UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Fairness and Artificial Intelligence

Naudts, L.; Vedder, A.

DOI
10.1017/9781009367783.006
Publication date
2025
Document Version
Final published version
Published in
The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence

License
CC BY

Link to publication

Citation for published version (APA):
Naudts, L., & Vedder, A. (2025). Fairness and Artificial Intelligence. In N. A. Smuha (Ed.), *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence* (pp. 79-100). Cambridge University Press. https://doi.org/10.1017/9781009367783.006

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:06 Oct 2026

4

Fairness and Artificial Intelligence

Laurens Naudts and Anton Vedder

4.1 INTRODUCTION

Within the increasing corpus of ethics codes regarding the responsible use of AI, the notion of fairness is often heralded as one of the leading principles. Although omnipresent within the AI governance debate, fairness remains an elusive concept. Often left unspecified and undefined, it is typically grouped together with the notion of justice. Following a mapping of AI policy documents commissioned by the Council of Europe, researchers found that the notions of justice and fairness show "the least variation, hence the highest degree of cross-geographical and cross-cultural stability."[1] Yet, once we attempt to interpret these notions concretely, we soon find that they are perhaps best referred to as essentially contested concepts: over the years, they have sparked constant debate among scholars and policymakers regarding their appropriate usage and position.[2] Even when some shared understanding concerning their meaning can be found on an abstract level, people may still disagree on their actual relation and realization. For instance, fairness and justice are often interpreted as demanding some type of equality. Yet equality, too, has been the subject of extensive discussions.

In this chapter, we aim to clear up some of the uncertainties surrounding these three concepts. Our goal, however, is not to put forward an exhaustive overview of the literature, nor to promote a decisive view of what these concepts should entail. Instead, we want to increase scholars' sensibilities as to the role these concepts can perform in the debate on AI and the (normative) considerations that come with that role. Taking one particular interpretation of fairness as our point of departure (fairness as nonarbitrariness), we first investigate the distinction and relationship

[1] In addition to the notion of privacy. Isaac Ben-Israel et al., "Towards regulation of AI systems: Global perspectives on the development of a legal framework on artificial intelligence systems based on the Council of Europe's Standards on Human Rights, Democracy and the Rule of Law" (Council of Europe, 2020) 2020/16 50, https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.

[2] W. B. Gallie, "IX.—Essentially contested concepts" (1956) *Proceedings of the Aristotelian Society*, 56: 167.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

between procedural and substantive conceptions of fairness (Section 4.2). We build upon this distinction to further analyze the relationship between fairness, justice, and equality (Section 4.3). We start with an exploration of Rawls' conception of justice as fairness, a theoretical framework that is both procedural and substantively egalitarian in nature. This analysis forms a stepping stone for the discussion of two distinct approaches toward justice and fairness. In particular, Rawls' outcome-oriented or distributive approach is critiqued from a relational perspective. In parallel, throughout both sections, we pay attention to the challenges these conceptions may face in light of technological innovations. In the final step, we consider the limitations of techno-solutionism and attempts to formalize fairness by design in particular (Section 4.4), before concluding (Section 4.5).

4.2 CONCEPTIONS OF FAIRNESS: PROCEDURAL AND SUBSTANTIVE

In our digital society, public and private actors increasingly rely on AI systems for the purposes of knowledge creation and application. In this function, data-driven technologies guide, streamline, and/or automate a host of decision-making processes. Given their ubiquity, these systems actively co-mediate people's living environment. Unsurprisingly then, it is expected for these systems to operate in correspondence to people's sense of social justice, which we understand here as their views on how a society should be structured, including the treatment, as well as the social and economic affordances citizens are owed.

Regarding the rules and normative concepts used to reflect upon the ideal structuring of society, a distinction can generally be made between procedural notions or rules and substantive ones. Though this distinction may be confusing and is equally subject to debate, substantive notions and rules directly refer to a particular political or normative goal or outcome a judgment or decision should effectuate.[3] Conversely, procedural concepts and rules describe *how* judgments and decisions in society should be made rather than prescribing *what* those judgments and decisions should ultimately be. Procedural notions thus appear seemingly normatively empty: they simply call for certain procedural constraints in making a policy, judgment, or decision, such as the consistency or the impartial application of a rule. In the following sections, we elaborate on the position fairness typically holds in these discussions. First, we discuss fairness understood as a purely procedural constraint (Section 4.2.1), and second, how perceptions of fairness are often informed by a particular substantive, normative outlook (Section 4.2.2). Finally, we illustrate how procedural constraints that are often claimed to be neutral nonetheless tend to reflect a specific normative position as well (Section 4.2.3).

[3] See, for instance: Christine M. Korsgaard, "Self-constitution in the ethics of Plato and Kant" in Christine M. Korsgaard (ed.), *The Constitution of Agency: Essays on Practical Reason and Moral Psychology* (Oxford University Press, 2008), 106–107, https://doi.org/10.1093/acprof:oso/9780199552733.003.0004, accessed February 15, 2023.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

4.2.1 *Fairness as a Procedural Constraint*

Fairness can be viewed as a property or set of properties of processes, that is, particular standards that a decision-making procedure or structure should meet.[4] Suppose a government and company want to explore the virtues of automation. A government wants to streamline the distribution of welfare benefits and a company seeks the same with its hiring process. Understood as a procedural value, fairness should teach us something about the conditions under which (a) the initial welfare or hiring policy was decided upon and (b) how that policy will be translated and applied to individuals by means of an automated procedure. A common approach to fairness in this regard is to view it as a demand for nonarbitrariness: a procedure is unfair when it arbitrarily favors or advantages one person or group or situation over others, or arbitrarily favors the claims of some over those of others.[5] In their analysis of AI-driven decision-making procedures, Creel and Hellmann evaluate three different, yet overlapping, understandings that could be given to the notion of arbitrariness, which we will also use here as a springboard for our discussion.[6]

First, one could argue that a decision is arbitrary when it is unpredictable. Under this view, AI-driven procedures would be fair only when their outcome is reasonably foreseeable and predictable for decision subjects. Yet, even in the case a hiring or welfare algorithm would be rendered explicable and reasonably foreseeable, would we still call it fair when its reasoning process placed underrepresented and marginalized communities at a disproportionate disadvantage?

