Wilfried Hinsch*

Framing Computational Fairness and Non-Discrimination

https://doi.org/10.1515/auk-2026-3001

Abstract: This article outlines a new framework of computational fairness for the analysis of data-driven statistical discrimination. It begins by identifying the limitations of the still dominant paradigm of *fairness through unawareness*. Building on this critique, the article proposes an approach that distinguishes concerns of procedural fairness from questions of distributive justice when evaluating potentially discriminatory procedures. In contrast to a widespread view, it argues that procedural computational fairness rests on a single formal criterion. A computational procedure is non-discriminatory, it is claimed, if it achieves equal predictive values across demographic groups. In addition, two non-procedural requirements of distributive justice are advanced for selective procedures: the proportionality of individual burdens and their compatibility with a just overall social distribution of benefits and burdens.

Keywords: statistical discrimination; 'fairness through unawareness'; procedural vs. distributive justice; conditional use vs. conditional procedure accuracy; 'impossibility of fairness'

Gert Wagner *in memoriam*

1 Statistical Discrimination

Companies and government agencies increasingly rely on advanced data-analytics and computational models for recruitment, credit ratings, supervision, and other purposes to make decisions that have a serious impact on people's lives. These models function through predictive algorithms that generate probabilistic assessments – typically in the form of scores – concerning, for example, the productivity of job

***Corresponding author: Wilfried Hinsch**, Universität zu Köln, Philosophisches Seminar/ Wissenschaftsforum zu Köln und Essen, Cologne, Germany, E-mail: Whinsch@uni-koeln.de

applicants, the credit worthiness of bank customers, or the likelihood that a person will engage in criminal behavior. It has become apparent and a matter of public concern that members of certain demographic groups – classes of individuals with shared attributes – frequently receive poorer ratings than what appears fair and reasonable. Thus, women are ranked lower than men when applying for leadership positions, residents of impoverished neighborhoods are deemd less reliable debtors, and Black people are classified as more likely to commit crimes. This reflects familiar patterns of illicit discrimination (see e.g., O'Neill 2016; Eubanks 2018).

The apparent unfairness in the outcomes of much computational scoring may seem surprising. After all, the devices that generate these assessments are passionless machines rather than humans. They lack the mean-spirited preferences and attitudes of sexists, racists, or those who simply disdain people from disadvantaged backgrounds. Indeed, the discriminatory patterns produced by computational scoring are not a matter of ordinary discrimination. They derive from *statistical discrimination* – a problem inherent in the way computational scoring models operate (Arrow 1972; Phelps 1972).

Such models generate probabilistic estimates regarding targeted personal traits and the expected future behavior of individuals. These estimates are derived from the individuals' membership in demographic groups that exhibit characteristic statistical patterns. Members of such groups share certain personal attributes that serve as predictors and provide the basis for conjectures about future conduct and performance. For example, if the group of individuals holding more than two credit cards contains a higher proportion of members with a history of payment difficulties than the group holding only one or two cards, then members of the first group will be assigned a correspondingly higher probability of insolvency.

The comparative evaluation of individuals on the basis of predictive features that rest on statistical patterns of classes of people who share those features was neither introduced by computational scoring nor is it, in general, impermissible. It is a universal cognitive strategy of rational expectation to infer from observable personal features, such as gender, to intangible traits and anticipated behavior on the basis of statistical correlations that support probability estimates (whether intuitively or formally calculated). If I have been attacked by dogs several times, I will naturally come to anticipate further attacks. Similarly, if women have experienced discrimination from employers in the past, it is reasonable for them to expect discrimination in the future. We move from what is known – or what we can easily find out – to what we do not yet know but may reasonably expect.

Statistical discrimination differs in important respects from ordinary forms of discrimination. It need not be grounded in objectionable attitudes, false beliefs, or

miscalculations. For instance, employers may or may not care about ethnicity or gender as such, yet they have a legitimate interest in obtaining reliable indications of the prospective performance of job candidates. If readily observable personal features provide statistical support for probability estimates concerning less tangible traits such as future productivity, it is reasonable for employers to take them into account in hiring decisions. The same applies to bank managers, security officers, and other decision-makers who make selective choices that impose non-negligible burdens or disadvantages on individuals. Their concern with characteristics such as gender, age, or place of residence does not necessarily stem from valuing these features in themselves, but rather from the need to infer other, intangible traits – such as professional performance or creditworthiness – that can only be assessed indirectly, by means of statistical and probabilistic reasoning on the basis of such features.

Nevertheless, computational scoring raises questions of fairness and justice even when it is not grounded in prejudice or false belief but in sound statistical correlations. The results of such scoring are probabilistic and thus inherently under- and over-inclusive. There will always be individuals who possess the targeted intangible feature that the model seeks to predict but remain undetected because they lack the respective predictive features – these are the so-called False Negatives (FNs). Conversely, there will be individuals who display the predictive features yet do not possess the targeted trait – the so-called False Positives (FPs). Consider, for example, the case in which women, more frequently than men, abandon promising professional careers for family-related reasons. Employers may then come to expect that female employees will resign from leadership positions earlier and, on that basis, hesitate to promote or even hire women in the first place. Such practices, however, appear unfair to the well-qualified and ambitious young woman who has never contemplated leaving her profession to raise children or support a spouse. Be fair, she may urge a prospective employer: *Don't judge me by my group!*

This example illustrates that the challenge of statistical discrimination lies less in the logic of probabilistic reasoning than in the informational basis upon which such reasoning rests. Probability estimates based on too little information are bound to be not only unreliable but also to some extent unfair. Note that the female applicant in the example is not simply a member of a single demographic category – the class of all women – but simultaneously of multiple overlapping groups defined by predictive features such as age, education, biography, and professional record, of which gender is only one. Even if all these features were taken into account as predictors, however, and even if they counted in favor of a candidate, the fact of being a woman would still count statistically against female applicants,

provided that women – in a greater number than men – exit professional careers at earlier stages. And one may still have qualms about this.

