Laboratoire d'Analyse et Modélisation
de Systèmes pour l'Aide à la Décision
UMR 7243

CAHIER DU LAMSADE

409
June 2024

Subjective fairness in algorithmic decision-support

Sarra Tajouri, Alexis Tsoukiàs

Subjective fairness in algorithmic decision-support

Sarra Tajouri, Alexis Tsoukiàs
CNRS-LAMSADE, PSL, Université Paris Dauphine

Abstract

The treatment of fairness in the decision-making literature usually consist in quantifying fairness through objective measures. This work takes a critical stance to highlight the limitations of these approaches (group fairness and individual fairness) using sociological insights. First, we expose how these metrics often fail to reflect societal realities. By neglecting crucial historical, cultural, and social factors, they fall short of capturing all discriminatory practices. Second, we redefine fairness as a subjective property moving from a top-down to a bottom-up approach. This shift allows the inclusion of diverse stakeholders' perceptions, recognizing that fairness is not merely about objective metrics but also about individuals' views on their treatment. Finally, we aim to use explanations as a mean to achieve fairness. Our approach employs explainable clustering to form groups based on individuals' subjective perceptions to ensure that individuals who see themselves as similar receive similar treatment. We emphasize the role of explanations in achieving fairness, focusing not only on procedural fairness but also on providing subjective explanations to convince stakeholders of their fair treatment.

1 Introduction

The pursuit of fairness, closely linked to the discourse on justice, has long captivated the minds of philosophers [64, 32, 18, 2, 62], sociologists [72, 10, 19], and economists [63, 3]. Notably, John Rawls' theory of justice as fairness, outlined in his seminal work "A Theory of Justice" [62], posits that societal arrangements should be structured to benefit the least advantaged members. Rawls argues for a hypothetical social contract, where individuals, behind a "veil of ignorance" about their own characteristics, would agree on principles of justice. He proposes two principles: the equal basic liberties for all and social and economic inequalities that are arranged to benefit the least advantaged, known as the *Difference Principle*. This work does not focus on fairness purely as a philosophical concept. Instead, we propose a novel approach to fairness in algorithmic decision-making that integrates the perceptions of individuals who collectively agree on what constitutes fairness.

Indeed, the pervasive use of algorithms in social and economic contexts has raised a significant concern. Many instances of algorithmic discrimination have been reported over the past decade including the investigation by *ProPublica* which exposed racial biases in Northpointe's COMPAS tool, a recidivism risk assessment algorithm [1]. These cases of algorithmic discrimination stem from models trained on either biased datasets [27], which lack diversity and fail to accurately represent society [67], or datasets that simply mirror the inequalities inherent in certain segments of society. Without adequate consideration for debiasing or corrective measures, these algorithms will perpetuate and exacerbate social injustices.

Consequently, fairness has become a fast-growing area of interest within the scientific community, giving rise to a proliferation of research papers [31, 40, 22, 56, 57] that have crystallized into a distinct field in computer science. This rise is due to the recognition that fairness is intrinsically linked to the collective well-being of our society. Numerous works have sought to formalize fairness within a decision process using mathematical constraints or metrics [40, 74, 71]. Although those techniques are well founded and produce promising results, they often fall short in incorporating the social dimension of fairness. In parallel, extensive work in social sciences has focused on this aspect, offering valuable insights that can significantly enrich the interdisciplinary perspective on fairness in computer science.

A decision process consists of multiple stages. When we discuss fairness, it is always with respect to a specific stage of the process and directed towards a particular stakeholder. As such, fair process refers to the impar-

tiality of procedures and methods used in making a decision. Whereas fair recommendation (suggestions of a decision-support system) and fair decision (final result) concern the outcomes. This distinction will be discussed notably in section 2. Following that, we expose some of the limitations of traditional fairness approaches, highlighting how they lack to align with societal realities, drawing on sociological perspectives. Then, we propose a novel approach viewing fairness in the decision-making process through a subjective lens. Central to our perspective is the inclusion of those impacted by a decision in the assessment of fairness, empowering them to determine if they are treated fairly. Our methodology for achieving subjective fairness employs explanations as a tool to construct justifications, aiming to convince stakeholders of the fairness in the process.

2 Decision framework

2.1 Decision, process and recommendation

Decision-making has been studied in several disciplines ranging from cognitive studies [14, 36] to management science [25] and from economy to organizational studies [66]. From an abstract mathematical perspective, any partitioning of set according to a subject's preferences is a decision problem [20]. In economics, a decision is an irreversible allocation of resources to individuals for an objective achievement such as maximizing utility, profits etc..

Contrary to viewing decisions as a single act, it should be regarded as a process. Simon [66] defined a decision process (DP) as a series of means and ends connected in a hierarchical chain to achieve an objective including predictions about decision behavior. According to the classical theory, accurate prediction of behavior can be achieved by considering the environment together with the strong assumption of *perfect rationality*. However, when dealing with imperfect competition and decision-making under uncertainty, such as not knowing all possible alternatives or external events, models of *bounded rationality*, subjectively defined and only valid within specific contexts, are more suitable to provide more realistic explanations of human decision-making behavior [65].

Decision-aiding process (DAP) is one form of decision process that involves more than one stakeholder. To simplify we can consider that it includes a decision-maker, who has domain knowledge concerning the decision process, and an analyst, who has technical expertise. The objective is

to reach a consensus between these two actors that responds to the initial problem of the decision-maker [69].

Then, the increasing availability of data and the expansion of computers capacity led to the rise of using automated decision making systems (ADMS) where algorithms could autonomously take decisions without human intervention [69]. From a computer science perspective, decisions can be viewed as the output of an algorithmic process with binary outcomes. ADMS are particularly suited for frequent generic decisions where speed is critical, supported by available collected data. Credit scoring is one example of ADMS where the decision problem is standardized: assigning a risk score to each candidate. This process is repetitive and depends on the output of software developed by the analyst(s). The decision-maker then uses these scores to make final decisions on whether to grant credit or not.

It's crucial to recognize that decisions carry responsibility and liability due to their potential unintended consequences [70]. But since the decisions are automated, the responsibility is often diffuse, shared among human stakeholders involved in the process design and implementation. ADMS actually propose a recommendation to the decision-maker who takes the responsibility to decide afterwards. However, most of the time, humans may lack the capacity to thoroughly evaluate each recommendation, leading them to accept suggestions without scrutiny. Consequently, the control over the final decision doesn't guarantee control over the outcome, as "recommendations are often treated as decisions"[70].

These advancements led to a proliferation of literature about the implications of such automated decision-making, including discussions on AI ethics, fairness, accountability, transparency etc.. In fact, algorithmic decisions have unveiled inherent biases in society and issues of discrimination against minorities that were embedded in the data we used to automate decisions. However, it is important to keep in mind that while correcting algorithmic outcomes is necessary, it alone cannot address broader societal issues. The most effective response to social challenges remains rooted in human and political action rather than purely technological solutions.

2.2 Fairness of what ?

Practically, some of the decision processes concern algorithms that suggest a certain action to be undertaken in a high-stake context. It could concern the distribution of financial resources, economic or educational opportunities, granting bail etc..

The EU AI act [21], the first regulation on artificial intelligence proposed

by the European commission, attempts to regulate the use of AI systems using a risk-based approach, distinguishing unacceptable, high and low risk. High-stake decisions fall under the definition advanced by the AI Act as "*systemic risk at Union level means a risk that is specific to the high-impact capabilities [...], having a significant impact on the internal market due to its reach, and with actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain*" (Article 3, (44d)) [21].

