想定してみよう。そして、これらの個人をランダムにA室とB室に割り当てるものと想定しよう。コイントスの例と同じように、この場合、個人の間での特徴の分布等によって、集団内での較正以外のあらゆる統計的な基準を、予測アルゴリズムが満たすことができない恐れがある。予測の基礎となる特徴について、どの個人のどんな特徴を対象とするかに関して不公平が存在するかもしれないものの、部屋への所属を理由として（*in virtue of their room membership*）、アルゴリズムは個人に対して不公平ではないのである。以上の議論は、私の議論が、コイントスがもつ偶然性に依拠したものではないことを示している。たとえ、予測を行う対象が偶然に委ねられたものではない、あるいは、個人が予測の基礎となる特徴を保有することになるプロセスが偶然に委ねられものでなかったとしても、予測アルゴリズムは、集団内での較正以外のすべての統計的な基準を満たすことができない可能性がある。

ここで、私の議論が限定的なものであることを強調しておきたい。私は、人、コイン、部屋のケースが現実的であり、COMPASのようなケースと完全に類似していると主張しているわけではない。私の例では、部屋への所属は、社会的に構築されたものでもなければ、歴史的な抑圧の基礎になったものでもない。対照的に、人種やジェンダー等は、少なくとも部分的には社会的に構築されており、社会的な抑圧の基礎になってきたものである。加えて、人間の行動は通常コイントスと同じ仕方で、偶然に委ねられたものではない。最後に、私の例では、2つの集団は同じ基準率を持つもかかわらず、基礎的なリスク分布が大きく異なる。対照的に、現実のケースでは、基準率が等しいグループは、少なくとも類似した基礎的なリスク分布を持っている。

しかし、私の議論は、私の用いる例が現実的なものであるかどうかに依拠したものではない。第一に、単純化や理想化は厄介な複雑な要因を排除して問題を明確することに役立つ。第二に、より重要な点だが、私は上記の統計的な基準のいずれもが（集団内での較正を除いて）公平性に必要（*necessary*）ではないと主張しているだけだ。そして、ある基準が公平性に必要ではないと結論づけるには、公平性は満たされているが基準を満たすことができないケースが1つあればよいのである。

4. 境界性と証拠

上記で検討された統計的な基準の多くは、Ayres(2002)[6]や Corbett-Davies and Goel(2018)[7]によって導入された、いわゆる「境界に達していないこと（infra-marginality)」と関連する理由から必要条件であることに失敗している。彼らによれば、アルゴリズムが公平であるためには、関連する集団間で境界事例を同じように扱わなければならない。たとえば、犯罪に関与している疑いが最も低いアフリカ系アメリカ人は、犯罪に関与している疑いが最も低い白人と同じ程度に疑われるべきである。しかし、上記の基準の多くは、非境界、あるいは境界に達していないケースにも関わる。

現実には、どのケースを境界事例と考えるべきかには、常に議論の余地がある。しかし、私の例では、コインの偏りが 0.5 付近にある人々が境界事例であり、コインの偏りが 0.5 より遠く離れた人々が、非境界事例に該当する。私の例では、A 室の人を B 室の人とまったく同じように扱い，コインの偏りに合わせてリスク・スコアを割り当て，それが 0.5 より大きい場合に限り表を予測した。しかし、このアルゴリズムは、集団内での較正を除いたすべての基準に反していた。これは、A 室の人々が、比較的「明らか(clear)」、あるいは、非境界事例である一方で、B 室の人々が、比較的明らかでない、あるいは、境界事例であるからだ。

これまで、私は集団内での較正を支持する積極的な論拠を与えてこなかった。ただ、この基準は、直観的に説得的があり、この基準を支持する動機付けは明瞭である。もし何らかのアルゴリズムがこの基準に反する場合、同一のリスク・スコアが、2 つの集団で異なる証拠上の効力を持つことを意味する。つまり、ある個人がリスク・スコアを付与されたとき、その個人がポジティブであるという確率は、その個人が属する集団によって左右されることになる。だが、これは、異なる集団への所属によって、個人を異なる仕方で扱うことのようにみえる。

たとえ、集団内での較正が公平性のための必要条件であっても、十分条件ではないだろう。というのも、おそらく、保護された集団に対して盲目である[8]等の、公平性のための他の条件が存在するからだ。

私が必要条件（*necessary*）であることを否定してきた、一部、ないしはすべての基準が、かかる基準に違反した際には、不公平性の一見自明な(*prima facie*)証拠を提供するという主張と、私が以上で得た結論は両立し得る。たとえば、(2)から(11)の基準の違反は、集団への所属を根拠としてアルゴリズムが予測を行っている等の証拠になりうるかもしれない。しかし、この証拠がどれほど強力なものであるかは、他の必要条件が何である等の想定に依

---

[6] Ayres, Ian. "Outcome tests of racial disparities in police practices." *Justice research and Policy* 4.1-2 (2002): 131-142.

