アルゴリズムによる差別にかんする予備的検討
──公平性基準に着目して

前田 春香

1. はじめに

本稿の目的は，アルゴリズム[1]に適用される数理的な公平の基準である公平性基準の欠点をまとめた上で，差別の規範理論によるその問い直しおよび補填を提案することである．

近年，機械学習や AI 技術の広範な使用に伴い，アルゴリズムは技術的論点のみならず社会的・倫理的な論点を含むさまざまな問題を提起している．人や人のデータを対象としたアルゴリズムに顕著にみられる問題として、次のようなものがある．2016 年，質問紙調査をもとに計算をおこなう再犯予測システムが黒人に不利なリスク判定をおこなったことが明らかになった（Angwin et al. 2016）．また，顔認識の正確な認証には人種や性別ごとに差があることや（Buolamwini 2017），Google の検索エンジンは”black girl”と検索すると過剰に性的なアウトプットを返すこと，などが確認されている（Noble 2018）．このような現象はアルゴリズムによる差別[2]とよばれている．

多くの倫理指針が公平ならびに反差別（anti-discrimination）をポリシーとして明記しているが（IEEE 2017; 人工知能学会　2017; OECD 2019），何が差別かは明記されていない．さらに，倫理指針も差別に対する研究も欧米が先行しているため，倫理指針における差別の事例は，無自覚のうちに文化的価値観が忍び込んだ結果，差別とされている可能性がある．商品としてのアルゴリズムがいとも容易く国境を超えることを考えると，異なる文化圏では，差別であったものが境界事例であるとされたり，差別でないとさえ判断されたりするかもしれない．翻って日本では，差別に対抗できる法律が男女雇用機会均等法と障害者差別解消法であるため，性別や障碍を理由とするもの以外の差別の防止は各団体の自助努力に任されることになる．このような状況下では，何が差別か

ということを，ある程度文化的偏りから独立して，理論的に判断できる基準が必要だと考えられる．

アルゴリズムによる差別に対処しようとする代表的な基準として，機械学習界隈で使用される，アルゴリズムのアウトプットに対して用いられる公平性基準を挙げることができる．本稿では公平性基準に着目し，2 節は偏りを明らかにしようとする公平性基準の解説とそれ自体の限界，また基準を適用するときの限界について説明する．次の 3 節では，偏りの中で悪質な差別を見分けることが必要であることを述べる．4 節では，その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．

2. 公平性基準の問題

公平性基準とは，数理的に公平な状態や偏っている状態を定義し，差別的な状況の指摘に役立てようというものである．通常用いられる公平性基準には三種類があり[3]，しかもそれぞれが固有の統計的限界を有する（Goel et al. 2018）．その三種類とは，反分類（anti-classification），キャリブレーション（calibration），分類の等しさ（classification parity）である．以後この節では，公平性基準における各手法の説明のため，以下の仮想事例を使用する．

［人種によってアウトプットに偏りのあるアルゴリズム］

住宅ローンの貸与可否の決定補助に使用されるリスク評価アルゴリズムがある．当該アルゴリズムは人種，職種，収入などを独立変数[4]とし，ローンが返済されないリスクを「可」または「否」のかたちで従属変数[5]として導く．なお，人種によってローンのリスクには偏りがあることがわかっており，具体的には白人以外が高リスク，白人が低リスクとされる傾向にある．その結果，白人以外は住宅の購入が難しくなっている．

反分類（anti-classification）とは，保護属性との関連が問題になりやすいことを踏まえて，保護属性やその代理変数（proxy）を使用しないという手法である[6]．保護属性とは，人種や性別といった，アウトプットと何らかの関連がある場

合に問題になりやすい属性のことをいう．代理変数とは，ある属性との関連性があまりに強いために，当該の変数が実質的にある属性を指し示すようなものをさす．

反分類という手法は，上の事例を次のように解決しようとする．すなわち，保護属性に当たる人種以外の属性である収入などを独立変数として従属変数である貸与リスクを導くことになる．

反分類の手法は次のような現象を防ぐことができる．人種とローン貸与のリスクの関係を調べたところ, 人種とローン貸与のリスクに関連があったとする．そのとき特定の人種だからローンのリスクが高いと考えたくなるかもしれないが，この理解は正しくない．これは擬似相関と呼ばれる現象で，収入という変数が人種とローン貸与リスクに関連があるため，人種とローン貸与リスクにも関連があるようにみえてしまう現象である．反分類は保護属性を使用しないことによって，このような誤りを防ごうとしているとみることができる．