Some examples of high-risk systems are the one that operates in the educational and professional sphere (i.e. admission to university and job hiring), in access to essential private services and essential public services or administration of justice and democratic processes such as systems used by a judicial authority to assist it in researching and interpreting facts and law.

In this paper, we focus on this type of decisions, that are generally binary with a "good" and a "bad" outcome. Ensuring fairness in these decisions is essential as they can have unintended consequences on people's lives and can deeply influence the future trajectory or stability of individuals, or even communities.

In addressing concerns about fairness in high-stakes decisions, we must consider what stage of the decision-making process we are evaluating for fairness. Is it the fairness of the process itself, the recommendations provided, or the final decisions made? As discussed in 2.1, many ADMS actually "suggest" a certain action, which is generally followed by the decision-maker, as observed in credit scoring and predictive justice scores. Therefore, when discussing fair outcomes in this context, it is primarily about fair recommendations.

Fair outcomes involves the absence of bias or discrimination and should be considerate of the interests of all stakeholders. However, it's important to note that there is no universal definition of fairness it depends on the notion adopted by the decision-maker, as will be further discussed in the next section.

Fairness of the process, on the other hand, requires that the decision-making process should be explainable, justifiable and perceived as meaningful for the analyst, useful for the decision-maker and legitimated by stakeholders [68]. The process might need to be understandable enough to be argued by the stakeholders and possibly be challenged or recused. For this to be possible, we consider that providing explanations for which the automated system made a given recommendation is fundamental and a first step

to ensure fairness [70].

In this paper, the proposed approach is positioned within the framework of process fairness in high-stakes decisions. However, this should not be interpreted as underestimating the importance of fairness of recommendations. A lot of research in social psychology explored the link between fair process and fair outcome, notably the work of Lind et al. (see [54, 8, 53]). They have demonstrated that fair processes can significantly influence how individuals react to outcomes. For instance, participants who were afforded the opportunity to express their opinions during the process tended to react more positively to the outcome compared to those who were not given this opportunity [8]. This remains out of the scope of this paper, we will only focus here on process fairness.

2.3 Fairness for whom ?

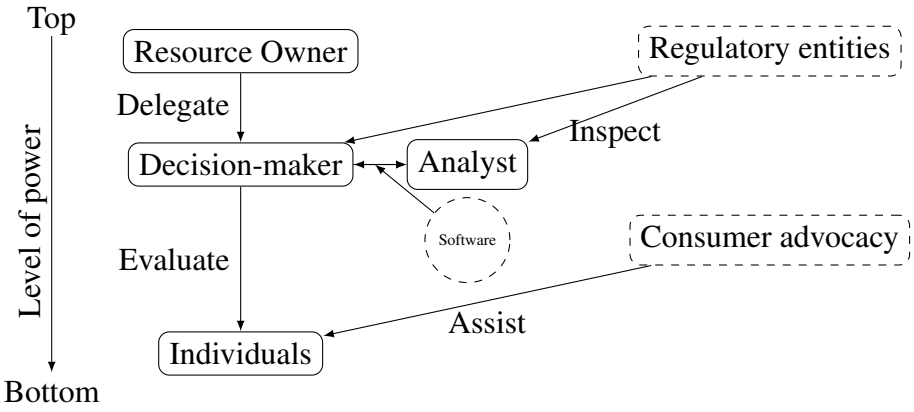

Figure 1: Stakeholders in a decision-making process

We refine the framework to clarify to whom we address fairness in the decision process. As discussed previously, we recognize the involvement of multiple stakeholders in a DP. In figure 1, we illustrate one possible configuration of stakeholders, in the case of high-stake decisions, and expose the power dynamics between them.

Power can be defined in various ways: according to Crozier [24], it's a dynamic relationship where one party can gain more than the other, yet neither is entirely destitute against the other. Dahl [26], on the other hand, views power as the ability of one party to ensure favorable terms of exchange for him/her in negotiations with the other.

Exploring the foundations of power, as suggested by Crozier [24], it's rooted in the assets, resources, and strengths of each stakeholder. However, it's not limited to these factors; the ability to take action (*possibilité d'action*) is also crucial. When A holds power over B, it's not merely to exert control but because A seeks to achieve an objective and B influences the possibility of its attainment. Therefore, power also lies in the *level of freedom* each stakeholder possesses and their ability to decline demands from others. Additionally, factors such as authority and other subjective attributes significantly impact freedom of action.

Adopting this paradigm, we outline the stakeholders involved in the decision-making process and position them along a power axis (depicted in Figure 1) to better understand the dynamics of their relationships and how do we want to move along this axis to approach the decision-making process differently.

At the top, we place the resource owner, representing the actor(s) with ownership rights over the resource to be distributed. They may possess financial, educational, or other forms of capital. It's important to note that the presence of this stakeholder is contingent on the case involving quantifiable resources controlled by individuals. For example, in the context of credit lending, the resource owner would be the investors and shareholders.

Below the resource owner, we find the decision-maker, tasked with allocating the resource based on interactions with analysts possessing technical knowledge to propose models (i.e. a software) and conduct proper evaluations for decision-making. At the lowest tier are the individuals impacted by the decision, often lacking influence or agency in the decision-making process. We specify "often" because in some cases users have the possibility of recourse to contest a decision and win the case, notably with the help of another eventual stakeholder who would be the consumer advocacy that help to assist citizens to claim their rights.

Additionally, decision actors are bound by a fiduciary duty to legal principals responsible for ensuring compliance with relevant legislation. Indeed, a decision goes through a legitimization process. What partially confer legitimacy to decisions made by one of the stakeholders are the established norms, rules and objectives. Indeed, actions are assessed against norms, results are compared to objectives, procedures are judged compliant with reference to the rules [44].

It is important to note that fairness is always established with respect to a specific stakeholder. We can consider that a process is fair if it is fair towards all stakeholders. However, achievement of fairness can sometimes be conflicting. For example, in credit allocation, "positive" recommendations

might be considered fair for socially deprived individuals who may not be financially solvable. Granting credit to these individuals can be viewed as an act of social justice. However, this can also be seen as unfair to the resource owner because it jeopardizes the financial stability of the banking institution. In this framework, we center our concerns about fairness towards the individuals impacted by the decision. Even though, fairness towards other stakeholders shouldn't be overlooked, the trade-off between fairness and performance will not be treated in this paper.

3 Critics towards traditional algorithmic fairness

3.1 Group fairness limitations

Some of the first work on fairness in the literature often center around what we call *group fairness* [40, 33, 74, 71]. This concept originates from the recognition that certain sub-populations have endured historical discrimination in our society and that was embedded in the data that we use to train models nowadays. Consequently, the goal of group fairness is to rectify biases by considering a protected attribute (e.g. gender, race, nationality), creating groups based on attribute modalities and mitigating outcomes within those groups.

For example, let's consider an employer that relies on a decision-aiding process, where given an application (i.e. a vector of individuals' characteristics) returns a prediction about whether a candidate would make a good employee. Applying group fairness in this process could be a way to ensure equal opportunities for all genders, acknowledging and correcting past gender biases.

Despite its intent, group fairness faces certain limitations when confronted with the complexity of historical biases and their translation into algorithmic decision-making. In the following, we set out the shortcomings we have observed.

3.1.1 Inherent incompatibility

There is an inherent incompatibility between some fairness metrics. This limitation has been extensively studied by [51]. While we won't focus on this aspect, it is worth mentioning their impossibility theorem that highlights the tension between different definitions of fairness, such as equalized odds,

equal opportunity, and calibration, when applied to risk prediction. The authors demonstrate that in certain scenarios and given a protected attribute, achieving one form of fairness may inevitably lead to unfairness in another dimension.