A first concern is one of procedural fairness. It may appear unfair to evaluate the productive capacities of a person on the basis of their gender, a characteristic that, to the best of our knowledge, has no direct causal bearing on individual productivity. To rely on such a feature would seem to violate a requirement of substantive relevance, namely that people's capacities should be assessed according to attributes which actually determine future performance rather than on the basis of merely statistical correlations.

A second concern is distributive justice. Hiring decisions that are grounded in statistical information about demographic groups tend to reinforce existing social inequalities between those groups. Members of groups the majority of which – for whatever reason – has been economically less successful than the majority of members of other groups, will be, for statistical and probabilistic reasons – regarded as less promising candidates for economic success and, hence, be offered less attractive positions. As a consequence, in comparison with other groups fewer members of those groups will attain these positions, hence, we will find women less often in leadership positions than men.

Both concerns of procedural fairness and distributive justice give rise to conceptual and normative questions that will be addressed in the following sections. Section 2 revisits the shortcomings of *fairness through unawareness*, a widely received strategy for combating discrimination that continues to shape much of the current public discourse and anti-discrimination law. Section 3 introduces a framework for understanding computational fairness and non-discrimination that situates wrongful discrimination within a broader landscape of injustices, distinguishing sharply between failures of procedural fairness and failures of distributive justice. Building on these distinctions, Section 4 turns to the formal requirements of procedural fairness. Contrary to prevailing assumptions that there is a plurality of valid but partly incompatible fairness criteria – the impossibility of (all-out) fairness thesis – it is argued that there exists a unique formal measure of procedural computational fairness which is based on the predictive values of a selective algorithm. Section 5 discusses the requirements of distributive justice which apply to selective procedures and practices.

2 Fairness Through Unawareness

The intuitive rationale underlying the anti-discrimination strategy *fairness through unawareness* is straightforward: if a decision-maker – or, for that matter, a computational scoring model – has no access to information about an individual's gender,

skin color, or ethnic origin, then discrimination on the basis of these features would appear impossible. According to the strategy, attributes such as gender, race, and other personal characteristics of groups that have historically suffered, and continue to suffer, discrimination are regarded as suspect grounds of discrimination. For this reason, they have been designated as 'protected attributes' that must not be taken into account when evaluating individuals or making selective decisions, such as in hiring or commercial lending. In line with this understanding, the scoring algorithms of Germany's largest credit bureau, *Schufa*, exclude gender from their set of predictive variables. National and international legal frameworks designed to shield groups of individuals (so-called protected groups) from unlawful discrimination typically include binding lists of such attributes. A notable example is provided by the *International Covenant on Civil and Political Rights*:

> In this respect, the law shall prohibit any discrimination and guarantee to all persons equal and effective protection against discrimination on any ground such as race, color, sex, language, religion, political or other opinion, national or social origin, property, birth or other status. (ICCPR 1966/1976, art. 26.2)

Fairness through unawareness and the legal concepts of 'protected attributes' and 'protected groups' which are based on this conception have significantly contributed to the fight against unfair discrimination since the 1960s. With the development of more and more sophisticated models of computational scoring, however, the shortcomings of the strategy and its legal implementation are becoming increasingly visible.[1]

Given the innumerable ways in which individuals make selective choices, and the wide range of predictive features that may be employed in computational scoring, the question arises: how can we establish a principled distinction between characteristics that constitute acceptable reasons for differential treatment and those that do not?

The attributes included on the list of publicly and legally recognized grounds of discrimination typically relate to groups that have most strikingly suffered illicit discrimination in the past and that continue to face unfair disadvantages today. This historical experience motivated demands to protect individuals with these characteristics. Recourse to historically disadvantaged demographic groups and to structural inequalities between groups, however, does little to illuminate the specific wrong of discrimination itself, since such an approach already presupposes criteria not only of unfair differential treatment of individuals but also of unjustified inequality between groups. Without such criteria we are running the risk of

1 See Cornacchia et al. 2023; Fabris et al. 2023; Ruf and Detyniecki 2020. For a more positive appraisal of *fairness through unawareness*, see Höltgen and Oliver 2025.

moving in circles by explaining discrimination as wrongful treatment of discriminated groups.

Moreover, explaining individual discrimination primarily in terms of group disadvantage distorts the social dynamics of discrimination. Unfair discriminatory practices do not require the prior existence of unjust structural inequalities but generate and reinforce such inequalities. This point is especially salient in the context of computational scoring, which typically relies on large numbers of predictive personal features which define highly specific groups of individuals that may be, unlike women or people of color, neither socially visible groups nor groups that were disadvantaged before the advent of computational scoring. Yet their members may nonetheless receive unfair scores and thereby suffer unjustified disadvantages.

We must also bear in mind that not all discriminatory practices rely on personal characteristics recognized as 'protected attributes'. Much social disadvantage stems from characteristics like social background, educational record, or personal lifestyle that are not included in the standard lists of suspect grounds of discrimination and legally protected attributes. Conversely, not all forms of unequal treatment based on protected attributes amount to unlawful discrimination. Under certain conditions, characteristics generally regarded as suspect may constitute legally defensible grounds for differential treatment. For example, charging young drivers higher insurance premiums is not considered unjustified age discrimination, since it reflects differences in claims risk. It is also not evident why a protected attribute such as gender should categorically be excluded as a statistically relevant predictor – say of economic performance or creditworthiness – if it functions not as the sole predictive feature, but merely as one among an indefinite number of others.

