The 37th Annual Conference of the Japanese Society for Artificial Intelligence, 2023

形式的公平性規準間の不可能性に関する考察

On the Impossibility between Formal Fairness Criteria

神嶌 敏弘 *1
Toshihiro Kamishima

*1産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

We first enumerate the formal fairness criteria that can be mathematically available, and show which criteria cannot be satisfied simultaneously. We discuss these criteria are mutually exclusive, even though the condition of these criteria is relaxed.

1. はじめに

機械学習技術は与信，採用，保険などの，個人の生活に大きく影響を与える決定にも利用されるようになった．そのため，社会的公平性を考慮しつつ予測する公平性配慮型機械学習 [神嶌 19, 神嶌 22] という技術が研究されている．これは，公平性規準を満たす制約の下で予測を行うものであるが，この公平性規準には多種多様なものが提案されている．そこで，本稿では，数理的にどれだけの公平性規準があり得るのかについて考察したあと，これらの規準間の関係について論じる．

2. 形式的公平性

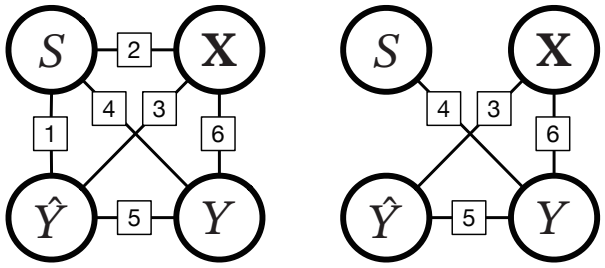
特徴と目的変数についていくつかの記号を定義したあと，公平性配慮型機械学習における公平性規準の考え方を示す．

2.1 準備

確率変数 S と $\mathbf{X}$ は，それぞれセンシティブ特徴 (sensitive feature) と非センシティブ特徴（もしくは，単に特徴；non-sensitive feature）を表す．公平性配慮型機械学習では，センシティブ特徴の表す性質に対して公平性を保証しつつ予測する．例えば，与信，採用，保険などの決定について扱うとき，社会的公平性の観点からその関与を排除すべき対象者の性別や人種といった個人属性情報を，このセンシティブ情報とする．なお，このセンシティブ特徴に何を設定するかは，データマイニングで扱うタスクと，法や規制などの社会的環境を考慮して与えるものとする．一方の非センシティブ特徴 $\mathbf{X}$ は，対象を表す特徴の中で，上記のセンシティブ特徴以外の全てを含む特徴ベクトルである．確率変数 Y は目的変数で，与信・採用・保険などの決定を表し，分析者はこの変数の表す内容に関心がある．この Y については，さらに結果の予測値を $\hat{Y}$，観測値を Y として区別する．形式的公平性 (formal fairness) とは，センシティブ特徴 S の目的変数の予測値 $\hat{Y}$ に与える影響が公平であるかを，数理的に定式化したものである．

2.2 形式的公平性の種類

公平性については，倫理学・法学・政治学や経済学で論じられてきたが，ここでは前者の文献 [Lippert-Rasmussen 06] に基づいて考察する．Discrimination という単語は『差別』と悪い意味に訳されることが多い．しかしながら，元は単なる区別であり，この区別が悪いものとなるのは，害ベース説 (harm-based account) と尊厳ベース説 (disrespect-based account) の二つの有力な説がある．害ベース説では，その区別が被差別者に損害を与えるので悪くなり，尊厳ベース説では，被差別者に対する軽蔑 がありそれが道徳的に不快なものであるため悪くなるというものである．前者は功利主義的，後者は義務論的な正義と関連付けられる．両者は，誰かを賞賛すると，暗黙的に他者に対する軽蔑を示すので尊厳ベース説では悪いが，特に他者に損害を与えてはいないので害ベース説では悪くはないといった違いがある．