Second, the arbitrariness of a process may lie in the fact that it was "unconstrained by ex-ante rules."[7] An automated system should not have the capacity to set aside the predefined rules it was designed to operate under. Likewise, government case workers or HR personnel acting as a human in the loop should not use their discretionary power to discard automated decisions to favor unemployed family members. Instead, they should maintain impartiality. Once a given ruleset has been put in place, it creates the legitimate expectation among individuals that those rules will be consistently applied. Without consistency, the system would also become unpredictable. Yet, when seen in isolation, most AI-driven applications operate on some

[4] T. M. Scanlon, "Rights, Goals, and Fairness" 85.

[5] See, for example: Scanlon (n 4); Jonathan Wolff, "Fairness, respect, and the egalitarian ethos" (1998) *Philosophy & Public Affairs*, 27: 97; Christopher McMahon, *Reasonableness and Fairness: A Historical Theory* (1st ed., Cambridge University Press, 2016), www.cambridge.org/core/product/identifier/9781316819340/type/book, accessed January 31, 2023.

[6] Creel and Hellmann do not necessarily position these three understandings as the sole interpretations that could be given to the notion of arbitrariness. Kathleen Creel and Deborah Hellman, "The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems" (2022) *Canadian Journal of Philosophy*, 52: 26, 34, 37–38. For their analysis of these definitions, and their limitations in light of AI-driven decision-making, reference can be made to the aforementioned work.

[7] Creel and Hellman (n 6).

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

predefined ruleset or instructions.[8] Even in the case of neural networks, unless some form of randomization is involved, there is some method to their madness. In fact, one of AI's boons is its ability to streamline the application of decision-making procedures uniformly and consistently. However, the same observation would apply: would we consider decisions fair when they are applied in a consistent, rule-bound, and reproducible manner, even when they place certain people or groups at a disproportionate social or economic disadvantage?

Finally, one could argue that arbitrariness is synonymous with irrationality.[9] Fairness as rationality partly corresponds to the principle of formal equal treatment found within the law.[10] Fairness as rationality mandates decision-makers to provide a rational and reasonable justification or motivation for the decisions they make. Historically, the principle of equal treatment was applied as a similar procedural and institutional benchmark toward good governance: whenever a policy, decision, or action creates a distinction between a (group of) people or situations, that differentiation had to be reasonably justified. Without such justification, a differentiating measure was seen as being in violation of the procedural postulate that "like situations should be treated alike."[11] This precept could be read as the instruction to apply rules consistently and predictably. However, where a differentiating measure is concerned, the like-cases axiom is often used to question not only the application of a rule but also that rule's content: did the decision-maker consider the differences between individuals, groups, or situations that were relevant or pertinent?[12] Yet, this conception might be too easily satisfied by AI-driven decisions. Indeed, is it often not the entire purpose of AI-driven analytics to find relevant points of distinction that can guide a decision? As observed by Wachter: "Since data science mainly focuses on correlation and not causation […] it can seemingly make any data point or attribute appear relevant."[13] However, those correlations can generate significant exclusionary harm: they can make the difference between a person's eligibility or disqualification for a welfare benefit or job position. Moreover, due to the scale and uniformity at which AI can be rolled out, such decisions do not affect single individuals but large groups of people. Perhaps then, we should also be guided by the

[8] Ibid., 28–29.

[9] Ibid., 28.

[10] H. L. A. Hart, *The Concept of Law* (Clarendon Press, 1961); Stefan Sottiaux, "Het Gelijkheidsbeginsel: Langs Oude Paden En Nieuwe Wegen (Artikel, 2008) [WorldCat.Org]" (2008) *Rechtskundig Weekblad*, 72: 690.

[11] See among others: Stefan Sottiaux, "Het Gelijkheidsbeginsel : Langs Oude Paden En Nieuwe Wegen (Artikel, 2008) [WorldCat.Org]" (2008) *Rechtskundig Weekblad*, 72: 690. Christopher McCrudden and Haris Kountouros, "Human Rights and European Equality Law," in *Equality Law in an Enlarged European Union: Understanding the Article 13 Directives*, ed. Helen Meenan (Cambridge University Press, 2007), 73–116, https://doi.org/10.1017/CBO9780511493898.004.

[12] Creel and Hellman (n 6).

[13] Sandra Wachter, "The theory of artificial immutability: protecting algorithmic groups under anti-discrimination law" (2022) *SSRN Electronic Journal*, 20, www.ssrn.com/abstract=4099100, accessed May 27, 2022.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

disadvantage a system will likely produce and not only by whether the differences relied upon to guide a procedure appear rational or nonarbitrary.[14]

Through our analysis of the notion of nonarbitrariness, a series of standards have been identified that could affect the fairness of a given decision-making procedure. In particular, fairness can refer to the need to motivate or justify a particular policy, rule, or decision, and to ensure the predictable and consistent application of a rule, that is, without partiality and favoritism. In principle, those standards can also be imposed on the rules governing the decision-making process itself. For example, when a law is designed or agreed upon, it should be informed by a plurality of voices rather than be the expression of a dominant majority. In other words, it should not arbitrarily exclude certain groups from having their say regarding a particular policy, judgment, or decision. Likewise, it was shown how those standards could also be rephrased as being an expression of the procedural axiom that "like cases ought to be treated alike." Given this definition, we might also understand why fairness is linked to other institutional safeguards, such as transparency, participation, and contestability. These procedural mechanisms enable citizens to gauge whether or not a given procedure was followed in a correct and consistent fashion and whether the justification provided took stock of those elements of the case deemed pertinent.