Note also that the exclusion of a protected attribute from computational scoring does not necessarily work to the advantage of the groups it is intended to protect. On average, women cause traffic accidents less often compared to men, yet because of gender blindness they pay the same for insurance. Considering credit scoring, the empirical evidence is inconclusive (Li 2018), but suggests that women – who, on average, appear to be more risk-averse and financially more disciplined than men – may often represent a lower credit risk (Goodman et al. 2016). Including the prohibited attribute of gender could, therefore, improve women's credit ratings.

Finally, from a practical standpoint, the strategy to protect individuals from discrimination by designating certain features as protected attributes appears futile, given the innumerable correlations among potentially predictive variables. Any dataset rich enough to contain all or most of available relevant information will, almost inevitably, include an indefinite number of proxies for every protected

attribute, thereby effectively re-identifying members of groups the exclusion was intended to shield.

To a certain degree, both legal and public discourse on discrimination acknowledge these misgivings. Thus, the framework of *fairness through unawareness* allows for the expansion of the list of protected attributes (for example, to include different gender identities or physical and mental impairments). It also permits differential treatment where this serves a legitimate purpose and does not impose disproportionate burdens. For example, Germany's *Allgemeines Gleichbehandlungsgesetz* from 2006 permits selective decisions when "the relevant provisions, criteria or procedures are objectively justified by a legitimate aim and the means of achieving that aim are appropriate and necessary" (AGG §3.2). Despite such qualifications, however, the prevailing discourse has yet to draw systematic conclusions from the shortcomings of *fairness through unawareness* and its protected-attribute model. In both public debate and legal practice, suspect features continue to be treated as defining markers of unfair discrimination and are broadly excluded from computational scoring – regardless of whether such exclusions promote or undermine legitimate stakeholder interests in fair procedures and just outcomes.

To conclude this section: *fairness through unawareness* rests on a misleading and inadequate conception of discrimination, and of statistical discrimination in particular. It is misleading insofar as it suggests that the wrong of discrimination can be captured by short lists of reasonably well-defined suspect attributes. And it is inadequate because it fails to provide the criteria necessary for drawing a reasonably clear line between permissible and impermissible forms of differential treatment. Whether a personal characteristic justifies unequal treatment cannot be determined in abstraction, but only in relation to the purpose and context of the selective procedure in question. For this reason, the distinction between protected and unprotected attributes – which lies at the core of *fairness through unawareness* – loses its fundamental normative significance. A different theoretical framework is therefore required for a proper understanding of statistical discrimination.

3 Discrimination and Justice

There are countless inequalities among individuals, and many reasons why some people are treated differently from others. Some inequalities, and some grounds for differential treatment, are harmless and morally irrelevant. And much inequality that is neither harmless nor negligible is nevertheless tolerable, or even desirable. Structural inequalities that arise from practices of illicit discrimination, however, are neither innocuous nor desirable. They are inherently unjust.

It is helpful to distinguish four types of moral concerns with regard to selective computational scoring. Two of them – legitimacy of purpose and equal treatment – are matters of procedural fairness, the two others – proportionality of individual burdens and outcomes consistent with a just social distribution of benefits and burdens – pertain to distributive justice. As we have seen, probabilistic scoring can never achieve perfect fairness or justice, since there will always be False Positives and False Negatives. Yet even an imperfectly just procedure must satisfy the requirements of both procedural fairness and distributive justice.[2]

Two assumptions will inform our discussion throughout: The first is that serious inequalities of income, wealth, or opportunity which result from human practices and institutions are in need of public justification. The second assumption is that the primary sources of moral claims are persons and not groups of persons. Unlike groups, natural persons are considered to be "self-authenticating sources of valid claims" (Rawls 1993, 32–3). Groups like soccer teams or political parties that qualify as collective agents may still have valid moral and legal claims. These claims must be explained, however, with recourse to the more basic claims, rights and duties of natural persons. Both assumptions may be contested, but their implications for the argument in this article are easily recognized and may not cause much controversy.

3.1 Social Practices of Discrimination

Predictive algorithms and computational models used to support selective procedures – such as recruitment or commercial lending – are abstract mathematical constructs that, in themselves, cannot be judged fair or unfair, just or unjust. Still, they become objects of moral assessment when embedded within social structures and practices that profoundly shape people's opportunities and life prospects. Discrimination would not exist apart from the unfairly discriminatory choices made by human agents. Yet such conduct is never merely a matter of isolated individual wrongdoing. Discrimination is an inherently social phenomenon, unfolding against a backdrop of shared beliefs, values, and attitudes that inform collective practices as much as individual actions.

The decision of a single employer to reject a well-qualified applicant because she is a woman is objectionable for a variety of reasons – for instance, as an expression of disrespect toward women, or as a failure of diligence and even-mindedness. Yet if it were only one employer (or only a few) acting on such a preference for

2 'Fairness' and 'justice' are often used interchangeably. To draw a clear distinction between different types of social justice concerns, however, I shall use 'fairness' to refer to procedural matters, and 'justice' to refer to criteria by which the outcomes of procedures are assessed.

male applicants, their isolated decisions would not, I submit, constitute the specific wrong of discrimination. If other employers did not share these attitudes or recruitment policies, the bias of one employer, though frustrating and unfair to the women concerned, would not give rise to the kind of disadvantages and burdens that characterize illicit discrimination. Indeed, a gender bias of only a handful of people may not impose exceptional burdens on women at all. An unfairly rejected female applicant could easily find employment with others. There is a crucial difference, however, between being rejected for no good reasons at some places and being rejected at many places or nearly everywhere. The wrongness of discrimination derives not simply from individual wrongdoing but from the cumulative effects of unfair practices and procedures. It, therefore, raises questions not only of transactional justice and procedural fairness to individuals but also of distributive justice between groups of people.