しかし，反分類で全ての誤認を防げるわけではない．擬似相関は実際には関係のないものを関係があると誤認することだが，実際には関係があるものを関係がないと誤認することもありうる．後者の誤認は本稿との関連で最も有害なものの一つである．しかもこの誤認は，反分類が要求する通りに人種変数を採用せず，ローン貸与リスクと収入の関連を調べたときに起こる．というのは，人種と富裕であるかどうか（＝ローンを返済できるかどうか）の結びつきが強固である場合，この反分類という手法を使用しても隠れた関連性によって結果的に偏ってしまうことが想定されるのである．この場合，経済状況を計測するものとしての収入が人種の代理変数として働いていることになる．これが意味することは，もしいかなる（保護）属性との関連も認めず，そのデータも取らないということであれば，アウトプットが偏っているかどうかさえ明らかにすることができないということである．このように反分類という手法は，現実の認識に際して重大な障壁をもたらす可能性があるのだ．

このようなわれわれの誤認を防ぐ一つの方法は，属性ごとに及ぼされた影響を計測し，その影響が偏っていないかという観点で帰結をみることである．帰結主義的な手法には，キャリブレーションと分類の等しさがある．本節ではキャリブレーションについて説明する．

キャリブレーションとは，見積もられたリスクを統制しても保護属性とアウトプットに関連がないことである（Corbett-Davies and Goel 2018 p. 2）．統計学における統制とは，当該アウトプットにかかわらない変数すべての影響を除去することをいう. すなわちここでキャリブレーションがおこなう手法としては，予測されたローン貸与リスクと人種や性別を始めとした保護属性に一切関連をもたせることなく，保護属性以外の変数にローン貸与リスクと関連をもたせることを意味する．結果として，黒人と白人は同じ確率でローンを貸与されることになる. この場合, 予測されたリスクと人種は関連をもたないといえるので，表面的には公平であり，誰も差別されていない状態が実現する，というわけである.

しかし果たして, 保護属性とアウトプットのいかなる関連をも許さない場合，当該データ活用は有用になりうるだろうか．これは反分類のときと同じような問題点である．つまり，保護属性以外の独立変数とリスク判定のアウトプットに関連をもたせたところで，保護属性とリスク判定のアウトプットに他の（保護属性以外の）変数を通して関連が存在するならば，結局人種による偏りが起こりうると考えられるからである．このように，アウトプットに考慮したからといって直ちに問題がなくなるというわけではない.

さらにはより深刻な点として，当該の基準を満たしたまま，差別行為をおこなうことが可能だということが指摘されている（Corbett-Davies and Goel 2018）. 例えば融資の可否を判断されるときに，返済できない可能性が高い高リスク集団と，無事に返済できる可能性が高い低リスク集団のデフォルト率を同率に調整し，リスクを同等にすることを考える．これは一見公平にみえるが，低リスク集団への融資を制限するために，意図的に収入などの個人的な要因を無視して, 住所などの情報と関連をもたせることが考えられる（Corbett-Davies and Goel 2018 p. 1）．ここで住所が問題になるのは，安い土地には経済力の低い集団が住み，高い土地には経済力の高い集団が住んでおり，経済力と人種が関係しているときに住所は人種の代理指標となりうるからだ．その結果，比較的リスクの高い地域に住んでいる少数派への融資を拒否するために，当該の判断を使用することができるのである．これはレッドライニング現象として広く知られている慣行である [7].

確かにこの手法は帰結を考慮には入れているが，その考慮の入れ方には問題があるといえる．この他，別のやり方で帰結を考慮に入れる基準がある．次に取り上げるのは，グループ間の偽陽性の差に着目する基準である分類の等しさである．

分類の等しさとは，同じ粒度の保護属性および保護属性でない属性の間で偽陽性および偽陰性の割合が同じであるべきだというものである[8]．実際に A であり，なおかつ A と予測される場合を真陽性，実際に¬A であり，なおかつ¬A と予測される場合を真陰性とする．この場合，実際には A であるのに¬A と予測される場合が偽陰性，実際には¬A であるのに A と予測される場合が偽陽性にあたる．この偽陰性と偽陽性が同程度になればよいため，例えば，上記のローン貸与においては肌の色が暗い人の偽陽性と肌の色が明るい人の偽陰性が同率になることなどがこの基準における公平であると考えられる．しかし，一部の文脈では，正確性と分類の等しさによる公平性を両立できないことが証明されている（Kleinberg et al. 2017）．もともとの集団で対象にしている確率にそもそも差がある場合（ここではローンの返済率），そのデータを用いて予測をしてもやはり人種間格差がみられるのは当然というわけだ（Simoiu et al. 2017）．したがって，このアウトプットは差別的なのかという問いを提起することになる．