3.1.2 Inner-group inequalities

Group fairness approaches often ignore inner-group inequalities which refer to disparities or variations that exist within a particular demographic group. Indeed, discrimination is blind to the definition of groups so it is important to consider intersectionality, which acknowledges that individuals have multiple dimensions of identity.

The concept of intersectionality, introduced by Crenshaw [23], highlights how overlapping systems of power impact the most marginalized individuals. While rooted in the analyses of black feminists, intersectionality has been embraced by various sociological perspectives, including Marxism, Weberianism, and Bourdieusian sociology [73]. It challenges traditional stratification theory by rejecting a singular focus on one social division (i.e. class), recognizing that differentiation occurs across various facets of social analysis. Individuals are actually positioned along socioeconomic grids of power, identificatory perspectives of belonging, and the normative value systems that shape their experiences within a complex societal framework [73]. This brief overview lays the groundwork for deeper exploration in future research.

Therefore, it appears unrealistic and unjustifiable to center the evaluation of a process's fairness solely on a single protected attribute. Such an approach is reductionist, oversimplifying people's identities. Even if fairness is attained within one dimension, it disregards other forms of discrimination. For instance, imagine that an employment interview process achieves statistical parity between men and women, meaning both groups receive an equal number of interviews. According to this metric, the process appears fair. However, this approach does not detect the fact that black women receive significantly fewer interviews than white women, leading to unfair practices.

Even though some have suggested constructing subgroups with various combinations of protected attributes [41], the challenge lies in their large number, making it difficult to consider all possibilities.

3.1.3 On the legitimacy of discrimination based on "merit"

The tension between maximizing an objective and predictive parity as a fairness metric has been discussed in economics as a failure of predictive parity to reflect taste-based discrimination [47]. In general, fairness metrics rely on the notion of merit, which can inadvertently legitimize and perpetuate discriminatory practices aligned with the decision-maker's objectives. Usually, concerns for fairness arise when societal ideals are in tension with a decision-maker's interests. The objective function's choice is crucial to the model. Most of the time it is chosen by policymakers who have ownership and control rights over the data and the algorithm [22]. Consequently, such metrics when applied on the objective function, may inadvertently reinforce the legitimacy of existing norms [47].

Usually, when we rely on merit to assess the entitlement of individuals to resources in an allocation process, we overlook the challenge of accurately measuring merit. The feature space employed often serves as a representation of latent or unmeasurable constructs. Friedler et al. [35] shed light on a compelling example in the context of college admissions: the decision should be made based on characteristics such as the intelligence, perseverance and motivation of candidates which are not directly observed. Rather, the feature space is constructed using proxies to those qualities such as IQ, school grades, extracurricular activities... Yet, these approximations may suffer from structural bias due to social and economic circumstances resulting in an incomplete quantification of the candidates' intrinsic characteristics.

If we try to analyse this from a social science perspective, we notice that defining merit is similar to understanding philosophical concepts like value and worth [55]. Habermas [39] highlighted that issues once thought purely philosophical now require consideration of their social context. Philosophy must recognize its connection to real-world contexts and history [43]. Following this principle, we take interest in sociology, especially the sociology of education, where there has been thorough examination of the nature and constraints of merit.

Bourdieu [9] showed how the educational system reproduces the class structure, reinforcing social inequalities and social stratification, and conceals the fact that it fulfils this function under the guise of neutrality. In fact, the reproduction of these hierarchies, founded on the hierarchy of donation ("dons" in french) and merit, serves as a legitimization for the perpetuation of the social order. However, the cultural capital and the capital of relationships inherited from the family "*are the condition, if not the main factor,*

for success" [9]. Therefore, the relative autonomy of the educational market only lends apparent justification to the meritocratic ideology. However, this overlooks that intelligence or academic diligence represent only one form of capital, often possessed alongside economic and social capital. Moreover, those with economic capital have greater chances of acquiring cultural capital, rendering educational credentials valuable primarily within the confines of the educational market [9].

3.1.4 How about fairness in other fields ?

In other disciplines besides the computer science literature, fairness exists independently of the existence of a protected group. Actually, there is no or little connection between how fairness is treated in the computer science community and the notion of fairness in social choice theory or welfare economics.

In social choice theory, the focus is on individual fairness through an axiomatic approach and the purpose is to distribute resources or make collective decisions in a fair manner by aggregating individual preferences. One of the principles of fair division in social choice ([11, 58, 12]) is envy-freeness: each person should receive a share that is, in their eyes, at least as good as the share received by any other agent. In other words, an allocation is considered envy-free if no agent would prefer someone else's bundle over their own. Contributions that study fair classification from fair division perspective are discussed in the related work section 6.

In welfare economics, Sen [64] extends Rawls' conception of primary goods to focus on what goods actually do for individuals. This shift forms the basis of the concept of 'capability equality'. He argues that an individual's well-being cannot be adequately measured solely by the resources they possess, but must also include their capability to use these resources to live a life they have reasons to value. This perspective ensures that the assessment of well-being and fairness includes the real opportunities available to individuals, reflecting their own perceptions and values.

Ed Diener has enriched the literature on psychology with a subjective well-being (SWB) model [28]. He showed that objective conditions such as health and wealth are not inherent and necessary to evaluate the SWB of people. Instead, SWB describes how people perceive the quality of their lives. This led to a proliferation of welfare economic research using subjective measures of happiness and life satisfaction [46]. It's worth mentioning that an individual's perception of fairness holds significance for their sense of well-being and can indeed impact social interactions, thus contributing to

the overall well-being of society.

Since we try here to adopt an interdisciplinary perspective on fairness and explore insights from other fields, the subjectivity comes as an area worth exploring. The preferences of individuals should play a role in determining the fairness of a process.

3.1.5 Non-universality

The categorization between protected and unprotected groups is often rooted in political constructs rather than scientifically grounded boundaries, so it presents challenges when we try to universally generalize this approach. The term "protected" usually refers to protection from anti-discrimination laws or policies notably with respect to gender, "race", ethnicity, religion, sexual orientation etc.. We note that the outdated notion of race as a scientific concept, has been widely discredited and is now recognized as entirely erroneous. Here we are referring to the sociological concept of "social race" as the socially constructed racial categories. Some of the political and social factors that shape how these categories are formed and recognized are societal perceptions, historical contexts, power dynamics, legal systems and cultural norms.

For instance, in the United States, racial categories in census are commonplace. The country has a long history of racial discrimination, redlining and partisan gerrymandering. During the eighteenth century, those in positions of political power perceived race as an inherent and obvious aspect of human identity, aligning with the ideals of the European Enlightenment. Ever since, race was an organising ideology and census of the population integrated racial categories, which varied through time [59]. However, after the civil rights movement, the purpose of racial categorization drastically changed. These information are now used to establish public policies that respect civil rights and to dismantle discriminatory mechanisms as residential discrimination, exclusion from certain occupation etc... [59]. Hence, in this context, racial data seem to be a good way to evaluate the fairness of government programs and to monitor compliance with anti-discrimination laws and regulations [49].

In parallel, in the majority of the EU member states, notably France and Germany, the distinction based on race is nonexistent. Collecting data on racial and ethnic origin is actually prohibited pursuant to the Racial Equality Directive (RED), which was adopted in 2000. It is an important legal instrument within the EU aimed at eliminating discrimination and ensuring equal treatment and opportunities for all individuals, irrespective of their racial and

ethnic differences. This aligns closely with the concept of "fairness through blindness" found in the scientific literature.

However, the directive has led to a unique approach in these countries, where the concept of race isn't officially recognized or used for data collection purposes. This approach has its roots in historical sensitivities and experiences, particularly in countries like Germany, where the legacy of Nazi ideology and its catastrophic consequences have heavily influenced national policies regarding racial categorization [16].