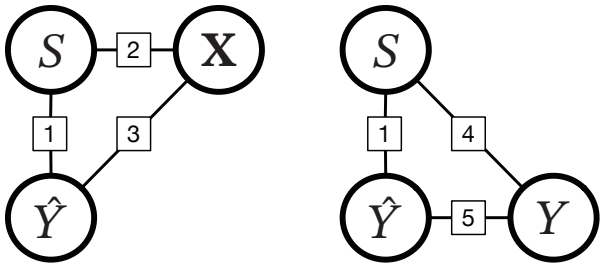
(a) 全連結の場合　(b) 辺 1 と 2 を削除した場合

図 1: 無向グラフィカルモデルによる連関ベース公平性の表現

公平性配慮型機械学習では，害ベース説に基づく差別を扱う．理由の一つは，統計技術を導入することが侮蔑的であると尊厳ベース説では定義することさえ可能で，もしこの定義を採用するのであれば対処は不可能であることである．もう一つは，司法の判断は，少数派と多数派の採用率などの定量的な差異を根拠とする害ベース説のものが多く，それに応えるべく，公平性配慮型機械学習は研究されているからである．

害ベース説は，何と比べたとき損害が生じたかによって二つに分かれる．一つは，少数派が不利に判定された現実に起こったことに対して，それがもし多数派であればという現実には起きなかった反事実と比べる．これは反実仮想公平性（もしくは，反事実公平性; counterfactual fairness) [Kusner 17] として定式化がなされている．もう一つは，ある理想的な配分と比較して損害を受けているかどうかを見るもので，連関ベース公平性（association-based fairness）と呼ばれている．次章では，この連関ベース公平性に着目し，どのような規準が考えられるかを論じる．

連絡先: ホームページ https://www.kamishima.net/



(a) S，$\hat{Y}$，および $\mathbf{X}$ の場合　　(b) S，$\hat{Y}$，および Y の場合

図 2: 3 変数の無向グラフィカルモデル

3. 連関ベース公平性規準

図 1 は，変数 S，$\mathbf{X}$，$\hat{Y}$，および Y の同時分布を，無向グラフィカルモデル（もしくは，マルコフ確率場）[Bishop 08, 8 章] で表現したものである．図 4. の全連結の場合では，どの変数対の間にも直接的な依存関係がある．図 1(b) のように，全連結から 1 と 2 番の辺を取り除くと，同時分布は次式のようになる．

$$\Pr[S, \hat{Y}, \mathbf{X}, Y] = f_1(S, Y) f_2(\mathbf{X}, \hat{Y}, Y) / Z \tag{1}$$

ただし，$f_i(\cdot)$ はポテンシャル関数，Z は正規化定数である．図 1(b) からは，$S \perp\!\!\!\perp (\mathbf{X}, \hat{Y}) \mid Y$ の条件付き独立性が成立することが読み取れる．さらに，この条件付き独立性から，$S \perp\!\!\!\perp \hat{Y} \mid Y$ と $S \perp\!\!\!\perp \mathbf{X} \mid Y$ の条件付き独立性が導出できる．

それでは，この無向グラフィカルによる表現を使って，2 変数，3 変数，そして 4 変数の場合について，どのような公平性規準が考えられるかを順に考察する．

3.1 2 変数の場合

変数 S と $\hat{Y}$ は必ず公平性規準には含まれなければならないので，これら 2 変数の関係を考える．なぜなら，形式的公平性規準はセンシティブ特徴 S の予測結果 $\hat{Y}$ への影響を扱うからである．変数が二つだけの場合は，これらが従属か独立かの 2 通りしかない．従属の場合は，センシティブ特徴が直接的に結果に影響しているため直接差別 (direct discrimination) [Pedreschi 08] や disparate treatment [Feldman 15] と呼ばれる不公平な状態に該当する．独立の場合は，間接差別 (indirect discrimination) [Pedreschi 08] や disparate impact [Feldman 15] も解消された状態であり，数理的には統計的均一性 (statistical parity) [Dwork 12] と呼ばれている規準である．

公平性とは無関係に，現実的ではない 2 変数間の独立性を挙げておく．もしこれらの独立性が 3 変数以上の場合の同時分布から導出されるなら，その分布は現実にはあり得ないものとなる．$S \perp\!\!\!\perp Y$ は過去の決定が全くセンシティブ特徴に依存しないというもので，差別的な決定が過去に全くなかったことを示しており，公平性配慮型を導入する必要はなくなるため，成立していないと仮定する．$\hat{Y} \perp\!\!\!\perp \mathbf{X}$ と $\mathbf{X} \perp\!\!\!\perp Y$ は，対象の性質を全く考慮されず，決定が無作為に行われたことを示しており，現実的ではない．$\hat{Y} \perp\!\!\!\perp Y$ も，これからの予測が，過去の決定と全く無関係に行われることを示しており非現実的である．$S \perp\!\!\!\perp \mathbf{X}$ はどちらも制御できない変数であり，現実的に成立することはないとする．よって，以上 5 通りの 2 変数間の独立性が導出できるような同時分布は現実にはありえないものと仮定する．