その時々で適切な基準を使い分ければ良いではないかという反論があるかもしれない．この反論に答えるため，以下では公平性基準そのものが有する実用上の限界が深刻な論争をもたらした，COMPAS というアルゴリズムの事例を参照したい．

COMPAS とは，Correctional Offender Management Profiling for Alternative Sanctions を正式名称とする再犯率予測ソフトウェアで，量刑判断や仮釈放判断，リハビリテーションプログラムの決定の補助に使用されている．COMPAS は，非営利機関 Propublica によって偽陰性・偽陽性の人種間格差を告発されて以来（Angwin et al. 2016），大きな問題としてみなされるようになった．Propublica は，再犯予測リスクが算出された人々の追跡調査を 2 年間にわたっておこない，肌の色が暗い人が再犯をすると誤って予測される確率が肌の色が明るい人のそれよりも少なくとも 2 倍高く，肌の色の明るい人は肌の色の暗い人よりも再犯

リスクが低いと誤ってラベル付けされる確率が高かったと結論づけた（Angwin et al. 2016; Larson et al. 2016）．一方 COMPAS の開発者である Northpointe 社（現 Equitable 社）は，キャリブレーションの基準に基づいてその告発に応答している（Dieterich et al. 2016）．つまり，人種にかかわらず同率で，再犯の予測が的中しているというわけである．

この公平性の基準はすべて相互に独立したもので，すべてを同時にみたすことが要請されているわけではない．しかし，第一に，分類の等しさとキャリブレーションを同時にみたすことができないことがあることはまず実用上の問題になりうる（Kleinberg et al. 2017; 神嶌・小宮山　2018）．そのことが明らかになったのが当該の論争であった．というのは，Propublica と Northpointe 社のこの論争は，どちらが間違っているというものではなく，参照している公平性基準が違うために起こっていることがわかっているのである（Sumpter 2019 pp. 83–92）．それだけに当該のソフトウェアに基づくこの対立は，公平性基準にとどまらない多数の論争を招いた．現在もこのソフトウェアは新しいバージョンを開発中である．

第二に，これが数理的に避けられない現象だからである．先述のように，異なる集団間であまりに異なる確率を有している（例えば肌の色の明るい人と肌の色の暗い人の再犯率があまりに違いすぎる）場合には，そのままデータを用いた場合当然の帰結として格差のある予測アウトプットが算出されるだろう（Ayres 2002 など）．よって，その他の基準や対応策が必要だと考えられる．

第三に, 第二の理由からしてこれは一回限りの不幸な事故ではなく, COMPAS で勃発したような論争が今後も起こりうるということが挙げられる．したがって，何が差別にあたるのかという問いは解決できずに残り続けることになるのだ．

3. 公平性基準をよりよくする方法──問い直しの提案

前節では，公平性基準が数理的に偏りを発見しようとする方法であることを指摘し，考えられる問題点を述べた．さらに，公平性基準を活用したときに問題が生じた具体的な事例として COMPAS の事例を挙げ，今後もこのような現

象が起こりうる可能性について言及した.

このような事態に際してわれわれには何ができるのだろうか．今後差別を指摘し防止するために，われわれが取りうる方法は三つ挙げられるだろう [8]．以下，折衷主義，どの基準がもっとも良いか決定する，根本的に枠組みを問い直す，の順に検討する.

一つ目は，折衷主義（pluralism）をとることである．この場合の折衷主義とは，分類の等しさなどの基準と，キャリブレーションに基づく基準を両立させながら，適切なときに適切なものを使う立場，というようにまとめられる．しかし，Long はこの立場を，不均等な誤認識の確率とキャリブレーションが両立するのみならず，アウトプットとなっている確率自体がキャリブレーションを必要としてしまうため, 問題を避ける立場であるとして批判している（Long 2020 p. 4）．また先述したように，そもそもの材料にしている確率（base rate）に差異がある場合（例えば肌の色が明るい人と肌の色が暗い人の再犯率がそもそも違う場合），アウトプットも偏るこの問題は infra-marginarity 問題と呼ばれ，当然起こるという意味で自然なことである（Simoiu et al. 2017）．先述の COMPAS の論争はまさにこの立場によって引き起こされたものであるといえ，したがって問題の防止に役に立つとはいえない.