France, similarly, has historically followed a republican model of "colorblindness" or "universalism," aiming to promote equality by avoiding the recognition of racial or ethnic distinctions within its legal frameworks. This stance is rooted in the idea that acknowledging such differences might lead to division and inequality [15].

Therefore, implementing a universal approach based on considering protected attributes poses significant challenges. While this method might hold relevance within the context of the historic and political situation in the USA, its applicability elsewhere is not straightforward.

3.1.6 Do not assume sense of belonging !

Constructing demographic groups is generally done by the decision-maker (or data collector) based on assuming and inferring the affiliations. The construction of demographic groups typically relies on the decisions made by those in positions of power, often relying on assumptions and perceived characteristics about individuals' identities. Thus, this process of assumption-making can be inherently flawed or oversimplified, and can perpetuate stereotypes and biases, potentially leading to discrimination or unequal treatment based on these constructed demographic groups.

Indeed, social identification and belonging to a community can be a matter of choice. An empirical study [60] aimed at investigating the relationship between the degree of choice in community membership and the subsequent levels of social identification showed that higher degree of choice is associated with higher levels of cohesiveness within a community. Membership to a community of interest may even be stronger than identification to a local neighborhood for example. Hence, deducing community affiliation becomes unsatisfactory due to the potential for individuals to opt for alternative group memberships beyond their designated assignment. For instance, a person classified as male in their civil status may choose not to align themselves with the male category and identify with a different gender group.

The choice of belonging to a community is political and not a natural

process. Taking the instance of race, it is indisputable that it emerges as an ideological construct. The illustration of the multiracial scenario amplifies this claim as individuals may belong to multiple racial categories. The imposition of inevitably invented racial categorizations is often resisted, as it contradicts self-perceptions and undermines the authenticity of personal feelings.

3.2 Individual fairness limitations

In contrast to group-based fairness, some works explored individual fairness according to the principle *"equals should be treated equally"* [31]. This notion aligns more closely with our definition of fairness, recognizing that the assessment of fairness should occur at an individual scale. Building on this approach, we identify two limitations that we aim to overcome.

First, this principle relies on the notion of "equals". But what can we consider "equals" ? In the literature, equality is measured through similarity on objective criteria such as income, wealth, education... It aims to ensure that individuals have the same opportunities and access to resources, if they are sufficiently similar on those objective criteria regardless of their background or (irrelevant) personal characteristics. However, this makes us question the legitimacy of those objective criteria and who has the right to fix them. We also point out that this principle fails to capture the subjectivity of equality, as "equals" is very different from "feeling equal". The former is a judgment held by the decision-maker while the latter takes into account the perceptions of the individuals impacted by the decision. Indeed, "subjective equality" focuses on how individuals experience equality, taking into account people's feelings and perceptions of fairness and justice. It recognizes that equality isn't solely about objective measures but also about how individuals perceive their treatment in society.
For example, two individuals with the same objective level of income may still feel unequal if one perceives their income as unfairly low due to discrimination or systemic barriers. Conversely, two individuals living in distinct environments with significantly different incomes may still perceive themselves as equal if their purchasing power remains similar.

Second, defining similarity between individuals is not an easy task and it has been one of the challenges of Dwork's work [31]. However, it has generally been assumed that this distance metric would be symmetric. To our knowledge, no approach in the fairness literature quantifies similarity using non-symmetric distance measures. Yet, this constitutes a crucial aspect of our work. Indeed, we believe that there is a non-negligible subjective dimen-

sion in the notion of similar (or equal) that must be reflected in the chosen similarity measure. Having a symmetric distance constrains the subjectivity of similarity, as we will always rely either on objective features or on one of the two parties to determine a real-value distance. Therefore, if we aim to quantify the similarity between two individuals, we must consider that they may not perceive their distance in the same way. We can get a sense of why this is important through an example.

Let's recall the property of individual fairness according to [31]: a mapping M satisfies individual fairness (IF) if for all $(x, y) \in X^2$, we have $D(M(x), M(y)) \leq d(x, y)$. In this example, let M being used for college admission decisions, yielding a probability that an applicant should be admitted. Suppose two applicants x and y that are objectively quite similar but y has slightly higher scores at exams so the objective distance between the two is $d(x, y) = 0.05$. And the mapping produces these scores $M(x) = 0.85$ and $M(y) = 0.9$. Then, the property $D(M(x), M(y)) \leq d(x, y)$ holds and M would be considered fair. However, x may perceive that she is closer to y because despite coming from a more disadvantaged high school and having less guidance and resources, x has exerted more effort than y to achieve similar scores, so x believes that $d_x(x, y) = 0.04$. Consequently, $D(M(x), M(y)) > d(x, y)$, leading x to perceive the treatment as unfair, while y adheres to the objective distance measure ($d_y(x, y) = 0.05$) and views the outcome as individually fair. The use of non-symmetric distance function in this example enables us to incorporate the perceptions of the individuals, which is the focus of this paper.

4 Subjective fairness

Building on the analysis of fairness measures' limitations, we try to redefine fairness with a subjective dimension, taking into consideration the perception of individuals.

One way to conceptualize this shift is by transitioning from a top-down to a bottom-up approach. When it comes to determining what constitutes equality, the responsibility should not rest exclusively with decision-makers. Instead, impacted individuals should be empowered to identify and report potential instances of discrimination, and have a say in evaluating whether they perceive their treatment as fair [37]. Our definition of fairness can be roughly summarized as a subjective extension of "equals should be treated equally".

Let's consider a set of individuals **I**, a set of issues $\mathcal{X}$ and a set of out-

comes $\mathbf{R}$. A decision-support system is a mapping $M : \mathbf{I} \times \mathcal{X} \mapsto \mathbf{R}$ such that $M(x, \psi) = r \in \mathbf{R}$ refers to the outcome r where individual $x \in \mathbf{I}$ is concerned by issue $\psi \in \mathcal{X}$.

We introduce a non-symmetric subjective similarity measure such that given an individual $x \in \mathbf{I}, \forall z \in \mathbf{I}$ $sim_x : \mathbf{I}^2 \mapsto [0, 1]$ describes the similarity between two individuals in the set $\mathbf{I}$ as perceived by x. From that we construct the set S_x of individuals that x considers similar to him such that

$$S_x = \{z | z \in \mathbf{I}, sim_x(x, z) \geq \delta\}$$

We then assume a similarity metric $T : \mathbf{R}^2 \mapsto [0, 1]$ describing the similarity between the treatment of individuals, and define $\phi(x, \psi)$ being the fact that x considers being treated fairly on purpose ψ. It could be binary $\phi(x, \psi) \in \{\textit{fair}, \textit{unfair}\}$ or take the form of a scale, a score, etc. To initiate this process, we put forth an initial definition of subjective fairness on an individual scale.

Definition 1 (Individual Subjective Fairness (ISF)) *Given $x \in \mathbf{I}$ and $\psi \in \mathcal{X}$, x considers herself to be treated fairly with respect to ψ if all individuals she considers similar to herself are treated similarly :*

$$\phi_{\delta,\epsilon}(x, \psi) \Leftrightarrow \forall y \in S_x, T(M(x, \psi), M(y, \psi)) > \epsilon \tag{1}$$

Definition 2 (Subjective fair process) *A decision process is $(\delta, \epsilon) - SF$ if all individuals involved in the decision process consider themselves being treated fairly:*

$$F_{\delta,\epsilon}(\mathbf{I}, \psi) \Leftrightarrow \forall x \in \mathbf{I}, \phi_{\delta,\epsilon}(x, \psi) = \textit{fair} \tag{2}$$

Example 1 *Let's consider two employees, Alice and Bob, both applying for a raise. Objectively, they have similar qualifications. For instance, both hold master's degrees and possess relevant job experience. However, their social status differ significantly. Alice is a black women, who comes from a privileged background, attended a prestigious school, and had financial support throughout her education. Bob is a white man, who comes from a precarious environment, had to work part-time jobs to pay for school, and attended a less renowned institution.*

Suppose Alice considers herself similar to Bob due to their shared qualifications. If Bob is accepted while Alice is not, it suggests disparate treatment. Alice may perceive the salary increase process as potentially sexist and/or racially discriminatory, given that the only discernible difference between her and Bob is their social backgrounds.