4.2.2 *Toward a Substantive Understanding of Fairness*

As the above analysis hints, certain standards imposed by a purely procedural understanding of fairness could be easily met where AI is relied upon to justify, guide, and apply decision-making rules. As any decision-making procedure can be seemingly justified on the basis of AI analytics, should we then deem every decision fair?

In the AI governance debate, the notion of fairness is seldom used purely procedurally. The presence of procedural safeguards, like a motivation, is typically considered a necessary but often an insufficient condition for fairness. When we criticize a decision and its underlying procedure, we usually look beyond its procedural components. People's fairness judgments might draw from their views on social justice: they consider the context in which a decision is made, the goals it aims to materialize and the (likely) disadvantage it may cause for those involved. In this context, Hart has argued that justice and fairness seemingly comprise two parts: "a uniform or constant feature, summarized in the precept 'Treat like cases alike' and a shifting or varying criterion used in determining when, for any given purpose,

[14] Of course, differences will play a role in our evaluation of decision-making procedures. We need to assess whether characteristics were reasonable or sensible in light of the task at hand. The point made, however, is that it might not be the only thing that should take up our attention. Wachter, for instance, argues that AI-guided decisions and procedures should be based on empirically coherent information that has a proven connection or an intuitive link to the procedure at hand Wachter (n 13). See also: Sandra Wachter and Brent Mittelstadt, "A right to reasonable inferences: Re-thinking data protection law in the age of big data and AI" (2019) *Columbia Business Law Review*, 2019: 494.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

cases are alike or different."[15] This varying criterion entails a particular political or moral outlook, a standard we use to evaluate whether a specific policy or rule contributes to the desired structuring of society.

For example, we could invoke a substantive notion of equality that a procedure should maintain or achieve. We might say that AI-driven procedures should not bar oppressed social groups from meaningfully engaging with their social environment or exercising meaningful control and agency over the conditions that govern their lives.[16] In so doing, we could also consider the exclusionary harm algorithms might introduce. Hiring and welfare programs, for instance, affect what Creel and Hellman refer to as "realms of opportunities:" the outcomes of these decisions give people more choices and access to alternative life paths.[17] In deciding upon eligibility criteria for a welfare benefit or job opportunity, we should then carefully consider whether the chosen parameters risk reflecting or perpetuating histories of disadvantage. From a data protection perspective, fairness might represent decision-makers' obligation to collect and process all data they use transparently.[18] Needless to say, articulating one's normative outlook is one thing. Translating those views into the making, structuring, and application of a rule is another. While a normative perspective might support us in the initial design of a decision-making procedure, the latter's ability to realize a set of predefined goals will often only show in practice. In that regard, the normative standard relied upon, and its procedural implementation should remain subject to corrections and criticisms.[19]

Of course, purely procedural constraints could maintain their value regardless of one's particular moral outlook: whether a society is structured alongside utilitarian or egalitarian principles, in both cases, consistency and predictability of a rule's application benefit and respects people's legitimate expectations. Given this intrinsic value, we might not want to discard the application of an established procedure outright as soon as the outcomes they produce conflict with our normative goals and ambitions.[20] The point, however, is that once a substantive or normative position has been taken, it can be used to scrutinize existing procedures where they fail to meet the desired normative outcome. Or, positively put, procedural constraints

[15] See also: Peter Westen, "The empty idea of equality" (1982) *Harvard Law Review*, 95: 537; Hart (n 10) 159–160.

[16] Iris Marion Young, *Justice and the Politics of Difference* (Princeton University Press, 1990); Naudts, *Fair or Unfair Differentiation? Reconsidering the Concept of Equality for the Regulation of Algorithmically Guided Decision-Making* (Doctoral Dissertation). (2023).

[17] Creel and Hellman (n 6) 22.

[18] Damian Clifford and Jef Ausloos, "Data protection and the role of fairness" (2018) *Yearbook of European Law*, 37: 130.

[19] See also: Westen (n 15); Hart (n 10) 159–160.

[20] On this point, see also: Christine M. Korsgaard, "Self-Constitution in the Ethics of Plato and Kant" in Christine M. Korsgaard (ed), *The Constitution of Agency: Essays on Practical Reason and Moral Psychology* (Oxford University Press, 2008) 106–108, https://doi.org/10.1093/acprof:oso/9780199552733.003.0004, accessed 15 February 2023.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

can now be modeled to better enable the realization of the specific substantive goals we wanted to realize. For example, we may argue that the more an AI application threatens to interfere with people's life choices, the more institutional safeguards we need to facilitate our review and evaluation of the techniques and procedures AI decision-makers employ and the normative values they have incorporated into their systems.[21] The relationship between procedural and substantive fairness mechanisms is, therefore, a reciprocal one.