とすれば二つ目に挙げられる方法は，どの基準が一番優れているかを決定することであろう．そもそも公平性基準における公平と，哲学における公平は，双方とも fairness という単語で示されるが, 意味内容は異なっている（Binns 2017 p. 3）．したがって，双方がどれほど合致しているかは公平性基準における公平に重要な示唆を提供するといえる．Hellman（2020）はこの立場に立って，哲学の観点から，偽陽性・偽陰性に基づく基準，すなわちここでいう分類の等しさとキャリブレーションを比較して，偽陽性と偽陰性を同時に考慮する基準が政治哲学のいう公平に沿っており，したがって優れていると述べる.

しかし，この公平性基準という土俵に乗る限り，数値だけで判断することの欠点を抱え込むことになるだろう．これは社会的な概念を数理的にのみ定義しようとして, 十分に落とし込みそこねる罠だという指摘がある（Selbst et al. 2019 pp.61−62）．先述した Hellman（2020 pp.820–833）は偽陰性と偽陽性を同時に利用することが有効であると述べながら，一方で，公平性基準は信念にかかわ

るものでないため，公平をはかる手段としては不適切であるとも論じている．

これをうけて三番目に提案することができるのは，数値によって公平かどうかをみる公平性基準という枠組そのものを問い直すことである．以下，差別を指摘するという目的に照らしたときの公平性基準という枠組みそのものの欠点について述べる．

第一には，そもそも現行の公平性基準は一般的な帰結主義と比べ，ごく一部しか帰結として考慮していないことが挙げられる．そのごく一部とは，アルゴリズムの当該の意思決定に直接影響を受ける人々である（Card and Smith 2020）．一度の自動の意思決定で直接影響を受ける人々は，全世界のうち微々たる人数であるし，また，時間としても一瞬にあたる．しかし，例えばサブプライムローン問題は長期的な影響を無視した政策によって引き起こされたものであるし，特定の人種に対して害を与えたものであった．このことをみれば，短期的な限られた情報だけを考慮するのでは不十分だといえるだろう（Card and Smith 2020 p. 8）．

第二に，当該の短期的な限られた情報は，偏ったデータから生まれたものかもしれない．第一の理由は直接のアウトプットの影響だけをみるという意味で限定的であったが，この理由はインプットされるデータのみを考慮することに由来する．これまでのアルゴリズム研究では，偏ったデータはそもそも偏った行動によって産出されたものであるため，偏ったデータを使用した意思決定は格差の再生産になるというものが主流であった（Barocas and Selbst 2016; O’Neil 2016 など）．この発想の前提にあるのは，偏ったデータがアルゴリズムに入力されるはるかに前から，歴史的に行動が偏ってきたという事実である（Noble 2018）．

確かに偏ったアウトプットの原因は，一部にはデータの偏りにある．Crawford（2016）が指摘するように，白人のデータセットしかない場合には，その属性をもたない他の人種のデータセットが少なくなり，それがプロダクトのアウトプットの偏りとなって表れるのだ．しかし，もしデータセットに偏りがなかったとしても，ただちに問題がなくなるわけではない．データセットに偏りがなくアルゴリズムに偏りがなくても，アウトプットに偏りがあることは考えられる．例えば男女同数のデータを取得し，その職業を調査する場合，女性には専業主婦が関連付けられ，男性にはエンジニアが関連付けられるかもしれない．この

とき，何か間違っているのだろうか．間違っているとすれば，いったい何が間違っているのだろうか．なお人種と経済的水準をはじめとするさまざまな要素には緊密な関係があるため，それらを反映してしまうことは少なくない．さらには，Long（2020）が指摘するように，どのようにキャリブレートするのが正しいのかという問いが再びやってくるだろう．そのとき公平性基準は，われわれが何をすべきかを教えてはくれないのだ（Hellman 2020）．結局，過去の歴史をうけて今後いかに修正するかという問題に帰着することになる．

第三に，文脈の中で生じる意味が見落とされている．ここでいう意味とは，データや事実にたいして付される価値判断のことである．データの偏りがなく，アルゴリズムに偏りもなく，アウトプットにも偏りがない状態を想定しよう．偏りがあったとしても，その意味内容からして当該の偏り自体には道徳的に問題がないケースである．人間の子どもは成人するにともなって頭の大きさが成長するため、年齢によって必要な帽子の大きさはもちろん異なるであろう．このような場合には一見問題がないようにみえるかもしれない．しかし，それでもなお十分ではない．われわれは道徳的に問題のないデータに基づいて差別をすることができる．現代では悪名高い骨相学は，人の性格が頭蓋骨の形に現れるというものであった．この発想は知性にも適用され，特定の人種は知性の面で劣るとされた（Gould 1996）．頭蓋骨の周囲の長さというデータそのものに道徳的悪質さは明らかにないが，それにわれわれは何らかの意味付けをし，人を貶めるために使用することができる[9]．Memmi が指摘するように，重要なのは差異ではなく，差異にどのような意味が与えられるかなのである（Memmi 1994）．であるならば，人間が差異に見出す意味を数値だけで捉えることはできない．同じような偏りであっても，文脈によってわれわれが見出す意味は様々に変化するだろう．この点で，数理的な基準のうちどれが一番よいかを決定することは，意味を見落とす可能性を孕む．