Now, let's suppose that Bob does not consider himself as similar to Alice. In such a case, Alice's outcome has no impact on Bob's perception of fairness. Since he doesn't view her as part of his group, he doesn't anticipate receiving the same treatment as her.

As we are dealing with subjective information, including feelings and opinions, we are confronted with defeasible information. For instance, perceptions of similarity may fluctuate due to evolving factors, leading to shifts in their understanding or evaluation of a given scenario.

Consequently, the determination of subjective fairness is not a static, one-time event but rather a dynamic process that necessitates justification for stakeholders to evaluate their perceptions effectively. In the following, we explore this point further.

5 Explanations to achieve subjective fairness

The issue of explainability is not independent of fairness. We believe that fairness could be achieved through explanations. By understanding the reasoning and the functioning of an algorithm, we gain the ability to identify instances of discrimination and rectify them, ensuring fairness towards all stakeholders. Multiple works operate between these two topics including [29, 76, 5].

First, we can establish how explanations could justify the fairness of a process in accordance with a normative standard. This is commonly known as *procedural fairness*. It adopts an objective stance, asserting that fairness is contingent upon adherence to procedural rules [52, 48]. Leventhal [52] has posited that procedural fairness often hinges on satisfying six key constraints. Some of these principles are:

- *The consistency rule* states that the process should remain consistent and uniform across all individuals, aligning closely with the concept of equality of opportunity.
- *The accuracy rule* dictates that the decision process should rely on the best available information, reflecting the principle of accountability.
- *The ethicality rule* requires that procedures must be compatible with the fundamental moral and ethical values accepted by the stakeholders.

Implicit in this approach is the notion that fairness can be assessed at a specific moment, guided by predefined rules, overseen by the decision-maker, and remains static without evolution.

Definition 3 (Fairness through objective explanations) *A process is fair if the explanations about how the process has been conducted either satisfy pre-established rules or a normative standard.*

Although fair procedures are commonly perceived as neutral and free from self-interested or ideological considerations, individuals make subjective evaluations of processes' fairness since it depends on their knowledge, prior preferences and bias. This is supported by [30], as they state that "*The idea that perceptions of procedural justice are subjectively determined is demonstrated in that appraisals of what is fair and unfair can vary across both individuals and circumstances. [...] people discount the importance of fair procedures when they were motivated to find support for (or arguments against) the legitimacy of a given outcome.*"

Objective explanations are particularly suitable to justify and support decisions for stakeholders not directly affected by them, but who are more interested in evaluating the process and ensuring regulatory compliance. However, in our framework for achieving subjective fairness (towards impacted individuals) through explanations, compliance with normative measures is necessary but not sufficient. Explanations should serve an additional role which is proving the legitimacy of the decision.

Following this idea, we consider that explanations are a social interaction process between two parties. It builds upon descriptive explanations to present an argument that convinces the stakeholders of the fairness and legitimacy of a decision. These arguments are defeasible, in the sense that they are not absolute and can be overturned when new information or a change in perspective is introduced. The level of acceptance of an argument is subjective and varies depending on the beliefs, attitudes, and biases of the audience. The same argument may be persuasive to some people but not to others [6].

Definition 4 (Subjective fairness through explanations) *A process is fair if the explanations about how the process has been conducted are convincing and accepted by all the population.*

Example 2 *In example 1 where Alice considers herself as similar to Bob, she expects that they will receive the same treatment. However, the employer decides to only grant a raise to Bob. One plausible explanation for this decision could be that Bob brought in a substantial client to the firm. According to their employment contract, employees who secure new clients and sign service agreement exceeding a certain amount are eligible for a promotion (normative explanation). Plus, a justification advanced by the*

decision-maker is that the firm has limited resources, and the raise is directly tied to the value of the signed agreement. If Alice accepts this argument, we can conclude that the process is subjectively fair for both Alice and Bob. Otherwise, Alice should advance counter-arguments to contest the decision. If there are no further explanations to respond to Alice's arguments, then the process is considered unfair.

5.1 Explainable clustering models

The choice of the clustering model must prioritize explainability. There are several methods to achieve this. First, a priori clustering which includes unsupervised machine learning algorithms. While effective, these methods rely on correlations rather than causal relationships and often depend on objective features, which may not capture subjective aspects [17]. Incorporating causality into fairness frameworks has been suggested to address these limitations [56]. Notably, semi-supervised learning frameworks that allow for the integration of user feedback to introduce subjectivity into the clustering process [42].

Fitted clustering is a method that involves collecting individual perceptions and preferences through surveys or interviews to create clusters reflecting shared subjective viewpoints. While theoretically robust and explainable, it is challenging to implement on a large scale and can lead to inconsistencies when integrating multiple perspectives. This situation is discussed in the next subsection. Similarly, sample-based clustering uses a sample of candidates to derive clusters. It is easier to implement but assumes the sample is representative of the entire population, which may not always be the case.

Additionally, Raboun et al. [61] introduced a dynamic clustering which addresses limitations in generating ratings by categorizing objects into predefined classes based on preference relations and reference profiles. It dynamically updates preferences with each new rating, ensuring explainability and consistency.

5.2 Decision on clusters

In the following, we propose a framework where we have recommendations of the decision-support system and the objective is to ensure that individuals who perceive themselves as similar receive the same treatment. Given subjective groups based on individuals' perceived characteristics, our focus is to make decisions that satisfy SF.

Let $\mathbf{I}$ be the population of individuals with $|\mathbf{I}| = n$. Each individual $i \in \mathbf{I}$ is associated with the set S_i of individuals he considers similar to himself, as presented in equation 4. We can denote the population as $X = \bigcup_{i \in \mathbf{I}} \{S_i\}$.

Assumption 1 *Each individual naturally perceives themselves as similar to themselves:* $\forall i \in \mathbf{I}$, $S_i \neq \emptyset$ *because* $sim_i(i, i) = 1$, *thus* $i \in S_i$. *This implies that we have a decision for every individual in the population.*

One individual can belong to more than one cluster. For example, consider two individuals $x, y \in \mathbf{I}^2$. Naturally, $x \in S_x$. Additionally, y considers x to be similar to her, so x is also included in S_y. However, y is not necessarily included in S_x because we use a non-symmetric similarity function.

Let $\mathbf{D}_\psi$ be a vector representing decisions concerning n individuals for a purpose ψ. This vector is n-dimensional and consists of binary elements, where $\mathbf{D}_\psi : \mathbf{I} \to \{0, 1\}$. Each element d_i in $\mathbf{D}_\psi = (d_1, \ldots, d_n)$ corresponds to the final decision rendered for individual i.

Let $\mathbf{R}_\psi$ be a vector representing recommendations suggested by the decision-support system for n individuals in the population, for a purpose ψ, $\mathbf{R}_\psi : \mathbf{I} \to \{0, 1\}$. Each element r_i in $\mathbf{R}_\psi = (r_1, \ldots, r_n)$ corresponds to the recommendation for individual i. Recommendations could also be scores.