4.2.3 *The Myth of Impartiality*

Earlier we said that procedural fairness notions appear normatively empty. For example, the belief that a given rule should not arbitrarily favor one group over others might be seen as a call toward impartiality. If a decision-making process must be impartial to be fair, does this not exclude the decision-making process of being informed by a substantive, and hence, partial normative outlook? Even though the opposite may sometimes be claimed, efforts to remain impartial are not as neutral as they appear at first sight.[22] For one, suppose an algorithmic system automates the imposition of traffic fines for speeding. Following a simple rule of logic, any person driving over the speed limit allocated to a given area must be handed the same fine. The system is impartial in the sense that without exception it will consistently apply the rules as they were written regardless of those who were at the wheel. It will not act more favorably toward politicians speeding than ordinary citizens for instance. At the same time, impartiality thus understood excludes the system from taking into account contextual factors that could favor leniency, as might be the case when a person violates the speed limit as they are rushing to the hospital to visit a sick relative. Second, in decision-making contexts made in relation to the distribution of publicly prized goods, such as job and welfare allocation, certain traits, such as a person's gender or ethnicity, are often identified as arbitrary. Consequently, any disadvantageous treatment on the basis of those characteristics is judged to be unfair. The designation of these characteristics as arbitrary, however, is not neutral either: it represents a so-called color-blind approach toward policy and decision-making. Such an approach might intuitively appear as a useful strategy in the pursuit of socially egalitarian goals, and it can be. For instance, in a hiring context, there is typically no reason to assume that a person's social background, ethnicity, or gender will affect their ability to perform a given job. At the same time, this color-blind mode of thinking can be critiqued for its tendency to favor merit-based criteria as the most appropriate differentiating metric instead. Under this view, criteria reflecting merit are (wrongfully) believed

[21] Creel and Hellman (n 6). See also: Jeremy Waldron, *One Another's Equals: The Basis of Human Equality* (Belknap Press: Harvard University Press, 2017).

[22] See also: Takis Tridimas, *The General Principles of EU Law* (2nd ed., Oxford University Press, 2006), p. 62.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

to be most objective and least biased.[23] In automating a hiring decision, designers may need to define what a "good employee" is, and they will look for technical definitions and classifications that further specify who such an employee may be. As observed by Young, such specifications are not scientifically objective, nor neutrally determined, but instead "they concern whether the person evaluated supports and internalizes specific values, follows implicit or explicit social rules of behavior, supports social purposes, or exhibits specific traits or character, behavior, or temperament that the [decision-maker] finds desirable."[24] Moreover, a person's social context and culture have a considerable influence on the way they discover, experience, and develop their talents, motivations, and preferences.[25] Where a person has had fewer opportunities to attain or develop a talent or skill due to their specific social condition, their chance of success is more limited than those who could.[26] A mechanical interpretation of fairness as impartiality obscures the differences that exist between people and their relationship with social context and group affinities: individual identities are discarded and rendered abstract in favor of "impartial" or "universal" criteria. The blind approach risks decontextualizing the disadvantage certain groups face due to their possession of, or association with, a given characteristic. Though neutral at first glance, the criteria chosen might therefore ultimately favor the dominant majority disadvantaging those minorities a color-blind approach was supposed to protect in the first place. At the same time, it also underestimates how certain characteristics are often a valuable component of one's identity.[27] Rather than render differences between people, such as their gender or ethnicity, invisible, differences could instead be accommodated and harnessed to eliminate the (social and distributive) disadvantage attached to them.[28] For example, a person's gender or ethnicity may become a relevant and nonarbitrary criterion if we want to redress the historical disadvantage faced by certain social groups by imposing positive or affirmative action measures on AI developers.

[23] Young (n 16) 201. See also: Michael J. Sandel, *The Tyranny of Merit: What's Become of the Common Good?* (Penguin Books, 2021).

[24] Young (n 16) 204.

[25] In this sense, Rawls observes, the principle of fair opportunity can only be imperfectly carried out: "the extent to which natural capacities develop and reach fruition is affected by all kinds of social conditions and class attitudes." John Rawls, *A Theory of Justice* (Harvard University Press (Belknap Press, 1971), p. 74.

[26] Richard J. Arneson, "Against Rawlsian equality of opportunity" (1999) *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 93: 77.

[27] See also: Sandra Fredman, "Substantive equality revisited" (2016) *International Journal of Constitutional Law*, 14: 712; Sandra Fredman, "Providing equality: Substantive equality and the positive duty to provide" (2005) *South African Journal on Human Rights*, 21: 163.

[28] This is also a criticism that can be leveled against "fairness-by-unawareness" design metrics. These metrics construct fairness as achieved when certain characteristics are not explicitly used in a prediction process. Fredman, "Substantive equality revisited" (n 27) 720. See also: Naudts (n 16).

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

4.3 JUSTICE, FAIRNESS, AND EQUALITY

In the previous section, we illustrated how a procedural understanding of fairness is often combined with a more substantive political or normative outlook. This outlook we might find in political philosophy, and theories of social justice in particular. In developing a theory of social justice, one investigates the relationship between the structure of society and the interests of its citizens.[29] The interplay and alignment between the legal, economic, and civil aspects of social life determine the social position as well as the burdens and benefits that the members of a given society will carry. A position will be taken as to how society can be structured so it best accommodates the interests of its citizens. Of course, different structures will affect people in different ways, and scholars have long theorized as to what structure would suit society the best. Egalitarian theories for instance denote the idea that people should enjoy (substantive) equality of some sort.[30] This may include the recognition of individuals as social equals in the relationships they maintain, or their ability to enjoy equal opportunities in their access to certain benefits. In order to explain the intricate relationship that exists between the notions of justice, fairness, and equality as a normative and political outlook, John Rawls is a good place to start.