3.2 Unreasonable Burdens

It is widely held that wrongful discrimination consists in adverse treatment on the basis of personal characteristics – such as gender – that ought to make no difference in selective practices like hiring or commercial lending. The personal attributes that guide differential evaluations and decisions shape both the disparate individual outcomes and the cumulative collective effects of such practices. For this reason, they rightly occupy a prominent place in any account of discrimination. Yet, as we have seen, they do not provide the ultimate criteria for distinguishing fair from unfair differential treatment. Explaining why a person is treated adversely is one matter; explaining what makes such treatment wrongful is another. The reason why a disadvantage based on specific personal characteristics is objectionable is not to be found in the respective features. Instead, it is grounded in the unreasonableness of the disadvantages and burdens which are imposed on persons because of these characteristics. Discriminatory practices can, therefore, be defined as practices that (a) track personal characteristics and that (b) lead to unreasonable disadvantages and burdens imposed on people with these features.

The disadvantages imposed on persons by discriminatory computational scoring may be unreasonable in two principal ways. Firstly, a scoring algorithm may fail to produce scores that justify placing a person at a disadvantage in the first place. Secondly, even where the scores would justify a disadvantage, the burdens resulting from it may still be disproportionate and unjust considering the broader social distribution of benefits and burdens. A model that fails to provide adequate reasons for differential treatment lacks procedural fairness. Such a model is incapable of justifying the outcomes it produces in terms of the distribution of benefits and burdens. Yet even a model that is procedurally fair – one that offers sound reasons for

treating individuals differently – may still generate individual or cumulative outcomes that fall short of reasonable demands of distributive justice. A comprehensive moral evaluation of selective practices must therefore integrate both perspectives: procedural fairness and distributive justice.

Since our subject is computational fairness and wrongful discrimination, rather than illicit practices in general, we may set aside scoring models that serve immoral or unlawful purposes. The focus here is on practices with legitimate aims, such as screening job applicants or assessing the creditworthiness of bank customers. Let us further assume that the targeted personal characteristics a computational model is designed to predict – such as future job performance or financial solvency – are features that employers or banks may reasonably consider when making hiring or lending decisions. Under these conditions, procedural fairness may seem to reduce to a question of adequate data input and sound probabilistic reasoning. And, indeed, if a model conforms to the principles of statistical inference, and if the data it processes contain reliable and representative information about the distribution of predictive and targeted characteristics within a given population, then the scores it generates will constitute reliable indicators of what employers or bank managers may reasonably expect from applicants, thereby providing legitimate grounds for hiring and credit decisions.

However, decision-making on the basis of fair evaluations also involves a formal requirement of equal treatment: individuals who equally satisfy the criteria deemed relevant to the purpose of a selective practice ought to be treated equally. This requirement applies to both computational and non-computational procedures. Note, however, that the meaning of 'relevant criteria' undergoes a fundamental shift when it comes to statistical and probabilistic scoring. In a non-computational context, 'relevant criteria' are defined substantively – that is, in semantic or causal terms – in relation to the purpose of a selective procedure. If the purpose is recruitment, relevant factors are those that contribute to professional performance, such as education, experience, or ambition. If the purpose is to assess creditworthiness, relevant factors are those that contribute to reliability and financial solvency, such as employment status, disposable income, or thriftiness. On such a substantive understanding of relevance, deciding between equally qualified applicants on the basis of gender clearly violates the equal treatment requirement, since it is difficult to see how the isolated characteristic of gender by itself could causally account for differences in either economic productivity or financial reliability.

In contrast, with computational scoring 'relevance' is no longer substantive but statistical and predictive relevance. On this understanding, a tangible personal characteristic is relevant if it correlates statistically with the targeted feature and supports an individual probability estimate regarding that feature: Different

individuals are considered equal cases if they share the same set of predictive features, regardless of whether those features bear a direct causal relation to the targeted trait or not.

Seen in this light, the mere fact that equally qualified women may receive lower scores than men in itself does not imply a violation of equal treatment. If gender is statistically predictive of a targeted feature such as job performance or creditworthiness – and empirical evidence suggests that it is – it may legitimately count in favor of or against a candidate, notwithstanding the fact that gender is not itself a causal determinant of productivity or creditworthiness.

A computational model that generates evaluative scores solely on the basis of predictive features might thus appear to satisfy the equal treatment requirement of procedural fairness almost by definition, since it never relies on features that are predictively irrelevant. Yet this appearance is misleading. Not all predictors possess the same predictive value, some produce more False Positives or False Negatives than others. Moreover, the predictive value of a given feature or set of features is not uniform across demographic groups. For instance, in recruitment, computational assessments of cognitive abilities often have a lower predictive value for women than for men (Rothstein and McDaniel 1992). Because fewer women than men occupy leadership positions, there is less data available on professionally successful women. This makes probability estimates for female applicants less reliable. Predictions of high or low professional performance will yield more False Positives or False Negatives for women than for men and, thus, expose women to a comparatively higher risk of being misclassified as unpromising candidates which seems unfair. The purely formal criterion of predictive relevance, thus, goes along with a risk-based conception of procedural fairness: Equal treatment does not any longer require identifying characteristics that are causally relevant to the purpose of a selective practice. Rather, it requires that the predictive features employed have equal predictive value across demographic groups, thereby ensuring equal chances of correct classification – or, equivalently, equal risk of misclassification.

4 Risk-Based Procedural Fairness

For any individual subjected to a scoring procedure, the error risk may be too high in absolute or in relative terms. A procedure may be unfair – but not necessarily discriminating – because it imposes an unreasonably high error risk on individuals irrespective of their group membership. And it may be unfair and also discriminating because it imposes unequal error risks on the members of one demographic group in comparison with another. There are two requirements of computational

		True Value	
		Positive	Negative
Predicted Value	Positive	*True Positives* TPs	*False Positives* FPs
	Negative	*False Negatives* FNs	*True Negatives* TNs

Figure 1: Confusion matrix.

fairness, then, one monadic, the other relational: Reasonable predictive success and equal predictive success across demographic groups.