3.2 3 変数の場合

3 変数の場合は，S と $\hat{Y}$ は除外できないため，考えられる $(S, \hat{Y}, \mathbf{X})$ と $(S, \hat{Y}, Y)$ の二つの場合について順に考察する．

3.2.1 S，$\hat{Y}$，および $\mathbf{X}$ の場合

$(S, \hat{Y}, \mathbf{X})$ の場合は，図 2(a) のグラフィカルモデルに相当する．1〜3 番の辺それぞれを削除した場合と，他の 2 辺を削除した場合を考えて，全部で 6 通りの場合が考えられる．1 番を削除すると，$S \perp\!\!\!\perp \hat{Y} \mid \mathbf{X}$ が成立する．これは，センシティブ特徴を取り除いてモデルを訓練する，すなわち目的変数は S とは独立で，$\mathbf{X}$ のみに依存させて，$\Pr[\hat{Y}|\mathbf{X}] = \Pr[\hat{Y}|S, \mathbf{X}]$ とすることで達成できる．これは（センシティブ特徴の）無視による公平性 (fairness through unawareness) [Dwork 12] にあたる．また，センシティブ特徴以外が同じ対象は同じ決定を受けるという treating like cases alike の原則を達成する個人公平性 (individual fairness) もこの条件付き独立性にあたる．すなわち，$\mathrm{Dom}(\mathbf{X})$ の全ての個人について，S と $\hat{Y}$ が文脈依存独立性 [Boutiller 96] が成立する

$$S \perp\!\!\!\perp \hat{Y} \mid \mathbf{X} = \mathbf{x},\ \forall \mathbf{x} \in \mathrm{Dom}(\mathbf{X}) \tag{2}$$

という条件と同値である．

次に 2 番を除外すると，$S \perp\!\!\!\perp \mathbf{X} \mid \hat{Y}$ が成立する．これは同じ予測がなされたときに，その対象の特徴はセンシティブ特徴によらず同じであるということであり，公平な状態といえよう．倫理学に対応する概念もなく，また公平性機械学習の文脈でも論じられたことはないと推察される．高次元の変数 $\mathbf{X}$ の独立性を評価する必要があるため，評価が難しく実用的ではない．同じ特徴の対象に同じ予測をする個人公平性とは逆向きなので，逆個人公平性と呼んでおく．

3 番を除外した $\mathbf{X} \perp\!\!\!\perp \hat{Y} \mid S$ も，非センシティブ特徴の情報は S を通じてのみ結果に影響している．そのため $\mathbf{X}$ から S を推定し，その推定 S に基づいて予測するという極めて不公平な状態といえる．1〜3 番の辺それぞれについて，その辺以外の 2 辺を削除すると，$(S, \hat{Y}) \perp\!\!\!\perp \mathbf{X}$，$(S, \mathbf{X}) \perp\!\!\!\perp \hat{Y}$，および $(\mathbf{X}, \hat{Y}) \perp\!\!\!\perp S$ の独立性を表す．これらの独立性からは，$S \perp\!\!\!\perp \mathbf{X}$ か $\hat{Y} \perp\!\!\!\perp \mathbf{X}$ の条件が導出されるため，現実にはあり得ない．

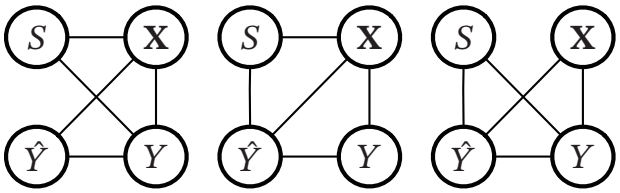
以上の議論から，$(S, \hat{Y}, \mathbf{X})$ の 3 変数の場合の公平性規準は，個人公平性（無視による公平性）と逆個人公平性となる．