人間が差異に見出す意味の対象は数値にとどまらない．というよりも，その対象はむしろ数値でないほうが一般的である．近年の機械学習の進歩によって，特に目覚ましい発展があったのは画像認識の分野であることを考えると[10]，画像に付与される解釈が人間のそれと異なっており，当該の解釈そのものが問題を生みうる．その顕著な例が，Google Photo が肌の色が暗い人をゴリラだとタ

グ付けした問題である（Zhang 2015）．いくらアルゴリズムからみて数値的に類似しているとしても，このタグ付けという行為そのものがいかに差別的であるか，多くの人々は説明なく理解できるであろう．アルゴリズムが与えた解釈は，明らかに不適切なものだとわかるからである．たとえ画像認識にかかわる問題であっても，どれほどの人数を正確に認識できたかというように数値化できるならば公平性基準でも扱うことができるが，このようにタグ付けをするという動作そのものに意味が付与されるようなケースは難しい．このような問題を公平性基準は扱うことができないのである．

人間がアルゴリズムを使用しておこなう差別と違って，Google Photo の例はアルゴリズムが直接差別をしているようにみえる．このようなアルゴリズムによる差別は製作者の意図したもので，したがって製作者に責任があるという反論があるかもしれない．しかし，アルゴリズムによる差別は人間[11]がまったく差別することを意図していない場合もある．しかも，もちろんアルゴリズムそのものも意図をもたないので，どちらにせよ，意図によって悪質な差別であるかどうかを区別することはできない．また，製作者に差別的な意図がなかったとしても，われわれはアルゴリズムによる差別を道徳的に悪質だと感ずることがある．したがって意図によらず，数理的な偏りだけに頼るのでもなく，何が差別なのかを問い続けることが必要である．

本稿では遠い帰結を考慮できないこと，意味を考慮できないことを現行の公平性基準の欠点として述べた．差別の観点からすると意味の問題はより差し迫ったものと考えられる．例えば Google の検索システムに”black girl”と入力すると，過剰に性的客体化されたアウトプットをするという報告がある（Noble 2018）．多くの人がこの現象の悪質さを認識するだろうが，この場合その悪質さというのは，先述した Google Photo の事例と同様関連付けられた意味に由来していると考えることができる．いかにアルゴリズムからして似ているようにみえていても，われわれにとっては似ていない場合，もしくは類似していても道徳的に問題がある場合が考えられる．単にアルゴリズムからみたその人と，われわれからみたその人が似ていないことと，その人に対して道徳的に問題がある見方がされていることは異なる．したがって，偏りがある状態と偏りに道徳的に問題がある状態は分けられなければならない．正確に認識されない問題の中にも，

とりわけ問題になるような認識のされ方が存在するのだ.

これらの理由から，遠い帰結や意味を含んだより広い視野から，アルゴリズムによる差別を研究することが求められているといえよう．データの偏りが他の情報と結び付けられて差別になることもある．一方で，データに偏りがあるだけでは一見して悪質であるにすぎない．ではそのデータの偏りは本当に悪質であるのか，いつどのように悪質でありうるか，その悪質さがどのように形作られるか，もしくはわれわれが悪質であると意味付けるにすぎないのか，目的によって悪質さが変化するのか（Barabas et al. 2018），などといった問いにはまだ答えられていないのだ.

これらは「もし予測が正確である場合には，対象者に対してのどのようなふるまいが許されるか」という問いにつながる．もちろん，道徳的に悪質な差別は正当化されない．しかし，確かに差異付けがアルゴリズムの本来の機能であることを踏まえ（Benjamin 2018），公平性基準よりも考慮する範囲を広げて，どのような差異処遇や偏りがどこまで正当化されるかを考える必要があるといえよう.

4. 差別の規範理論の導入

3 節では，公平性基準がアルゴリズムによる差別に対して十分に有効であると考えられない状況下で何をどうすべきかについて検討した．その結果，どのような差異処遇や偏りが道徳的に正当化されるのかを考える必要があるという結論に到達した．次の問題は，どのような差異処遇や偏りが道徳的に正当化されうるかを，何によって明らかにするのがよいかである.