4.3.1 *Justice as Fairness*

In his book *A Theory of Justice*, Rawls defines justice as fairness.[31] For Rawls, the subject of justice is the basic structure of society. These institutions are the political constitution and the principal economic and social arrangements. They determine people's life prospects: their duties and rights, the burdens, and benefits they carry. In our digital society, AI applications are technological artifacts that co-mediate the basic structure of society: they affect the options we are presented (e.g., recommender systems), the relationships we enter into (e.g., AI-driven social media), and/or the opportunities we have access to (e.g., hiring and welfare algorithms).[32] While AI-driven applications must adhere to the demands of justice, the concept of fairness is, however, fundamental to arrive at a proper conception of justice.[33] More specifically, Rawls argues that the principles of justice can only arise out of an agreement made under fair conditions: "A practice will strike the parties as fair if none feels

[29] Philip Pettit, *Judging Justice. An Introduction to Contemporary Political Philosophy* (Routledge & Kegan Paul, 1980).

[30] See also Richard Arneson, "Egalitarianism" in Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2013, Metaphysics Research Lab, Stanford University, 2013), https://plato.stanford.edu/archives/sum2013/entries/egalitarianism/, accessed February 8, 2023.

[31] Rawls, *A Theory of Justice* (n 25).

[32] Iason Gabriel, "Towards a theory of justice for artificial intelligence" (2022) *Daedalus*, 151: 12.

[33] John Rawls, "Justice as fairness" (1958) *The Philosophical Review*, 67: 164, 178.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

that, by participating in it, they or any of the others are taken advantage of, or forced to give in to claims which they do not regard as legitimate."[34] It is this position of initial equality, where free and rational persons choose what course of action best suits the structure of society, from which principles of justice may arise.[35] Put differently, fairness does not directly inform the regulation, design, and development of AI, the principles of justice do so, but these principles are chosen from a fair bargaining position. While fairness could thus be perceived as a procedural decision-making constraint, the principles that follow from this position are substantive. And as the principles of justice are substantive in nature, Rawls argues, justice as fairness is not procedurally neutral either.

One major concern Rawls had was the deep inequalities that arise between people due to the different social positions they are born into, the differences in their natural talents and abilities, and the differences in the luck they have over the course of their life.[36] The basic structure of society favors certain starting positions over others, and the principles of justice should correct as much as possible for the inequalities people may incur as a result thereof. Rawls' intuitive understanding regarding the emergence of entrenched social inequality, which AI applications tend to reinforce, could therefore function as a solid basis for AI governance.[37]

In his *A Theory of Justice*, Rawls proposes (among others) the difference principle, which stipulates that once a society has been able to realize basic equal liberties to all and fair equality of opportunity in social and economic areas of life, social and economic inequalities can only be justified when they are to the benefit of those least advantaged within society. As AI applications not only take over social inequality but also have a tendency to reinforce and perpetuate the historical disadvantage faced by marginalized or otherwise oppressed communities, the difference principle could encourage regulators and decision-makers, when a comparison is made between alternative regulatory and design options, to choose for those policy or design options that are most likely to benefit the least advantaged within society. In this context, one could contend that justice should not only mitigate

[34] Ibid. Fairness is guaranteed as a result of the fair conditions under which people are able to reach an agreement regarding the principles of justice. They are the outcome of an original agreement in a suitably defined initial situation. The participants of this initial situation – or the original position – decide upon the principles that will govern their association. While the participants are rational and in the pursuit of their own interests, they are also each other's equals. They view themselves and each other as a source of legitimate claims. In addition, the parties that partake in this hypothetical original position are situated behind a veil of ignorance. No participant knows their place in society, their natural talents. They do not know the details of their life. From this position, they are to derive the appropriate principles of justice. Rawls, *A Theory of Justice* (n 25) chapter 3, The Original Position, and p. 119.

[35] Rawls, *A Theory of Justice* (n 25) 11.

[36] Rawls, *A Theory of Justice* (n 25).

[37] See also: Gabriel (n 32) 10; Jamie Grace, "'AI theory of justice': Using Rawlsian approaches to better legislate on machine learning in government" (2020) *SSRN Electronic Journal*, www.ssrn.com/abstract=3588256, accessed August 10, 2022.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

and avoid the replication of social and economic injustice but also pursue more ambitious transformative goals.[38] AI should be positively harnessed to break down institutional barriers that bar those least advantaged from participating in social and economic life.[39]

4.3.2 *Distributive Accounts of Fairness*

Like conceptions of fairness, people's understanding of what justice is, and requires, is subject to dispute. Rawls' understanding of justice for instance is distributive in nature. His principles of justice govern the distribution of the so-called primary goods: basic rights and liberties; freedom of movement and free choice of occupation against a background of diverse opportunities; powers and prerogatives of offices and positions of authorities and responsibility; income and wealth; and the social bases of self-respect.[40] These primary goods are what "free and equal persons need as citizens."[41] A distributive approach toward fairness may also be found in the work of Hart, who considered fairness to be a notion relevant (among others) to the way classes of people are treated when some burden or disadvantage must be *distributed* among them. In this regard, unfairness is a property not only of a procedure but also of the shares produced by that procedure.[42] Characteristic of the distributive paradigm is that it formulates questions of justice as questions of distribution. In general terms, purely distributive-oriented theories argue that any advantage and disadvantage within society can be explained in terms of people's possession of, or access to, certain material (e.g., wealth and income) or nonmaterial goods (e.g., opportunities and social positions).[43] Likewise, social and economic inequalities can be evaluated in light of the theory's proposed or desired distribution of those goods it has identified as "justice-relevant."[44] Inequality between people can be

[38] Gabriel (n 32) 9–10. See also: Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor* (Macmillan Publishers, 2018); Caroline Criado Perez, *Invisible Women: Exposing Data Bias in a World Designed for Men* (1st ed., Chatto & Windus, 2019); Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York University Press, 2018), www.degruyter.com/document/doi/10.18574/9781479833641/html, accessed December 8, 2021.

[39] See also: Gabriel (n 32) 9–10.

[40] John Rawls, *Justice as Fairness: A Restatement* (Kelly Erin ed., Belknap Press: Harvard University Press, 2001), pp. 58–59.