4.1 Recognition Rates and Predictive Values[3]

The rationale of demanding reasonable and equitable predictive success for computational scoring seems obvious enough. Predictive success, however, may be defined and measured in different ways. Two prominent measures are recognition rates and predictive values in the sense of the *Conditional Procedure Accuracy* (CPA) and the *Conditional Use Accuracy* (CUA) of a computational model.[4] Recognition rates and predictive values are calculated from the entries in the confusion matrix of a predictive model (Figure 1).

A confusion matrix is a mapping of the predicted and true values of a target variable that tells us how often the predicted value matches the true value and how often it does not. It is based on a binary classification. The value of the target variable is either *P* (Positive) or *N* (Negative) and the prediction may be correct or wrong. Thus, we have *True Positives* (TP), *False Positives* (FP), *True Negatives* (TN), and *False Negatives* (FN). Since we are dealing with predictive models which generate continuous probabilities and not binary values, a selective computational

3 The following sections revisit my argument in Hinsch 2023, 248–50, and explain why the earlier article's claim that the combined recognition rates of a scoring model provide an appropriate index of fairness was mistaken.

4 With regard to computational fairness, recognition rates and predictive values should always be understood in terms of *Conditional Procedure Accuracy* and *Conditional Use Accuracy*, i.e., as combining both positive and negative recognition rates (TPR & TNR) and predictive values (PPV & NPV). The positive and negative rates and values are logically independent and can vary freely against each other. All rates and values may raise questions of social justice. The risk of an unjustified disadvantage due to False Negatives must be considered just as much as the risk (or chance) of unjustified advantage due to False Positives (Hinsch 2023, 246). The focus in this article, however, is solely on unfair disadvantages that arise from the rate of False Positives.

procedure has to incorporate a mapping function which defines a threshold and translates probabilities into binary values. In any case, computational fairness is not a matter of true or false probability estimates – which, after all, may have no truth value – but a matter of reasonably reliable predictions which are supported by sound estimates.

If the focus is solely on cases in which an individual is wrongly predicted to possess a characteristic that justifies adverse treatment, the procedural fairness of a computational model must be evaluated by reference to the risk that a person will be misclassified as a Positive. The question, then, is how best to measure this risk.

At first glance, both recognition rates and predictive values appear equally plausible candidates and *Conditional Procedure Accuracy* (CPA) and *Conditional Use Accuracy* (CUA) are widely regarded as equally reasonable measures of the risk of individual misclassification and thus of exposure to unjustified disadvantage. In practice, however, recognition rates and predictive values cannot typically be maximized simultaneously. As a rule, it is impossible to achieve parity with respect to both measures. Since models can be tuned to improve one measure only at the cost of the other, trade-offs are inevitable. The crucial question, thus, is whether the risk of misclassification – and with it the fairness of the model – should be assessed in terms of recognition rates or in terms of predictive values.

The problem is neatly summarized in the impossibility of fairness thesis. If CPA- and CUA-parity are equally valid criteria of fairness which, in general, cannot be satisfied together, unambiguous procedural fairness is nothing that we may reasonably expect from a computational scoring model (Hardt et al. 2016; Kleinberg et al. 2016; Chouldechova 2017; Barocas et al. 2023). From the viewpoint of legal regulation, the impossibility thesis has the unfortunate consequence to invite what may be called *fairness shopping*. Since it is hard to see how specific contextual features of an application could decide about a model's procedural fairness, in practice, developers will have a free choice. For any given application they may either optimize a model's combined recognition rates or its predictive values in order to meet regulatory standards of non-discrimination, e.g., the self-certification requirements of the EU-AI-Act. And we may safely assume that, as a rule, the preferred corporate choices will reflect economic interests rather than considerations of social justice.

From a moral point of view, it may be good news, then, that the impossibility of fairness thesis proves to be mistaken. CPA and CUA differ profoundly in their suitability as conditions of procedural fairness, both in relation to the calculation of individual error risk (monadic fairness) and in relation to the appropriate parity requirements (relational fairness). Only CUA yields a free-standing measure of

procedural fairness and constitutes an indisputable necessary condition of non-discrimination.[5]

4.2 CPA and CUA as Monadic Measures of Reasonable Error Risks

Consider the *False Positive Rate* (FPR) of a predictive algorithm, i.e., the ratio of False Positives to the sum of False Positives and True Negatives (FP/FP + TN = FP/*N*). FPR is a conditional probability conditioned on the actual characteristics of persons (*P* v *N*). The risk assessments we derive from it are thus confined to persons who are already known to be *N*s. Off-hand, using FPR as a measure of reasonable success rates would, therefore, not seem to make much sense: who would use probabilistic scoring to make individual assessments, if it is already known who are the *Negatives* and who are the *Positives*. Moreover, FPR would miscalculate the error risk for the *Negatives*, because for the members of this group the probability of being an FP, if predicted *P*, is *by definition* 1 and not the False Positive Rate.

We may still derive a sound probability estimate for error risks. In fact, it is not necessary to know whether somebody is a *P* or *N* in order to calculate the risk of a misclassification if we know the base-rate (BR) or prevalence of the target variable *P* (α) in a representative test sample. Given that there is no risk of an unjustified disadvantage for *P*s, it is $(1 - \alpha)^*$FPR. Thus, we can arrive at a meaningful individual probability estimate using recognition rates. It may, however, not be the best estimate that we can derive from the information contained in the confusion matrix of a scoring model. Following Carnap's Principle of Total Evidence (Carnap 1950; Hempel 1965), we must consider all available information to arrive at reliable probability estimates for individuals. A CPA-measure of reasonable predictive success does not meet this requirement. It ignores the available information about the group of individuals who have been tested positive (TP v FP). If this information is considered, we can calculate the predictive value of a model and its *False Discovery Rate* (FDR = FP/TP + FP). While FDR conforms to Carnap's condition of sound probability estimates for individual cases, the *False Positive Rate* does not. And it would be a mere coincidence if FDR would equal $(1 - \alpha)^*$FPR if we take it that the group of people who are classified as positive (TP v FP) is a proper subset of the population *P* v *N*.