ここでいう差異処遇とは，属性によって異なる処遇のうち道徳的に悪質でないものをさす．女性を女子トイレに行かせ，男性を男子トイレに行かせることは（トランスジェンダーなどの人を除いて）道徳的に悪質でない差異処遇である一方で，能力が同等である男女に異なる給与を与え，なおかつ当該の給与が女性の方が低いのは道徳的に悪質な差異処遇であり，すなわち差別である．前者は，保護属性によって処遇を変えているという意味で偏ってはいるが，これを道徳的に悪質だと断ずる人はいないだろう.

男女でトイレを分けることと，男女で違う給与を支給することはどう違うのだろうか．差別の規範理論は，この問いに対して二通りの説明を与える．一つ目は，道徳的に悪質な差異処遇である後者は，前者と違って被差別者（ここでは女性）に害を与えるから悪質だというものである（Knight 2013; Lippert-Rasmussen 2013）．二つ目は，被差別者の尊厳を尊重していないから悪質だというものである（Hellman 2008 など）．人間がアルゴリズムを使って差別をするという立場に立つ場合には，偏りが害をもたらす場合に悪質であるため公平性基準で考慮できる一部のアルゴリズムによる差別に対応することがわかるだろう．さらに，後者の説によれば，前節で指摘した意味がなぜ重要な意味を持つのかに理論的な基礎づけを与え，アルゴリズムによる差別にたいするわれわれの直観をよりよく説明することができる．

以下のような状況下でわれわれがすべてのステレオタイプを直ちに道徳的な悪と断ずるかを考えることは，データに対する価値判断と個人の尊厳の関連を多少なりと明らかにすることに役立つだろう．ある集団 A が実際になんらかの傾向性をもっているものとする．そして，当該の集団の外にいる集団 B がそれを観測して，集団 B の価値判断に関連する集団 A にたいするイメージ，すなわちステレオタイプをもったとする．

このとき，3 節の観点からいうと，全てのステレオタイプが道徳的に悪質であるわけではない（Hellman 2005 など）．日本人が長時間働くという傾向性から，日本人はまじめだ，ひいては日本人はそのことによってよい人々であると考えるとき，それは悪質でないステレオタイプといえる．このことに異論はないであろう．人間がステレオタイプを他の人間に適用する場合，いろいろなパターンが考えられるが，3 節によれば論点をさしあたり 1 つとすることができる．

重要なのは，この論点はアルゴリズムによる差別であっても有効だと考えられることである．われわれは，アルゴリズムによる差別は人間によるそれと性質が違っても，悪いと感じる傾向にある．幾多の事例が炎上し告発されていることはその証左であるといえよう．このような事例を問題だと感じる人にとって，例えばアルゴリズムによって差別的な仕方でおこなわれた関連付けの内容は，ただの文字列や画像ではないのだ．データに付与される価値判断を考慮に

入れることは，そのような直観に沿った説明を可能にする．

差別の規範理論の規範理論を参照する意義は以下の通りである．第一に，今まで発見できなかった，もしくは単発的なものとして理解されている差別事例を悪質な差別の事例の一つとして体系化することができる．アルゴリズム研究はこれまで，アルゴリズムによる差別によって格差が広がることに目を向けてきたが（Barocas and Selbst 2016 など），差別の規範理論によれば，格差を広げること以外による悪質さがありうることが示唆されている．公平性基準にはあてはまらないが，広くアルゴリズムによる差別として認識され，話題になっている事例があるのは先述のとおりである．アルゴリズムによる差別の問題として捉えるならば，まず散発的な事例を一括した上で考察することが必要であろう．

第二に，当該理論が政治哲学における平等主義を背景にしたものであることが挙げられる．差別の理論は他にも多数あるが，平等主義を背景にしていることは差別の規範理論の大きな特徴である．近年アルゴリズム研究と公平を一つのテーマとしてきた政治哲学が接近しているが（Binns 2017; Long 2020; Hellman 2020 など），何がよいかではなく何が悪いかという視点から，望ましいアルゴリズムやその使い方を指摘できる可能性がある．

5. おわりに

本稿では，現状のアルゴリズムによる差別に対処しようとする公平性基準が不完全であり，その基礎づけやその欠点を補うものとして差別の規範理論が有効であることを指摘した．具体的には，1 節で社会的背景を説明し，2 節で一般に使用されている公平性基準に関しての検討をおこなった．3 節でアルゴリズムによる差別に適用しようとする際の公平性基準の欠点を説明し，意味を考慮できないことが欠点となりうることを指摘した．4 節ではそのような観点から，公平性基準を補うような望ましい道具立てとしての差別の規範理論と，差別の規範理論を参照する意義について述べた．