The objective here is to define the decision set $\mathbf{D}_\psi$. It is important that the decision-making process remains explainable, as we need to justify decisions to individuals. In the following discussion, we present a solution that serves as a starting point and should be further enriched with explanations. We outline various recommendations that can be constructed and specify the explanation requirements for each of them.

Decision from individuals to sets Let's consider that we have n clusters ($S_i, \ \forall i \in \mathbf{I}$). Each individual in the clusters carries a recommendation suggested by the ADMS. It is very likely to have different recommendations within a same cluster. In this case, SF is not satisfied. We propose a relaxed version of ISF where, rather than comparing x with every other individual individually, we compare x with the collective outcomes of the individuals in her cluster as a group. We can imagine that for x, we seek an outcome that is as similar as possible to the aggregated outcomes of individuals in her cluster.

Definition 5 (Relaxed ISF) *Given an individual* $x \in \mathbf{I}$ *and purpose* $\psi \in \mathcal{X}$, x *considers herself to be treated fairly with respect to* ψ *if "most" individuals she considers similar to herself are treated similarly with respect to* ψ *as she*

is treated:

$$\phi'_{\delta,\epsilon}(x,\psi) \Leftrightarrow \forall y \in S_x,\, T(M(x,\psi), agg(M(y,\psi)\ \forall y \in S_x) > \epsilon \qquad (3)$$

One way to establish a single recommendation for every set S_i is to aggregate the recommendations of all individuals in the set, as illustrated in figure 2. Let $\mathbf{R}_{\text{set}}$ be a n-dimensional vector representing aggregated recommendations for each cluster $S_i,\ \forall i \in \mathbf{I}$, for a purpose ψ. The vector $\mathbf{R}_{\text{set}} : X \to \{0,1\}$. Each element r_{S_i} in $\mathbf{R}_{\text{set}} = (r_{S_1}, \ldots, r_{S_n})$ corresponds to the recommendation for the set S_i, $r_{S_i} = 1$ if a certain proportion of individuals in S_i are labeled 1. It can be parameterized by θ, if we want absolute majority, $\theta = \frac{1}{2}$

$$\mathbf{R}_{\text{set}} \text{ is defined such as } r_{S_i} = \begin{cases} 1 & \text{if } \frac{1}{|S_i|} \sum_{x_j \in S_i} r_{x_j} > \theta \\ 0 & \text{otherwise} \end{cases}$$

If each set S_i has a unique label, we can encounter three scenarios:

- $T(r_x, r_{S_x}) > \epsilon$ and $\forall y \in S_x,\ T(r_y, r_x) > \epsilon \Rightarrow$ ISF is satisfied
- $T(r_x, r_{S_x}) > \epsilon$ and $\exists y \in S_x,\ T(r_y, r_x) \leq \epsilon \Rightarrow$ Relaxed ISF is satisfied
- $T(r_x, r_{S_x}) \leq \epsilon \Rightarrow$ Neither ISF or relaxed ISF is satisfied

The recommendations r_x proposed by the ADMS should be explainable. This involves providing explanations of the causal relationships between the subjective information provided by individuals and the final outcomes generated by the system. Additionally, the chosen aggregation method determines r_{S_x} and thus should be justified. These explanations enable to identify inaccuracies in the ADMS recommendations or erroneous or manipulative behavior of individuals. For example, explanations could give sufficient reasons that prove the dissimilarity between x and the individuals in his cluster. Here after we present another way to make individual decisions by comparison to other clusters.

Decision from sets to individuals Let's suppose we have a single recommendation for every set S_i, as computed above. Ideally, we could assign the outcome of the group to all individuals in that group: $\forall i \in S_x, d_i = r_{S_x}$. However, in reality, this process is not always straightforward for certain individuals who belong to multiple groups with conflicting outcomes.

For instance, consider the case where $x \in S_x$ with $r_{S_x} = 1$, but also $x \in S_y$ where $r_{S_y} = 0$. How should we label d_x in this situation? This scenario is illustrated in Figure 3. Here after we reconstruct the set of decisions on each individual i such as we aggregate all of the decisions that

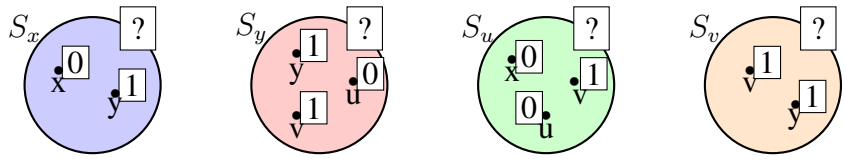

Figure 2: Example of recommended labels for individuals: some assignations are natural : $r_{S_v} = 1$ whereas it is not clear what to assign to S_y, S_u and S_x. According to majority voting $r_{S_x} = 0$, $r_{S_y} = 1$ and $r_{S_u} = 0$

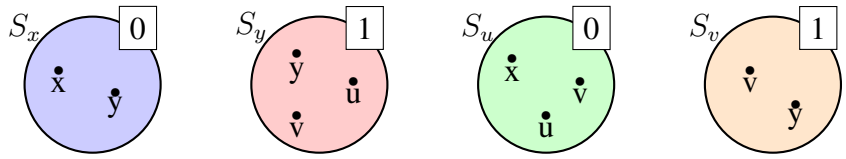

Figure 3: Example of recommended labels for clusters: some assignations are natural : $d_x = 0$ whereas it is not clear what to assign to y, u and v. According to majority voting $d_y = 1$, $d_u = 0$ and $d_v = 1$

has been made on the sets in which i belongs. The idea is to assign to d_i the majority recommendation across all clusters that i belongs to, eventually parameterized by θ as follow:

$$\mathbf{D}_\psi \text{ is defined such as } d_i = \begin{cases} 1 & \frac{1}{|\{S_j \in X | i \in S_j\}|} \sum_{\{S_j \in X | i \in S_j\}} r_{S_j} > \theta \\ 0 & \text{otherwise} \end{cases}$$

As such, this method helps mitigate manipulative behavior to some extent. For instance, if individual x strategically places themselves in a cluster with a higher likelihood of a favorable outcome, the fact that we take into consideration the recommendations of other clusters to which x belongs helps neutralize x's dishonesty. This approach results in having three possibly conflicting outcomes: r_x, r_{S_x} and d_x. When r_x and r_{S_x} are dissimilar we can either have:

- $T(r_x, d_x) > \epsilon$: ISF criteria are not satisfied but could be justified with d_x as we have to convince x that they actually identify with a group to which there are enough reasons to be dissimilar to.
- $T(r_x, d_x) \leq \epsilon$: ISF criteria are not satisfied, and the recommendation of x is inconsistent with the recommendations of individuals in her cluster and individuals that put x in their cluster. This may indicate that the ADMS has produced erroneous outcome and justifica-

tions about how this recommendation has been produced are required to clarify the decision that should be made.

The explanations to treat conflicts have a subjective dimension, as they're tailored to individuals. We can talk about subjective explanations, but this will be further explored in future work. Beyond providing explanations, individuals should have the ability to engage in dialogue based on these explanations, making the decision process dynamic and responsive to both perspectives.

5.3 Discussion

The majority rule can pose limitations and potentially lead to unfair outcomes, especially when a majority of individuals can impose a great loss on a minority. Moreover, the majority rule is susceptible to inconsistencies and paradoxes, as highlighted by concepts such as the Condorcet paradox and Arrow's impossibility theorem [13].