5 Say, a fairness measure is free-standing, if it can be used without recourse to another measure.

4.3 Asymmetric Information

An instructive complication may seem to arise in situations of asymmetric information. Think of an airline passenger who knows that he is not carrying a bomb and informs airport security about this. Still, they may not be sure that the bearded man from the Middle East with peace in his heart is telling the truth. Since he knows that he is neither a terrorist nor a liar, the passenger can calculate his personal risk of a misclassification as a terrorist once he knows the recognition rates of the model that may misclassify him at the airport. His estimate is: p_i (FP|FP&TN) = x. The security staff, however, not knowing whether the passenger is N or P, would calculate the passenger's risk of a wrong classification on the basis of FPR differently as $(1-\alpha)^*$FPR. Since p_i (FP|FP&TN) = $(1-\alpha)^*$FPR would be a mere coincidence, we most likely get two different risk assessments in the case of asymmetric information. Which one should be decisive? Should the airport security treat passengers based on what they have reason to be uncertain about or based on what passengers have reason to believe about themselves? The claim that the risk of a computational misclassification must be calculated on the basis of the actual outcome appears well-grounded only on the assumption that in the case of asymmetric information the stakeholder perspective prevails. A general demand, however, that we must never act on what we have reason to expect from others if it differs from what they have reason to expect from themselves, would not seem defensible.[6]

4.4 CPA and CUA-Parity and Non-Discrimination[7]

In general, both unequal recognition rates and predictive values across different demographic groups indicate higher risks of false predictions – and, thereby, statistically unjustified disadvantages – for the members of one group. Thus, CPA- and CUA-parity may seem equally plausible criteria of relational procedural fairness, but they are not.

Let us, this time, first consider the parity condition for predictive values. If the predictive values of a model, say to assess the creditworthiness of bank customers, are lower for women than for men, then women face a higher risk of being disadvantaged than men – either through the rejection of a loan application or higher

6 Note also that the condition of at least one-sided reliable knowledge is not fulfilled in many cases. Lie detectors and security screenings to uncover terrorists are special cases in this regard. If we think of creditworthiness or future professional productivity, the uncertainty is symmetrical. Indeed, a bank may have more reliable information to calculate the probability that a requested loan will be paid-off in due time than their customers.

7 In practical terms, 'parity' need not require strict equality of predictive values or recognition rates across different demographic groups but allow for a range of tolerable differences.

interest rates on an approved loan. Thus, parity of predictive values between men and women – indeed across all demographic groups – is an obvious and indisputable requirement of relational computational fairness and non-discrimination.

To see why this is not equally true for CPA-parity, consider the interdependencies that exist between the CPA and CUA values of a computational model in relation to the base rate or prevalence of a target characteristic in different demographic groups. If a model has equal predictive values across demographic groups, its recognition rates necessarily follow the base rate of the target characteristic within each group. Any change in the combined recognition rates (TPR & TNR) implies a change in the error distribution (TP/FP & TN/FN) of the model. Parity in recognition rates can only be expected if base rates are equal in all demographic groups if the predictive values of a model remain constant. In practice, however, the assumption of equal base rates across demographic groups is rarely met.[8]

It is true, though, that the predictive values of a model also follow its recognition rates if again we assume equal base-rates. There can also be no change in the combined predictive values (PPV & NPV) without a corresponding change in combined recognition rates (TPR & TNR), provided that the base rate remains constant. Given this symmetrical dependency – with equal base rates the recognition rates follow the predictive values but the predictive values also follow the recognition rates – both CUA- and CPA-parity may appear to be equally suitable measures of error risk parity. Observing unequal recognition rates, say between women and men, when it comes to commercial lending, one may very well wonder whether it can be fair that more women than men are misjudged, and seemingly for the only reason that they are women?

We must not assume, though, that unequal recognition rates are necessarily due to an unfair scoring procedure. They may also reflect unequal base rates of the target characteristic in the groups of men and women. Unequal recognition rates only indicate a lack of relational fairness if they correlate with unequal predictive values. Indeed, if the predictive values in two groups are the same but the base-rates of the target characteristic differ, unequal recognition rates are a necessary condition of a statistically sound and procedurally fair scoring procedure.

In a fair predictive model with equal predictive values across all groups, no group is treated better or worse simply because its members share an arbitrary trait. Disparate treatment only occurs when the feature in question is actually predictive, i.e., it correlates positively with the target characteristic. For example, a fair credit scoring system that assigns equal predictive values to both genders would

8 Also note the possibility of a *reductio ad absurdum*. If equal base-rates for a target feature could be assumed across demographic groups, predictive scoring would become meaningless.

rate women lower than men only if the actual loan repayment rates among women were lower than among men.

The reason why CPA-parity cannot be a free-standing measure of relational fairness is that CPA-parity may result from an incidental concurrence of unequal base rates and unequal predictive values of a model in two groups. When predictive values differ between groups, we cannot tell whether recognition rate differences are due to unequal base rates or biased predictive performance. In contrast, predictive values can reveal whether a model is tracking the base rate fairly or introducing bias. We must therefore always consult predictive values to make that determination. A CPA-measure of comparative fairness would thus seem redundant since it provides an unbiased estimate of individual error risks only on the condition that CUA-parity is already guaranteed.

We thus conclude that CPA-parity is not a necessary condition and free-standing criterion of procedural computational fairness. Nevertheless, unequal recognition rates in different demographic groups raise questions of social justice which must be addressed. The combined recognition rates of a model describe the distributive effects it has on the composition of demographic groups. A scoring model, for instance, that is meant to screen groups of applicants for executive positions will identify less women than men if the combined recognition rates in both groups differ. High numbers of misclassifications for women will lead to fewer women in the group of executives. This gives rise to issues of distributive justice irrespective of whether the difference in the recognition rates is due to a model's unequal predictive value in the respective groups or not.