3.2.2 S，$\hat{Y}$，および Y の場合

$(S, \hat{Y}, Y)$ の 3 変数の場合は，図 2(b) のグラフィカルモデルで表現でき，やはり 6 通りの場合が考えられる．まず辺 1 を削除した $S \perp\!\!\!\perp \hat{Y} \mid Y$ は均等オッズ (equalized odds) [Hardt 16] は呼ばれている．観測された決定それぞれについて，その観測と同じ予測がなされる割合が全てのセンシティブ特徴の値について等しいというものである．決定が 2 値変数のときは偽正率と真正率がセンシティブ特徴の値によらず等しいという条件にあたるもので，ProPublica が再犯率予測スコア COMPAS が不公平であると主張したときの規準 [Angwin 16] として著名である．

次に辺 4 を削除した $S \perp\!\!\!\perp Y \mid \hat{Y}$ の条件は十分性 (sufficiency) と呼ばれている．予測値 $\hat{Y}$ を知ったあとで，実際にそれと同じ値が観測される割合が，センシティブ特徴によらず同じであるというものである．上記の ProPublica の指摘に対して，裁判所は均等オッズではなく，この十分性を満たすように COMPAS スコアは設計されていると反論 [Flores 16] している．

最後の 5 番を削除した $\hat{Y} \perp\!\!\!\perp Y \mid S$ は，観測された決定の情報のうち S に依存したもののみを通じてのみ予測をするので，極めて不公平といえる．さらに 2 辺を削除した三つのグラフィ



(a) $S \perp\!\!\!\perp \hat{Y} \mid (\mathbf{X}, Y)$ (b) $S \perp\!\!\!\perp Y \mid (\mathbf{X}, \hat{Y})$ (c) $S \perp\!\!\!\perp \mathbf{X} \mid (Y, \hat{Y})$

図 3: 4 変数の無向グラフィカルモデルの抜粋

カルモデルは，$\hat{Y} \perp\!\!\!\perp Y$ か $S \perp\!\!\!\perp Y$ のいずれかが導出されるため，仮定により除外する．

以上の議論から，$(S, \hat{Y}, Y)$ の 3 変数の場合の公平性規準は，均等オッズと十分性の二つとなる．

3.3 4 変数の場合

4 変数の場合は，全部で $2^6 - 2 = 62$ 通りあるが，これら全てについて検討した．3.1 節の現実的でない 2 変数間の独立性と，3.2 節で不公平と考えられた $\mathbf{X} \perp\!\!\!\perp \hat{Y} \mid S$ と $\hat{Y} \perp\!\!\!\perp Y \mid S$ が導出される場合を除くと 23 通りの場合が残る．そのうち，個人公平性と均等オッズが成り立つ場合が 6 通りずつ，逆個人公平性と十分性が成り立つ場合が 5 通りずつ，どの公平性も成り立たないものが 5 通りあった．なお，個人公平性と均等オッズが同時成立する場合が 2 通り，逆個人公平性と十分性が同時成立する場合が 2 通り存在しており，これら二つの場合のみ二つの公平性規準が同時に満たせることが分かった．

これらの同時分布から，図 3 の三つを取り上げる．図 3(a) で示した条件 $S \perp\!\!\!\perp \hat{Y} \mid (\mathbf{X}, Y)$ は，個人公平性と均等オッズが同時に成立しており，個人均等オッズと呼べるだろう．図 3(b) は，十分性 $S \perp\!\!\!\perp Y \mid \hat{Y}$ の条件部を個人との同時分布にした $S \perp\!\!\!\perp Y \mid (\mathbf{X}, \hat{Y})$ であるため，個人十分性と呼べるだろう．ここで注目すべきは，前者は均等オッズか個人公平性の条件が満たされれば一般に成立し，後者も十分性が満たされれば一般に満たされる．ここで一般にと述べたのは，特殊な独立性が成立していと成り立たない場合があるためである．例えば，図 3(a) で，$\mathbf{X}$ から Y への辺 6 を削除した場合，すなわち過去の決定がセンシティブ変数によってのみ行われる極めて不公平な場合である．このとき均等オッズと個人公平性を共に満たされるが，条件 $S \perp\!\!\!\perp \hat{Y} \mid (\mathbf{X}, Y)$ は満たされない．