Other forms of aggregation may be considered. For instance, proportionality involves aggregating individual outcomes based on their proportional contribution or significance within a group. This can be achieved by assigning weights to individuals based on specific criteria. For example, individual recommendation r_x could have different weight according to the similarity between the recommendation of individual x and others in her cluster y. As such, if $T(r_x, \text{agg}(r_y, \forall y \in S_x))$ is high, it means that x accurately constructed their cluster. Consequently, we tend to trust x's recommendation and give them higher weight if they are present in the clusters of the individuals in her group. Another alternative could be a conflict resolution strategy that reconcile conflicting recommendations based on rules such as favoring the bad outcome over the good one. Also, we could look at individual objective attributes and specify rules or heuristics to resolve conflicts. For example, some rules could serve as a veto to a recommendation.

6 Related work

Following the substantial increase in literature on fairness in artificial intelligence, numerous works have adopted a critical stance, advocating for an interdisciplinary perspective and emphasizing certain shortcomings in the approaches taken [7, 34, 42, 45, 38]. In this vein, Binns [7] argues that individual and group fairness are not inherently conflicting but rather represent different approaches to addressing the same moral and political concerns.

They also contend that group fairness approaches may overlook discrimination stemming from intersectionality or groups not yet protected by anti-discrimination laws. Moreover, Fleisher [34] highlighted the limitations of individual fairness notably the fact that the similarity measure is generally chosen by the decision-maker who holds their own bias and how moral values highly influence the choice of relevant features to determine similarity. This work expands upon these ideas by integrating them with sociological insights, adding depth to the analysis.

Recognizing these limitations has driven some works to reconceptualize fairness beyond mere statistical metrics. Notably, Zafar et al. [75] proposed a notion of fairness that is not based on parity (i.e. equality of outcomes or treatment) but on preferences. However, these preferences are related to sensitive attribute groups rather than individuals. Consequently, it ensures envy-freeness at the group level, guaranteeing that no group of users would be better off by changing their group membership. Therefore, this approach still operates at a group level, which is a concern we aim to avoid in this work.

Concurrent work by Balcan et al. [4] introduced an approach for fair classification tasks drawing from the literature of fair division, particularly using envy-freeness. It is adapted for scenarios with multiple potential outcomes, not just binary ones. Unlike the work of Dwork et al. [31], which relies on a similarity function, this approach requires access to individuals' utility functions (preferences). The objective is similar to other machine learning approaches, aiming to minimize a loss function while satisfying some fairness constraint. Within this framework, the constraint is envy-freeness, which represents an individual measure of fairness that is based on the preferences of individuals rather than being a statistical measure chosen by the decision maker.

Other works that connect fairness in decision-making with the literature of fair division in social choice include the Preference-Informed Individual Fairness (PIIF) framework of Kim et al. [50]. They introduce a relaxation of individual fairness (IF) and envy-freeness (EF), whereby the primary requirement is to satisfy IF, yet it remains flexible to accommodate individuals' preferences. However, it's worth noting that envy-freeness may not always be suitable for binary outcome problems. Conversely, our framework is situated within the context of high-stakes decisions where one outcome is universally perceived as "good" and the other as "bad", thus each individual will naturally prefer the "good" outcome.

While our work shares the objective of eliminating envy, our approach diverges significantly, as we do not rely on individuals' preferences or their

utilities over policies but use explanations to guide the perceived fairness of individuals. Indeed, individuals do not feel envy towards other if the explanations about how the decision has been made convince them of their fair treatment. The fairness of a process is established if all individuals accept both their own outcomes and those of others and is situated in a dynamic framework where evolving arguments can influence perceptions of similarity and acceptance.

7 Conclusion and future work

We introduced an innovative approach that extends beyond traditional objective measures of fairness by incorporating the subjective perceptions of individuals impacted by algorithmic decisions. This new framework aligns more closely with societal realities and empower individuals to determine their fair treatment. We then propose a novel definition of fairness and present our methodology that uses explanations as a tool.

As this paper is primarily conceptual on the notion of subjective fairness, future work will focus on a more comprehensive and rigorous conceptualization of explanations and justifications. Additionally, we will explore clustering methods and develop a robust explanation framework aimed at achieving subjective fairness.

We also aim to address the challenge of treating conflicting fairness towards different stakeholders and managing conflicting explanations. Our objective is to take into account the perceptions of all stakeholders to ensure fairness for everyone. By incorporating diverse viewpoints, we develop a more inclusive and fair decision-making process that finds the trade-off between the varying interests of all parties involved.

Furthermore, we plan to conduct experiments using a real dataset to evaluate the practical applicability and fairness of our approach in real-world scenarios.

References

[1] Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias risk assessments in criminal sentencing. ProPublica, May **23** (2016)

[2] Arneson, R.J.: A defense of equal opportunity for welfare. Philosophical Studies: An International Journal for Philosophy in the Ana-

lytic Tradition **62**(2), 187–195 (1991), http://www.jstor.org/stable/4320204

[3] Arrow, K.J.: THE THEORY OF DISCRIMINATION, pp. 1–33. Princeton University Press, Princeton (1974). https://doi.org/doi:10.1515/9781400867066-003, https://doi.org/10.1515/9781400867066-003

[4] Balcan, M.F.F., Dick, T., Noothigattu, R., Procaccia, A.D.: Envy-free classification. Advances in Neural Information Processing Systems **32** (2019)

[5] Begley, T., Schwedes, T., Frye, C., Feige, I.: Explainability for fair machine learning. arXiv preprint **abs/2010.07389** (2020)

[6] Bench-Capon, T.J., Dunne, P.E.: Argumentation in artificial intelligence. Artificial intelligence **171**(10-15), 619–641 (2007)

[7] Binns, R.: On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 514–524 (2020)

[8] Van den Bos, K., Wilke, H.A., Lind, E.A., Vermunt, R.: Evaluating outcomes by means of the fair process effect: Evidence for different processes in fairness and satisfaction judgments. Journal of Personality and Social Psychology **74**(6), 1493 (1998)

[9] Bourdieu, P.: Reproduction culturelle et reproduction sociale. Social Science Information **10**(2), 45–79 (1971)

[10] Bourdieu, P., Passeron, J.C.: Reproduction in education, society and culture, vol. 4. Sage (1990)

[11] Brams, S.J., Taylor, A.D.: Fair Division: From cake-cutting to dispute resolution. Cambridge University Press (1996)

[12] Brandt, F., Conitzer, V., Endriss, U., Lang, J., Procaccia, A.D.: Handbook of computational social choice. Cambridge University Press, n.p. (2016)

[13] Brighouse, H., Fleurbaey, M.: Democracy and proportionality. The Journal of Political Philosophy **18**, 137—-155 (2010)

[14] Burke, A.S.: Improving prosecutorial decision making: Some lessons of cognitive science. Wm. & Mary L. Rev. **47**, 1587 (2005)

[15] Chapman, H., Frader, L.L.: Race in France: Interdisciplinary perspectives on the politics of difference. Berghahn Books, n.p. (2004)

[16] Chin, R., Fehrenbach, H., Eley, G., Grossmann, A.: After the Nazi racial state: difference and democracy in Germany and Europe. University of Michigan Press, n.p. (2010)

[17] Cliff, N.: Some cautions concerning the application of causal modeling methods. Multivariate behavioral research **18**(1), 115–126 (1983)

[18] Cohen, G.A.: On the currency of egalitarian justice. Ethics **99**(4), 906–944 (1989)

[19] Collins, P.H.: Intersectionality as critical social theory. Duke University Press (2019)

[20] Colorni, A., Tsoukiàs, A.: What is a decision problem? European Journal of Operational Research **314**(1), 255–267 (2024)

[21] Commission, E.: EU AI Act. Council of the EU, Press release (December 2023), https://artificialintelligenceact.eu/ai-act-explorer/