アルゴリズムによる差別の問題や，反差別をめざすべきであることは各所で周知が図られている．公平性の定義について拡大が必要であることは指摘され

ていたが（Whittaker et al. 2018），アルゴリズムによる差別の悪質さを考えるためにどのような道具立てが有効といえるかの指摘はなかった．本稿はアルゴリズムによる差異処遇が悪質であるかどうかを判断することにつながる．

しかし，残された課題もある．本稿では公平性基準の欠点として，意味などを含めた総合的な視点が十分でないことを指摘した．だが，例えばデータサイエンスなどの関連分野でどれほど意味が考慮されるのかは明らかではない．次に，何が差別かを指摘するために差別の規範理論が有効であると示した．ただし差別の規範理論は一般に人間が人間を差別する場合を想定しているため，アルゴリズムが差別的なアウトプットをしている事例にそのまま適用することが可能かということは当然問題とされるべきである．例えば，アルゴリズムの差別的なアウトプットが，製作者の意図や予測を超えて観察された場合にはどうなるのか．これは，誰（もしくは何）に差別の悪さが帰されるのかという議論につながるであろう．

暫定的な結論としては，われわれのアルゴリズムによる差別にたいする直観を適切に反映するには，差別の規範理論による問い直しが有効であるといえよう．

註

1. アルゴリズムとは，本稿では機械学習を含む，含まないによらず，自動化された意思決定システムをさす（Koene et al. 2018 p. 39）．
2. 本稿におけるアルゴリズムによる差別とは，アルゴリズムが直接人を差別しているようにみえる現象をさす．また，差別とは道徳的に悪質な差異処遇をさす．道徳的に悪質でない差異処遇は差異処遇と表記する．詳しくは 4 節を参照．
3. ここで対象にしているのは集団公平性である．集団公平性とは，人種や性別といった，その属性によってアウトプットに偏りがあると問題になりやすい保護属性をもった人々と，保護属性をもっていない人々の間で差異がないことをさす．
4. 独立変数とは，y=f(x)としたときの変数 x のこと．原因をさす．
5. 従属変数とは，y=f(x)としたときの変数 y のこと．
6. Hellman（2020）によれば，保護属性を使用しなければ良いという広範な誤解がある．保護属性を使用しなければ良いというこの発想は，この節で述べるように正しくない．
7. レッドライニングは現在のアメリカ法では違法である．しかし 2015 年，ハドソンシティ貯蓄銀行が，肌の色が暗い人とヒスパニック系の人々に対しての住宅ローンを意図的に避けていたことなどにたいし，司法省および消費者金融保護局と 3300 万ドルで和解している．

8. 他にグループ間の真陽性の差に着目するものとして，equality of oppotunity が，グループ間の偽陰性の差に着目するものとして demographic parity がある．
9. Whittaker ら（2018）は，システム自体が公正であっても不公正（な目的のため）に使用できることを指摘している．
10. 機械学習の手法は，大きく分けて人間が当該画像が何であるかをラベルという形で教える教師あり学習，ラベルをつけずにデータを学習させる教師なし学習，良い・悪いのフィードバックのみをおこなう強化学習に大別される. 2016年, 人間の入力なく Google のアルゴリズムが自動で猫を認識することに成功したが，これが画期的だったのは，猫の画像と「これが猫である」と示すラベルが与えられなかったにもかかわらず，アルゴリズムが自動で猫だと認識したためである．
11. 製作の段階でも使用の段階でもさまざまな人間が関与しうるが，この場合の人間は製作者をさす.

参考文献

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). “Machine Bias”, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Retrieved November 30, 2020).

Ayres, I. (2002). “Outcome Tests of Racial Disparities in Police Practices”, in *Justice Research and Policy*, 4(1–2), pp. 131–142.

Barabas, C., Dinakar, K., Ito, J., Virza, M. and Zittrain, J. (2018). “Intervention over Predictions: Reframing the Ethical Debate for Actual Risk Assessment”, in *Journal of Machine Learning Research*, 81, pp. 1–15.

Barocas, S. and Selbst, A.D. (2016). “Big Data’s Disparate Impact”, in *California Law Review,* 104, pp. 671–732.

Benjamin, J.J. (2018). “Complex Intentions: A Methodology for Contemporary Design Practice”, in *Proceeding of the 2018 Designing Interactive Systems Conference*, pp. 347–350.