5 Distributive Justice

There are two main issues of distributive justice with regard to potentially discriminatory procedures. Both raise questions of proportionality. The first issue relates to the proportionality of individual burdens: disadvantages imposed on persons on the basis of a scoring result may be unreasonable when compared with what is gained by using a particular scoring model. In the field of commercial lending, for instance, one may wonder whether the use of a narrow list of attributes related to the credit history of an individual for the assessment of personal creditworthiness outweighs the disadvantages imposed on young people who receive low scores due to their lack of credit history (Blanken and Klinger 2023).

The second issue is not about individual hardship but about collective outcomes and, more specifically, about the cumulative distributive effects of scoring procedures on the composition of demographic groups. A computational model may operate in a procedurally fair and non-discriminatory way and still lead to

structural inequalities that, like the disproportional share of men in senior management positions, are objectionable from a distributive point of view.

Disproportional individual burdens, though unjust, are not a matter of procedural fairness if they do not result from a denial of equal treatment. After all, the consequences of a negative evaluation – e.g., no job offer or a denial of credit – are the same for all *True* and *False Positives*. We must look, however, beyond the immediate outcomes of a procedure and also consider less tangible future disadvantages – say a decline in expected income or lost opportunities. These disadvantages create *marginal personal burdens* that are not the same for all persons. For instance, a low credit rating due to frequent moves may be far more damaging for someone with low income and education than for a well-paid professional. The severity of burdens varies in accordance with a person's socioeconomic position and even a procedurally fair scoring will typically have uneven consequences in different demographic groups. For this reason, the proportionality of the average burden imposed on people by a selective practice is not an adequate measure of distributive justice. The greater marginal burdens on the most disadvantaged persons must also be reasonable. To determine a threshold of maximal marginal burdens, however, is beyond the reach of philosophical argument. Given the narrow limits of utilitarian cost-benefit calculations, the issue of proportionality has to be settled by informed, but still intuitive, common sense, and, ultimately, by a fair political process.

Demands of fair and just equality address asymmetries both between individuals and between demographic groups. Group-based asymmetries amount to unjust structural inequalities when they persist over time and concern goods such as income or social positions, the distribution of which has a profound impact on people's lives. Not every structural inequality is a moral problem. Few would be troubled, for instance, if more green-eyed than blue-eyed people would go for holidays in the mountains; and we certainly do not seek a 'fair share' of quack doctors in our hospitals. Matters look different, however, when we learn that senior management positions are disproportionately occupied by men rather than women, or that women are overrepresented in the lowest income groups.

The challenge, again, is to distinguish unjust from morally unobjectionable structural inequalities. Moreover, not all structural injustice stems from unfair or discriminatory practices. The criteria of fairness in individual treatment and justice in structural outcomes are logically independent. We may be concerned, for example, about the disproportionately low share of women in senior executive positions even if hiring and promotion procedures were free of bias. Conversely, it may be unfair that Peter occupies a higher position in the corporate hierarchy than the better-qualified Mary, even if women were not underrepresented in leading positions.

Much unjust structural inequality in the distribution of goods and positions results from discriminatory practices and would not exist if all selective procedures conformed to the requirements of procedural fairness and equal treatment. Yet not all structural inequalities that raise moral questions can be reduced to procedural unfairness and, therefore, eliminated through legal reforms designed to secure equal treatment. Even fair procedures and unbiased selective decisions can give rise to structural inequalities that violate reasonable principles of distributive justice. Consider, for example, the gender pay gap in contrast with the broader phenomenon of income inequality in market economies.

The primary income distribution produced by a competitive market is the unintended cumulative outcome of countless individual transactions, not the foreseeable result of any single, generally followed discriminatory practice. Competitive markets generate inequalities of income and wealth even in the absence of discrimination and under the counterfactual assumption that all transactions are fair. Still, much primary income inequality in market economies may be expected to be unjust – not because of procedural flaws, but on grounds of distributive justice. One reason is that market incomes at the bottom of the distribution may fail to secure the necessities of life. Moreover, even among groups above the poverty line, existing wage differentials often bear little relation to actual contributions to economic productivity and may therefore appear unjustified.

By contrast, much of the existing gender pay gap is the foreseeable result of well-documented discriminatory practices in the labor market. The gap would be noticeably smaller if gender bias were absent and hiring decisions and wage negotiations consistently adhered to principles of procedural fairness and equal treatment (see Weichselbaumer and Winter-Ebmer 2005; Castilla 2015; Blau and Kahn 2017). Admittedly, it is difficult to determine precisely how much of the existing economic inequality is attributable to bias and discrimination, and how much would persist even under conditions of perfect procedural fairness. The distinction may therefore appear largely theoretical. Yet it remains analytically and politically important. It highlights that considerations of social justice apply to structural inequalities that cannot be reduced to questions of procedural fairness, and it reminds us that eliminating unjust structural inequalities requires not only reforming unfair practices to secure equal treatment but also adopting redistributive and other measures that directly address unjust distributions.

One final remark on group fairness. Some accounts of unjust structural inequality in the literature on non-discrimination employ notions of group fairness and group discrimination. The claim is that procedures which may appear non-discriminatory at the individual level can nonetheless be unfairly discriminating at the group level if they generate asymmetric distributions of important benefits and burdens across demographic groups. This line of argument, however, rests on a

misconception. What non-discrimination requires as a necessary condition of procedural fairness is the equal treatment of individuals who are subject to a selective procedure, not the equal treatment of the demographic groups to which they happen to belong. Individuals, not groups, are recruited, promoted, or granted credit, even if the criteria by which they are evaluated correlate with group membership. Moreover, to lodge a justice complaint about structural inequality on the basis of an additional requirement of group fairness – when individual-level fairness has already been secured – would commit us to the puzzling view that a group of people might be treated unfairly even though none of its members has been treated unfairly.