[22] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., Huq, A.: Algorithmic decision making and the cost of fairness. In: Proceedings of KDD-17. pp. 797–806. ACMPress, New York, NY, USA (2017)

[23] Crenshaw, K.: Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stanford Law Review **43**(6), 1241–1299 (1991), http://www.jstor.org/stable/1229039

[24] CROZIER, M., FRIEDBERG, E.: Le pouvoir comme fondement de l'action organisée. Théories de L'Organisation. Personnes, Groupes, Systèmes et Environnement **3**, 133 (1990)

[25] Daellenbach, H., McNickle, D., Dye, S.: Management science: decision-making through systems thinking. Bloomsbury Publishing (2017)

[26] Dahl, R.A., Birnbaum, P.: Qui gouverne ? Analyse politique, Librairie Armand Colin, Paris (1971)

[27] Danks, D., London, A.J.: Algorithmic bias in autonomous systems. In: Ijcai. vol. 17, pp. 4691–4697 (2017)

[28] Diener, E.: Subjective well-being. Psychological bulletin **95**(3), 542 (1984)

[29] Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K., Dugan, C.: Explaining models: an empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th international conference on intelligent user interfaces. pp. 275–285. ACMPress, New York, NY, USA (2019)

[30] Doherty, D., Wolak, J.: When do the ends justify the means? evaluating procedural fairness. Political Behavior **34**, 301–323 (2012)

[31] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226. ACMPress, New York, NY, USA (2012)

[32] Dworkin, R.: What is equality? part 1: Equality of welfare. Philosophy & public affairs pp. 185–246 (1981)

[33] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. pp. 259–268. ACMPress, New York, NY, USA (2015)

[34] Fleisher, W.: What's fair about individual fairness? In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. pp. 480–490 (2021)

[35] Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. Communications of the ACM **64**(4), 136–143 (2021)

[36] Frydman, C., Camerer, C.F.: The psychology and neuroscience of financial decision making. Trends in cognitive sciences **20**(9), 661–675 (2016)

[37] Gentelet, K., Bahary-Dionne, A.: Stratégies des premiers peuples au canada concernant les données numériques: décolonisation et souveraineté. Tic&société **15**(1| 1er semestre 2021), 189–208 (2021)

[38] Gölz, P., Kahng, A., Procaccia, A.D.: Paradoxes in fair machine learning. Advances in Neural Information Processing Systems **32** (2019)

[39] Habermas, J.: Moral consciousness and communicative action. MIT press (1990)

[40] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)

[41] Hébert-Johnson, U., Kim, M., Reingold, O., Rothblum, G.: Multicalibration: Calibration for the (computationally-identifiable) masses. In: International Conference on Machine Learning. pp. 1939–1948. PMLR (2018)

[42] Hoffmann, A.L.: Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society **22**(7), 900–915 (2019)

[43] Honneth, A., Joas, H.: Communicative action: essays on Jürgen Habermas's The theory of communicative action. MIT Press (1991)

[44] Jameux, C.: Pouvoir et décision, politique et stratégie. Sciences de la Société **38**(1), 159–163 (1996). https://doi.org/10.3406/sciso.1996.1288, https://www.persee.fr/doc/sciso_1168-1446_1996_num_38_1_1288, included in a thematic issue : Pouvoir et dynamique des organisations (1). Etat des lieux et des savoirs

[45] John-Mathews, J.M., Cardon, D., Balagué, C.: From reality to world. a critical perspective on ai fairness. Journal of Business Ethics **178**(4), 945–959 (2022)

[46] Kahneman, D., Krueger, A.B., Schkade, D., Schwarz, N., Stone, A.: Toward national well-being accounts. American Economic Review **94**(2), 429–434 (2004)

[47] Kasy, M., Abebe, R.: Fairness, equality, and power in algorithmic decision-making. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 576–586. ACMPress, New York, NY, USA (2021)

[48] Kaufmann, H.: Legality and harmfulness of a bystander's failure to intervene as determinants of moral judgment. Altruism and helping behavior pp. 21–57 (1970)

[49] Kertzer, D.I., Arel, D.: Census and identity: The politics of race, ethnicity, and language in national censuses. No. 1 in New Perspectives on Anthropological and Social Demography, Cambridge University Press (2002)

[50] Kim, M.P., Korolova, A., Rothblum, G.N., Yona, G.: Preference-informed fairness. arXiv preprint arXiv:1904.01793 (2019)

[51] Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 (2016)

[52] Leventhal, G.S.: What should be done with equity theory? new approaches to the study of fairness in social relationships. In: Social exchange: Advances in theory and research, pp. 27–55. Springer (1980)

[53] Lind, E.A., Kanfer, R., Earley, P.C.: Voice, control, and procedural justice: Instrumental and noninstrumental concerns in fairness judgments. Journal of Personality and Social psychology **59**(5), 952 (1990)

[54] Lind, E.A., Tyler, T.R.: The social psychology of procedural justice. Springer Science & Business Media (1988)

[55] Madan, A.: Sociologising merit. Economic and Political Weekly pp. 3044–3050 (2007)

[56] Makhlouf, K., Zhioua, S., Palamidessi, C.: Survey on causal-based machine learning fairness notions. arXiv preprint arXiv:2010.09553 (2020)

[57] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) **54**(6), 1–35 (2021)

[58] Moulin, H.: Fair division and collective welfare. MIT press (2004)

[59] Nobles, M.: Racial categorization and censuses. Census and identity: The politics of race, ethnicity, and language in national censuses p. 43 (2002)

[60] Obst, P.L., White, K.M.: Choosing to belong: The influence of choice on social identification and psychological sense of community. Journal of community psychology **35**(1), 77–90 (2007)

[61] Raboun, O., Chojnacki, E., Tsoukiàs, A.: Dynamic-r: a "challenge-free" method for rating problem statements. Annals of Operations Research **325**(2), 845–873 (2023)

[62] Rawls, J.: A theory of justice: Revised edition. Harvard university press (2001Ò)

[63] Roemer, J.E.: Theories of distributive justice. Harvard University Press (1996)

[64] Sen, A.: Equality of what? Cambridge: Cambridge University Press. (1979)

[65] Simon, H.A.: Rational decision making in business organizations. The American economic review **69**(4), 493–513 (1979)

[66] Simon, H.A.: Administrative behavior. Simon and Schuster (2013)

[67] Suresh, H., Guttag, J.V.: A framework for understanding unintended consequences of machine learning. arXiv preprint arXiv:1901.10002 **2**(8) (2019)

[68] Tsoukiàs, A.: On the concept of decision aiding process: an operational perspective. Annals of Operations Research **154**(1), 3–27 (2007)

[69] Tsoukiàs, A.: From decision theory to decision aiding methodology. European journal of operational research **187**(1), 138–161 (2008)

[70] Tsoukiàs, A.: Social responsibility of algorithms: an overview. EURO Working Group on DSS: A Tour of the DSS Developments Over the Last 30 Years pp. 153–166 (2021)

[71] Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. p. 1–7. FairWare '18, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3194770.3194776, https://doi.org/10.1145/3194770.3194776

[72] Weber, M.: The distribution of power within the political community: Class, status, party. Economy and society **2**, 926–940 (1978)

[73] Yuval-Davis, N.: Situated intersectionality and social inequality. Raisons politiques pp. 91–100 (2015)

[74] Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th international conference on world wide web. pp. 1171–1180 (2017)

[75] Zafar, M.B., Valera, I., Rodriguez, M., Gummadi, K., Weller, A.: From parity to preference-based notions of fairness in classification. Advances in neural information processing systems **30** (2017)

[76] Zhao, Y., Wang, Y., Derr, T.: Fairness and explainability: Bridging the gap towards fair model explanations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 11363–11371 (2023)