[41] John Rawls, "The basic liberties and their priority" in Sterling M. Mcmurrin (ed.) (1981), p. 89; Rawls, *Justice as Fairness: A Restatement* (n 40) 60.

[42] An additional area of social life where fairness is mandated is the situation where a person has been done some injury and must be compensated. Hart (n 10) 159.

[43] Young (n 16) 8.

[44] Thomas Pogge, "Relational conceptions of justice: Responsibilities for health outcomes" in Sudhir Anand, Fabienne Peter, and Amartya Sen (eds), *Public Health, Ethics, and Equity* (Oxford University Press, 2004), p. 147; Christian Schemmel, "Distributive and relational equality": (2011) *Politics, Philosophy & Economics* 127, http://journals.sagepub.com/doi/10.1177/1470594X11416774, accessed August 4, 2020.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

justified as long as it contributes to the desired state of affairs. If it does not, however, mechanisms of redistribution must be introduced to accommodate unjustified disadvantages.[45]

Distributive notions of fairness have an intuitive appeal as AI-driven decisions are often deployed in areas that can constrain people in their access to publicly prized goods, such as education, credit, or welfare benefits.[46] Hence, when fairness is to become integrated into technological applications, the tendency may be for design solutions to focus on the distributive shares algorithms produce and, conversely, to correct AI applications when they fail to provide the desired outcome.[47]

4.3.3 *Relational Accounts of Fairness*

Though issues of distribution are important, relational scholars have critiqued the dominance of the distributive paradigm as the normative lens through which questions of injustice are framed.[48] They believe additional emphasis must be placed on the relationships people hold and how people ought to treat one another as part of the relationships they maintain with others, such as their peers, institutions, and corporations. Distributive views on fairness might be concerned with transforming social structures, institutions, and relations, but their reason for doing so lies in the outcomes these changes would produce.[49] Moreover, as Young explains, certain phenomena such as rights, opportunities, and power are better explained as a

[45] Thomas W. Pogge, "Three problems with Contractarian-Consequentialist ways of assessing social institutions" (1995) *Social Philosophy and Policy*, 12: 241. Young (n 16) 24–25.

[46] Reuben Binns, "Fairness in machine learning: Lessons from political philosophy" (2018) *Proceedings of Machine Learning Research*.

[47] Atoosa Kasirzadeh, "Algorithmic fairness and structural injustice: Insights from feminist political philosophy" (2022) *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, http://arxiv.org/abs/2206.00945, accessed February 3, 2023; Pratik Gajane and Mykola Pechenizkiy, "On formalizing fairness in prediction with machine learning" (2017) arXiv:1710.03184 [cs, stat], http://arxiv.org/abs/1710.03184, accessed July 23, 2018; presented during the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (Stockholm, 2018).

[48] Young (n 16); Carina Fourie, Fabian Schuppert, and Ivo Walliman-Helmer (eds), *Social Equality: On What It Means to Be Equals* (Oxford University Press, 2015). For a relational perspective on AI, see: Abeba Birhane, "Algorithmic injustice: A relational ethics approach" (2021) *Patterns*, 2: 100205; Salomé Viljoen, "A relational theory of data governance" (2021) *The Yale Law Journal*, 82; Kasirzadeh (n 47); Virginia Dignum, "Responsible Artificial Intelligence: Recommendations and Lessons Learned," in *Responsible AI in Africa: Challenges and Opportunities*, ed. Damian Okaibedi Eke, Kutoma Wakunuma, and Simisola Akintoye (Cham: Springer International Publishing, 2023), 195–214, https://doi.org/10.1007/978-3-031-08215-3_9; Virginia Dignum, "Relational artificial intelligence" (2022) arXiv:2202.07446 [cs], http://arxiv.org/abs/2202.07446, accessed February 17, 2022; Naudts (n 16); Laurens Naudts, "The Digital Faces of Oppression and Domination: A Relational and Egalitarian Perspective on the Data-driven Society and its Regulation." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA (2024), 701–12. https://doi.org/10.1145/3630106.3658934.

[49] Schemmel (n 44); Pogge (n 44).

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

function of social processes, rather than thing-like items that are subject to distribution.[50] Likewise, inequality cannot solely be explained or evaluated in terms of people's access to certain goods. Instead, inequality arises and exists, and hence is formed, within the various relationships people maintain. For example, people cannot participate as social equals and have an equal say in political decision-making processes when prejudicial world views negatively stereotype them. They might have "equal political liberties" on paper, but not in practice.

When fairness not only mandates "impartial treatment" in relation to distributive ideals but also requires a specific type of relational treatment, the concept's normative reach goes even further.[51] AI applications are inherently relational. On the one hand, decision-makers hold a position of power over decision-subjects, and hence, relational fairness could constrain the type of actions and behaviors AI developers may impose onto decision-subjects. At the same time, data-driven applications, when applied onto people, divide the population into broad, but nonetheless consequential categories based upon generalized statements concerning similarities people allegedly share.[52] Relational approaches toward fairness will specify the conditions under which people should be treated as part of and within AI procedures.

Take for instance the relational injustice of cultural imperialism. According to Young, cultural imperialism involves the social practice in which a (dominant) group's experience and culture is universalized and established as the norm.[53] A group or actor is able to universalize their world views when they have access to the most important "means of interpretation and communication."[54] The process of cultural imperialism stereotypes and marks out the perspectives and lived experiences of those who do not belong to the universal or dominant group as an "Other."[55] Because AI-applications constitute a modern means of interpretation and communication in our digital society, they in turn afford power to those who hold control

[50] Young (n 16) 25–31.