Binns, R. (2017). “Fareness in Machine Learning: Lessons from Political Philosophy”, in *Journal of Machine Learning Research*, 81, pp. 1–11.

Buolamwini, J.A. (2017). “Gender Shades: Intersectional Phenotypic and Demographic Evaluation of Face Datasets and Gender Classifiers”, Massachusetts Institute of Technology.

Card, D., and Smith, N.A. (2020). “On Consequentalism and Fairness”, in *Frontiers in Artificial Intelligence,* 3(4), pp. 1–11.

Corbett-Davies, S. and Goel, S. (2018). “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning”, *arXiv*,1808.00023v2[cs.CY].

Crawford, K. (2016). “Opinion | Artificial Intelligence's White Guy Problem”, *The New York Times*. https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html (Retrieved November 19, 2020).

Dieterich, W., Mendoza, C. and Brennan, T. (2016). “COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity”, http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Goel, S., Shroff, R., Skeem, J.L and Slobogin, C. (2018). “The Accuracy, Equity, and Jurisprudence of Criminal Risk Assessment”, in *SSRN Electronic Journal*, pp.1–21.

Gould, S.J. (1996). *The Mismeasure of Man*, W. W. Norton & Company.

Hellman, D. (2005). “Racial Profiling and the Meaning of Racial Categories”, in *Contemporary Debates in Applied Ethics*, Andrew I. Cohen and Christpher H.

Wellman (eds.), Wiley-Blackwell, pp. 232-244.

——— (2008). *When is Discriminarion Wrong?*, Harvard University Press.

——— (2020). “Measuring Algorithmic Fairness”, in *Virginia Law Review,* 106(4), pp. 811–866.

人工知能学会. (2017). “人工知能学会倫理指針”, 人工知能学会, http://ai-elsi.org/wp-content/uploads/2017/02/%E4%BA%BA%E5%B7%A5%E7%9F%A5%E8%83%BD%E5%AD%A6%E4%BC%9A%E5%80%AB%E7%90%86%E6%8C%87%E9%87%9D.pdf.

神嶌敏弘, 小宮山純平. (2019). 「機械学習・データマイニングにおける公平性」『人工知能』, 34(2), 196–204 頁.

Kleinberg, J., Mullainaha, S. and Raghavan M. (2017). “Inherent Trade-Offs in the Fair Determination of Risk Scores”, in *Proceedings 8th Conference on Innovation in Theoretical Computer Science*, pp. 1–23.

Knight, C. (2013). “The Injustice of Discrimination”, in *South African Journal of Philosophy*, *32*(1), pp. 47–59.

Koene, A., Dorthwaite, L., and Seth, S. (2018). “IEEE P7003™ standard for algorithmic bias considerations”, IEEE. http://fairware.cs.umass.edu/papers/Koene.pdf

Larson, J., Mattu, S., Kirchner, L. and Angwin, J. (2016). “How We Analysed the COMPAS Redicisism Algorithm”, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (Retrieved November 30, 2020).

Lippert-Rasmussen, K. (2013). *Born Free and Equal?: A Philosophical Inquiry into the Nature of Discrimination.* Oxford University Press.

Long, R. (2020). “Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness”, *arXiv,*2007.02890 [cs.CY］

Memmi, A. (1994). *Le Racisme*, Gallimard.

Noble, S.U. (2018). *Algorithms of Oppression,* New York University Press.

IEEE. (2017). “Ethically Alined Design”, https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

OECD. (2019). “Ethics Guidelines for Trustworthy AI”, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

O’Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threaten Democracy*, Crown.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S. and Vertesi, J. (2019). “Fairness and Abstruction in Sociotechnical Systems”, in *FAT 19' :Proceedings of the Fairness, Accountability and Transparency*, pp. 59–68.

Simoiu, C., Corbett-Davies, S. and Goel, S. (2017). “The problem of Infra−marginality in outcome tests for discrimination”, in *Annals of Applied Statistics*, *11*(3), pp. 1193−1216.

Sumpter, D. (2018). *Outnumbered from Facebook and Google to Fake News and Filter-Bubbles: The Algorithms That Control Our Lives*, Bloomsbury .（千葉敏生・橋本篤史訳，2019 年，『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）

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R. and Schwartz, O. (2018). “AI Now Report 2018”, https://ainowinstitute.org/AI_Now_2018_Report.pdf

Zhang, M. (2015). “Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software”, *the Forbes*. https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/#616ed5e4713d (Retrieved September 10, 2019).