図 3(c) は，逆個人公平性 $S \perp\!\!\!\perp \mathbf{X} \mid \hat{Y}$ の条件部に Y を加えた $S \perp\!\!\!\perp \mathbf{X} \mid (Y, \hat{Y})$ である．予測結果ごとではなく，予測結果と過去の観察結果との全ての組合せにおいて，$\mathbf{X}$ の分布がセンシティブ特徴の影響を受けずに等しいという，かなり複雑な条件である．誤分類を考慮した逆個人公平性と呼べなくもないが，条件は複雑であり，逆個人公平性よりもさらに実用的な意味はないだろう．

4. 公平性規準のまとめ

従来から議論されてきた形式的公平性規準には，統計的均一性，個人公平性，均等オッズ，および十分性があった．さらに逆個人公平性 $S \perp\!\!\!\perp \mathbf{X} \mid \hat{Y}$，個人均等オッズ $S \perp\!\!\!\perp \hat{Y} \mid (\mathbf{X}, Y)$，個人十分性 $S \perp\!\!\!\perp Y \mid (\hat{Y}, \mathbf{X})$，誤分類を考慮した逆個人公平性 $S \perp\!\!\!\perp \mathbf{X} \mid (\hat{Y}, Y)$ の規準が考えられることが，以上の議論から分かった．さらに，同時に成立しうるのは個人公平性と均等オッズ，逆個人公平性と十分性の 2 通りの場合に限られ，他の規準は同時には成立させることはできないことも明らかとなった．

2 変数と 3 変数で定義される 6 種の形式的公平性規準の性質について表 1 にまとめた．まず，2 行目の『単位』から説明する．これは，公平性が満たされる単位が個人ごとかグループごとかということである．例えば，同じ経歴・能力の男性と女性がいたとしたとき両者が採用される確率が等しいのが個人単位の公平性である．一方で，男性グループ全体と女性グループ全体で同じ確率で採用されるのがグループ単位の公平性である．個人の経歴や能力といった特徴は $\mathbf{X}$ で表されているため，この $\mathbf{X}$ で条件付けされている個人公平性のみが個人単位の公平性にあたる．なお，均等オッズと十分性については，一般には個人単位での公平性が成立することを 3.3 節で論じたため，括弧書きで『個人』と加えた．

3 行目の『無視』は，センシティブ情報を取得しないことで達成する無視による公平性 (fairness through unawareness) か，取得したセンシティブ情報に基づいて補正することで達成する考慮による公平性 (fairness through awareness) とを示している．無視による公平性を達成できるのは個人公平性のみであることは，3.2.1 節で論じた．ところが，日本でも 2020 年 7 月に JIS 規格の履歴書から性別欄が任意記述になったことから，採用などで性別や人種などの情報を取得しないこと公平性に重要であると広く考えられているように思われる．しかしながら，この無視による公平性ではセンシティブ特徴から予測結果への直接効果，すなわち図 4. の辺 1 のみが削除して，個人公平性を満たすだけである．そのため，個人公平性とは同時には満たせない統計的一致性を達成するには，センシティブ特徴の取得が必須となる．このことは文献 [Žliobaitė 16] で論じられており，EU の General Data Protection Regulation でセンシティブ情報の取得を許す条件に，採用などの決定で公平性を保証することを加える必要があることを指摘している．

最後の『仮定』は文献 [Friedler 21] で提案されたものなので，簡単に説明する．特徴変数と目的変数を表現する空間には，これらの変数に含まれる情報を理想的に表現している構成空間 (construct space) と，この理想的な表現に外乱が加わって実際に観測される情報を表現する観測空間 (observed space) とがある．構成空間ではセンシティブな情報に依存しない理想的な決定が行われるが，観測空間では特徴が変化し，決定過程も理想的なものではなくなる．しかしながら，構成空間の情報は観測できないので，予測タスクを行う世界に対して何らかの仮定が必要になり，その仮定を『世界観仮定』と呼ぶ．“What You See Is What You Get (WYSIWYG)” とは，構成空間と観測空間で個人間の相対的な位置関係は変化しないという仮定である．この仮定の下では，観測空間で得た決定結果は信頼でき，その結果に対する誤差の公平性を論じることになる．もう一つは “We All Equal (WAE)” という仮定である．センシティブ情報で分けたグループごとに歪んだ情報が観測空間では観測されている．しかしながら，構成空間ではグループごとに本質的には等しいというのが，この WAE 仮定である．均等オッズや十分性が WYSIWYG を仮定し，統計的一致性が WAE を仮定していることは，文献 [Friedler 21] と同意見であるが，個人公平性を WYSIWYG 仮定としている点については同意できない．そのため，予測結果を観測結果と変えることが個人公平性でもありうるので，これは WAE 仮定であるとの考えで表にまとめている．