Nevertheless, there are good reasons to be concerned about structural inequalities that do not stem from considerations of equal treatment in social transactions or from an elusive conception of group fairness. Moral misgivings about structural inequalities in the composition of groups that do not derive from unfair procedures call for a further explanation that goes beyond considerations of transactional equal treatment and involves substantive principles of distributive justice.

One important reason to care about the composition of certain demographic groups – such as those in senior management or political leadership – derives from ideas of fair political representation and fair opportunities for collective action. The members of these groups regularly make decisions that profoundly affect the lives of others. Since we care about these decisions, we must also care about the composition of the groups making them, knowing that their perspectives and judgments are shaped by particular interests that tend to align with demographic categories. From a public point of view that accords equal weight to the interests of all members of society, persons have not only a claim to equal treatment irrespective of their group membership. They also have a claim that the groups to which they belong enjoy a fair opportunity to advance the interests of their members which suggests a principle of proportional representation as a requirement of distributive justice.

6 Conclusions

To conclude, the proposed framework of computational fairness rests on a sharp conceptual distinction between procedural requirements of equal treatment for persons subject to selective procedures and requirements of distributive justice. It holds that, in order to satisfy the equal-treatment requirement of procedural fairness, a computational model must meet a single formal condition: parity of predictive values across demographic groups. In addition, the framework articulates two non-procedural requirements of distributive justice applicable to

selective procedures more generally: first, that the burdens imposed on individuals be proportionate; and second, that these burdens be compatible with a just overall distribution of benefits and burdens in society.[9]

References

AGG. 2006. *Allgemeines Gleichbehandlungsgesetz* from 2006.

Arrow, Kenneth. 1972. "Models of Job Discrimination." In *Racial Discrimination in Economic Life*, edited by A. H. Pascal, 83—102. Lexington: Lexington Books.

Barocas, Solon, Hardt Moritz, and Arvind Narayanan. 2023. *Fairness in Machine Learning. Limitations and Opportunities*. Cambridge: MIT Press.

Blanken, Duygu Damar, and Helena Klinger. 2023. *Alterdiskriminierung bei der Kreditvergabe. Abschlussbericht*. Institut für Finanzdienstleistungen e.V im Auftrag der Antidiskriminierungsstelle des Bundes. www.iff-hamburg.de/wp-content/uploads/2023/09/ADS-Abschlussbericht (accessed April 16, 2026).

Blau, Francine D., and Lawrence M. Kahn. 2017. "The Gender Wage Gap: Extent, Trends, and Explanations." *Journal of Economic Literature* 55 (3): 789—865.

Carnap, Rudolf. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Castilla, Emilio J. 2015. "Accounting for the Gap: A Firm Study Manipulating Organizational Accountability and Transparency." *Organization Science* 26 (2): 311—33.

Chouldechova, Alexandra. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data* 5 (3): 153—63.

Cornacchia, G., V. W. Anelli, F. Narducci, A. Ragone, and E. Di Sciascio. 2023. "Auditing Fairness Under Unawareness Through Counterfactual Reasoning." *Information Processing & Management* 60. https://www.sciencedirect.com/science/article/pii/S0306457322003259 (accessed April 16, 2026).

Eubanks, Virginia. 2018. *Automating Inequality. How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Fabris, A., A. Esuli, A. Moreo, and F. Sebastiani. 2023. "Measuring Fairness Under Unawareness of Sensitive Attributes: A Quantification-Based Approach." *Journal of Artificial Intelligence Research* 76: 1117—80.

Goodman, Laurie, Jun Zhu, and Bing Bai. 2016. *Women are Better than Men at Paying Their Mortgages*. Washington: Urban Institute.

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *arXiv*.

Hempel, Carl. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hinsch, Wilfried. 2023. "Differences that Make a Difference. Computational Profiling and Fairness to Individuals." In *The Cambridge Handbook of Responsible Artifcial Intelligence*, edited by S. Vöneky, P. Kellmeyer, O. Müller, and W. Burkard, 229—51. Cambridge: Cambridge University Press.

9 My understanding of the subject has benefited greatly from discussions on algorithmic discrimination with the late Gert Wagner, Felix Rebitschek, and Helena Burkard. I am also grateful to Jörn Lamla for his comments on an earlier draft, and to Anton Leist and Julian Culp for their insightful feedback on the final manuscript.

Höltgen, Benedikt, and Nuria Oliver. 2025. “Reconsidering Fairness Through Unawareness from the Perspective of Model Multiplicity.” *arXiv*.
ICCPR. 1966/76. International Covenant on Civil and Political Rights (adopted 16 Dec. 1966, entered into force 23 Mar. 1976).
Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” *arXiv*.
Li, Geng. 2018. *Gender Related Differences in Credit Use and Credit Scores*. Washington: FEDS Notes.
O’Neill, Cathy. 2016. *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy*. New York: Penguin Books.
Phelps, Edmond. 1972. “The Statistical Theory of Racism and Sexism.” *American Economic Review* 62 (4): 659—61.
Rawls, John. 1993. *Political Liberalism*. New York: Columbia University Press.
Rothstein, H. R., and M. A. McDaniel. 1992. “Differential Validity by Sex in Employment Settings.” *Journal of Business and Psychology* 7 (1): 45—62.
Ruf, Boris, and Marcin Detyniecki. 2020. “Active Fairness Instead of Unawareness.” *arXiv*.
Weichselbaumer, Doris, and Rudolf Winter-Ebmer. 2005. “A Meta-Analysis of the International Gender Wage Gap.” *Journal of Economic Surveys* 19 (3): 479—511.