[51] See, for example: Schemmel (n 44); John Baker, "Conceptions and dimensions of social equality" in Carina Fourie, Fabian Schuppert, and Ivo Walliman-Helmer (eds), *Social Equality: On What It Means to Be Equals* (Oxford University Press, 2015); Marie Garrau and Cécile Laborde, "Relational equality, non-domination, and vulnerability" in Carina Fourie, Fabian Schuppert, and Ivo Walliman-Helmer (eds), *Social Equality: On What It Means to Be Equals* (Oxford University Press, 2015).

[52] See also: Viljoen (n 48).

[53] Young (n 16) 59. See also: María C. Lugones and Elizabeth V. Spelman, "Have we got a theory for you! Feminist theory, cultural imperialism and the demand for 'the woman's voice'" (1983) *Women's Studies International Forum*, 6: 573; For a more in-depth application of this notion onto AI, as well as Young's other "faces of oppression," see also: Laurens Naudts, The Digital Faces of Oppression and Domination: A Relational and Egalitarian Perspective on the Data-driven Society and its Regulation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA (2024), 701–12. https://doi.org/10.1145/3630106.3658934.

[54] Nancy Fraser, "Talking about needs: Interpretive contests as political conflicts in welfare-state societies" (1989) *Ethics*, 99: 291; Nancy Fraser, "Toward a discourse ethic of solidarity" (1985) *PRAXIS International*, 5: 425.

[55] Young (n 16) 59. See also: WEB Du Bois, *The Souls of Black Folk* (Oxford University Press, 2007).

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

over AI: AI-driven technologies can discover and/or apply (new) knowledge and give those with access to them the opportunity to interpret and structure society. They give those in power the capacity to shape the world in accordance with their perspective, experiences, and meanings and to encode and naturalize a specific ordering of the world.[56] For example, in the field of computer vision methods are sought to understand the visual world via recognition systems. In order to do so AI must be trained on the basis of vast amounts of images or other pictorial material. To be of any use; however, these images must be classified as to what they contain. Though certain classification acts seemingly appear devoid of risk (e.g., does a picture contain a motorbike), others are not.[57] Computer vision systems that look to define and classify socially constructed categories, such as gender, race, and sexuality, tend to wrongfully present these categories as universal and detectable, often to the detriment of those not captured by the universal rule.[58] Facial recognition systems and body scanners at airports that have been built based on the gender binary risk treating trans-, non-binary, and gender nonconforming persons as nonrecognizable human beings.[59] In a similar vein, algorithmic systems may incorporate stereotyped beliefs concerning a given group. This was the case in the Netherlands, where certain risk scoring algorithms used during the evaluation of childcare benefit applications operated on the prejudicial assumption that ethnic minorities and people living in poverty were more

[56] The notion classification is used here in a broad sense. It not only refers to the ways in which an algorithmic decision-making process may group together individuals. It also refers to instances where data are classified or labelled in a training set. Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (Yale University Press, 2021) 128 and 139, https://doi.org/10.12987/9780300252392; Kate Crawford and Trevor Paglen, "Excavating AI: The politics of images in machine learning training sets" (*Excavating AI*, September 19, 2019), www.excavating.ai, accessed February 7, 2020.

[57] On the role of power in image data sets, see also the work by Milagros Miceli et al.: Milagros Miceli et al., "Documenting computer vision datasets: An invitation to reflexive data practices," *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2021), https://dl.acm.org/doi/10.1145/3442188.3445880, accessed March 10, 2021; Milagros Miceli, Julian Posada, and Tianling Yang, "Studying up machine learning data: Why talk about bias when we mean power?" (2021) arXiv:2109.08131 [cs], http://arxiv.org/abs/2109.08131, accessed October 4, 2021; Milagros Miceli, "AI's symbolic power: Classification in the age of automation" (2019); Milagros Miceli and Julian Posada, "A question of power: How task instructions shape training data" (2020) *Symposium on Biases in Human Computation and Crowdsourcing* (BHCC2020, November 11, 2020), https://sites.google.com/sheffield.ac.uk/bhcc-2020/program?authuser=0; Milagros Miceli, Martin Schuessler, and Tianling Yang, "Between subjectivity and imposition: Power dynamics in data annotation for computer vision" (2020) *Proceedings of the ACM on Human-Computer Interaction*, 4: 1.

[58] Crawford (n 56) 144. See also: Miceli et al. (n 57); Miceli and Posada, "A question of power: How task instructions shape training data" (n 57); Milagros Miceli and Julian Posada, "Wisdom for the crowd: Discoursive power in annotation instructions for computer vision" (arXiv, May 23, 2021), http://arxiv.org/abs/2105.10990, accessed August 29, 2022.

[59] Lucas Waldron and Medina Brenda, "TSA's body scanners are gender binary. Humans are not." (*ProPublica*), www.propublica.org/article/tsa-transgender-travelers-scanners-invasive-searches-often-wait-on-the-other-side?token=7bjY-MRzWk5Ed4DCZRvFVYwt2HBrAFXd, accessed February 7, 2022. See also: Os Keyes, "The misgendering machines: Trans/HCI implications of automatic gender recognition" (2018) *Proceedings of the ACM on Human-Computer Interaction*, 2: 1.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

likely to commit fraud.[60] The same holds true for highly subjective target variables, such as the specification of the "ideal employee" in hiring algorithms. As aforementioned, technical specifications may gain an aura of objectivity once they become incorporated within a decision-making chain and larger social ecosystem.[61]