公平性規準のほとんどは同時には満たせないので，表 1 に示した性質を考慮しつつ，公平性規準を選択する必要がある．

表 1: 公平性規準のまとめ

名称	個人公平性 $S \perp\!\!\!\perp \hat{Y} \mid \mathbf{X}$	逆個人公平性 $S \perp\!\!\!\perp \mathbf{X} \mid \hat{Y}$	均等オッズ $S \perp\!\!\!\perp \hat{Y} \mid Y$	十分性 $S \perp\!\!\!\perp Y \mid \hat{Y}$	統計的均一性 $S \perp\!\!\!\perp \hat{Y}$
単位	個人	グループ	グループ（個人）	グループ（個人）	グループ
無視	無視	考慮	考慮	考慮	考慮
仮定	WAE	WAE	WYSIWYG	WYSIWYG	WAE

5. おわりに

本研究では，関連ベースの形式的公平性基準について，数理的に可能なものを列挙した．その結果，従来から利用されている規準の他に，逆個人公平性と呼ぶべき規準も存在しうることが分かった．そして，個人公平性と均等オッズ，そして十分性と逆個人公平性は同時に成立しうるが，他の対では同時には成立しないことが分かった．また，均等オッズや十分性は個人化した条件に，また逆個人公平性も誤分類を考慮した条件に拡張可能であることが分かった．今後はこれらの規準の意味付けや性質と，これらの規準間の関係についてより考察を深めたい．

謝辞：本研究は JSPS 科研費 JP24500194，JP15K00327，JP18H03300，および 21H03504 の助成を受けた．

参考文献

[Angwin 16] Angwin, J., Larson, J., Mattu, S., and Kirchner, L.: Machine Bias (2016), ⟨`https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`⟩

[Bishop 08] Bishop, C. M.: パターン認識と機械学習 — ベイズ理論による統計的予測, 上下, 丸善出版 (2007–2008), [監訳：元田 浩 他；訳：神嶌 敏弘 他]

[Boutiller 96] Boutiller, C., Friedman, N., Goldszmidt, M., and Koller, D.: Context-Specific Independence in Bayesian Networks, in *Uncertainty in Artificial Intelligence 12*, pp. 115–123 (1996)

[Dwork 12] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.: Fairness Through Awareness, in *Proc. of the 3rd Innovations in Theoretical Computer Science Conf.*, pp. 214–226 (2012)

[Feldman 15] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S.: Certifying and Removing Disparate Impact, in *Proc. of the 21st ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 259–268 (2015)

[Flores 16] Flores, A. W., Bechtel, K., and Lowenkamp, C. T.: False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.", *Federal Probation Journal*, Vol. 80, No. 2 (2016)

[Friedler 21] Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S.: The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making, *Communications of the ACM*, Vol. 64, (2021)

[Hardt 16] Hardt, M., Price, E., and Srebro, N.: Equality of Opportunity in Supervised Learning, in *Advances in Neural Information Processing Systems 29* (2016)

[神嶌 19] 神嶌 敏弘, 小宮山 淳平：機械学習・データマイニングにおける公平性, 人工知能, Vol. 34, No. 2, pp. 196–204 (2019)

[神嶌 22] 神嶌 敏弘：私のブックマーク「人工知能と公平性」, 人工知能, Vol. 37, No. 2, pp. 230–233 (2022)

[Kusner 17] Kusner, M., Loftus, J., Russell, C., and Silva, R.: Counterfactual Fairness, in *Advances in Neural Information Processing Systems 30* (2017)

[Lippert-Rasmussen 06] Lippert-Rasmussen, K.: The Badness of Discrimination, *Ethical Theory and Moral Practice*, Vol. 9, pp. 167–185 (2006)

[Pedreschi 08] Pedreschi, D., Ruggieri, S., and Turini, F.: Discrimination-aware Data Mining, in *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp. 560–568 (2008)

[Žliobaitė 16] Žliobaitė, I. and Custers, B.: Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models, *Artificial Intelligence and Law*, Vol. 24, pp. 183–201 (2016)