[7] Corbett-Davies, Sam, and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning." *arXiv preprint arXiv:1808.00023* (2018).

[8] 性別や人種などの差別的取り扱いの原因になりうる区分のこと。

拠しているのである。

5. 含意

最後に、概念的な論点と実践的な論点について指摘して締めくくりたい。概念的な論点は次の通りである。予測アルゴリズムが、不正とみなされるような結果を生み出す決定のために用いられたとしよう。この場合でも、集団への所属を理由として、当該のアルゴリズムが、個人に対して不公平であったり、バイアスがかかっているということを意味するわけではない。不公平さやバイアスは、社会の背景的条件等の、アルゴリズム外の要素に起因している可能性がある。実践的な論点は次の通りである。上述の結果として、不公平さやバイアスへの最善の応答は、予測アルゴリズム自体を修正するのではなく、社会の背景的条件等のアルゴリズム外の要素への介入になりうるということだ。

われわれには、公平で正確な予測に加えて、正しい判断と社会全体の正しさという複数の目的がある。そして、これらの複数の目的を達成するために、予測アルゴリズム自体に過剰な責任を負わせるべきではない。もちろん、予測アルゴリズムが公平で正確な予測という目的を達成することを確実なものにすべきである。しかし、可能なかぎり、その他の目的を達成するために、アルゴリズム外の要因にも追加の介入を行うべきである。

しかし、アルゴリズム内・外の介入によって不公平さに対処することは、政治的な理由をはじめとして、様々な理由から実現不可能かもしれない。第一に、補償問題が好例だが、不平等や歴史的な不正義を矯正するために最も効果的な政策手段は、余りに論争的なため、実現可能性が低いかもしれない。第二に、拘束力を持たない私企業のように、往々にして、予測アルゴリズムを設計する担当者には、アルゴリズム外の介入を行う権限が与えられていない。第三に、予測アルゴリズムの使用によって避けがたい負の副作用が生じる可能性がある。たとえば、予測アルゴリズムそれ自体は公平であったとしても、予測アルゴリズムの使用はある集団に対する有害なステレオタイプを強化するかもしれない。

アルゴリズムの使用が一部の不利な立場にある集団に有害な結果をもたらし、アルゴリズム外の要因への介入によって、この有害な結果を緩和することが不可能な場合でも、こうした事実は、必ずしも予測アルゴリズム自体が不公平であることを意味しない。むしろ、予測アルゴリズム自体の公平性が問題のすべてではないという事実へ注意を促すものである。

以上の洞察は、ここでは簡単に触れることしかできないが、いくつかの重要で難しい問題を提起している。第一に、これらの有害な結果をどのように考えるべきか。こうした有害な結果は、ある種の不公平となるのか、それとも、不公平ではなく、不公平とは異なる点で道徳的に問題であるのか。

たとえこれらの有害な結果が不公平の一種とみなされたとしても、私がこれまで退けてきた統計的な基準について再検討する理由にはならないだろう。というのも、ある統計的な基準の違反が有害な結果をもたらすかどうかは、偽陽性等の診断を有害、ないしは有益な影

響に結び付ける「収益構造（payoff structure）」に依拠しているからだ。そしてこの収益構造は、予測アルゴリズムがいかに決定を下す際に使用されるのか、また、社会の背景的条件等によって変化するものである。したがって、統計的な基準はいずれも、そうした統計的な基準に違反することで、必然的にある集団やその成員に有害な影響をもたらすものではない。ゆえに、統計的な基準はいずれも、公平性にとって必要不可欠なものとはならないのである。

第二に、このような有害な結果が不公平であるかどうかに関わらず、こうした有害な結果に対して、アルゴリズムの運用を担当する者はどのように対応すべきか。アルゴリズム自体が公平であることを保証すれば、その結果を考慮しないことが許されるのか。それとも、有害な結果を緩和するために、本質的に公平なアルゴリズムを放棄または修正することが求められる場合もあるのか。このような問題に取り組むことは、この本論の範囲を超えているだろう。しかし、少なくとも、それ自体は公平なアルゴリズムが、やむを得ず有害な結果をもたらすような方法で使用されているいくつかのケースでは、当該のアルゴリズムの使用を断念する十分な理由が存在することは明らかだと、私は考える。

これまでの議論を要約すると、おそらく、集団内での較正を除いて、どの統計的な基準も、予測アルゴリズムの公平性の必要条件ではないと私は主張してきた。しかし、実際に予測アルゴリズムをどのように設計すべきかは、アルゴリズム自体の公平性以外の要因にも依拠している。ある場合には、アルゴリズム外で適切な介入を行いつつ、単にアルゴリズムの公平性を確保するだけで、われわれが望む結果を得ることができるかもしれない。しかし、他の場合では、分配をはじめとした様々な結果を達成するように、アルゴリズムを設計しなければならない。しかしながら、その方法は、倫理的な配慮と、単純な公式に還元できない複雑で多面的な経験的要素の両方に依拠したものとなるだろう。

（福家佑亮）