Under a relational view, these acts, and regardless of the outcomes they may produce, are unjust because they impose representational harms onto people: they generalize, misrepresent, and deindividualize persons. From a relational perspective, these decisions may be unjustified because they interfere with people's capacity to learn, develop, exercise, and express skills, capacities, and experiences in socially meaningful and recognized ways (self-development) and their capacity to exercise control over, and participate in determining, their own options, choices, and the conditions of their actions (capacity to self-determination).[62] They do so however, not by depriving a particular good to people, but by rendering the experiences and voices of certain (groups of) people invisible and unheard. Unlike outcome-focused definitions of justice, whose violation may appear as more immediate and apparent, these representational or relational harms are less observable due to the opacity and complexity of AI.[63]

If we also focus on the way AI-developers treat people as part of AI procedures, a relational understanding of fairness will give additional guidance as to the way these applications can be structured. For instance, procedural safeguards could be implemented to facilitate people's ability to exercise self-control and self-development when they are likely to be affected by AI. This may be achieved by promoting diversity and inclusion within the development, deployment, and monitoring of decision-making systems as to ensure AI-developers are confronted by a plurality of views and the lived experiences of others, rather than socially dominant conventions.[64] Given the power they hold, AI-developers should carefully consider their normative assumptions.[65] Procedural safeguards may attempt to equalize power asymmetries within the digital environment and help those affected by AI to regain, or have increased, control over those structures that govern and shape their choices and options in socially meaningful and recognized ways. The relational lens

[60] "Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal." (Amnesty International, 2021), www.amnesty.nl/content/uploads/2021/10/20211014_FINAL_Xenophobic-Machines.pdf?x42580; "Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms" (*Amnesty International*, October 25, 2021), www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-benefit-scandal/, accessed August 15, 2022; Jan Kleinnijenhuis, "Hoe de Belastingdienst lage inkomens profileerde in de jacht op fraude" (*Trouw*, November 22, 2021), www.trouw.nl/gs-bbb66add, accessed November 23, 2021.

[61] Crawford (n 56) chapter 4, Classification.

[62] For an application of both notions onto AI, see also Naudts (n 16 and 48), drawing from the work of Young (n 16).

[63] Solon Barocas, Moritz Hardt, and Arvind Narayanan, "Fairness and machine learning" 253, chapter Introduction.

[64] See also Naudts (n 48).

[65] See, for instance: Miceli et al. (n 57).

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

may contribute to the democratization of modern means of interpretation and communication to realize the transformative potential of technologies.

4.4 LIMITATIONS OF TECHNO SOLUTIONISM

From a technical perspective, computer scientists have explored more formalized approaches toward fairness. These efforts attempt to abstract and embed a given fairness notion into the design of a computational procedure. The goal is to develop a "reasoning" and "learning" processes that will operate in such a way that the ultimate outcome of these systems corresponds to what was defined beforehand as fair.[66] While these approaches are laudable, it is also important to understand their limitations. Hence, they should not be seen as the only solution toward the realization of fairness in the AI-environment.

4.4.1 *Choosing Fairness*

During the development of AI systems, a choice must be made as to the fairness metric that will be incorporated. Since fairness is a concept subject to debate, there has been an influx of various fairness metrics.[67] Yet, as should be clear from previous sections, defining fairness is a value-laden and consequential exercise. And even though there is room for certain fairness conceptions to complement or enrich one another, others might conflict. In other words, trade-offs will need to be made in deciding what type of fairness will be integrated, if the technical and mathematical formalization thereof would already be possible in the first place.[68]

Wachter and others distinguish between bias preserving and bias transforming metrics and support the latter to achieve substantive equality, such as fair equality of opportunity and the ability to redress disadvantage faced by historically oppressed social groups.[69] Bias-preserving metrics tend to lock in historical bias present within society and cannot effectuate social change.[70] In related research, Abu-Elyounes

[66] Laurens Naudts, "Towards accountability: The articulation and formalization of fairness in machine learning" (2018) *SSRN Electronic Journal*, www.ssrn.com/abstract=3298847, accessed July 30, 2020.

[67] Gajane and Pechenizkiy (n 47); Doaa Abu Elyounes, "Contextual fairness: A legal and policy analysis of algorithmic fairness" (September 1, 2019), https://papers.ssrn.com/abstract=3478296, accessed February 5, 2023; Sandra Wachter, Brent Mittelstadt, and Chris Russell, "Bias preservation in machine learning: The legality of fairness metrics under EU non-discrimination law" (Social Science Research Network, 2021) *SSRN Scholarly Paper* 3792772, https://papers.ssrn.com/abstract=3792772, accessed April 28, 2022.

[68] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent trade-offs in the fair determination of risk scores" (2016) arXiv:1609.05807 [cs, stat], http://arxiv.org/abs/1609.05807, accessed October 11, 2020. See also Section 1.4.2 The Limitations of Abstraction.

[69] Wachter, Mittelstadt, and Russell (n 67).

[70] Ibid.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

suggested that different fairness metrics can be linked to different legal mechanisms.[71] Roughly speaking, she makes a distinction between individual fairness, group fairness, and causal reasoning fairness metrics. The first aim to achieve fairness toward the individual regardless of their group affiliation and is closely associated with the ideal of generating equal opportunity. Group fairness notions aim to achieve fairness to the group an individual belongs to, which is more likely to be considered as positive or affirmative action. Finally, due process may be realized through causal reasoning notions that emphasize the close relationship between attributes of relevance and outcomes.[72] This correspondence between fairness metrics and the law could affect system developers and policymakers' design choices.[73] For example, affirmative action measures can be politically divisive. The law might mandate decision-makers to implement positive action measures but limit their obligation to do so only for specific social groups and within areas such as employment or education because they are deemed critical for people's social and economic participation. Thus, the law might (indirectly) specify which fairness metrics are technologically fit for purpose in which policy domains.

Regardless of technical and legal constraints, formalized approaches may still be too narrowly construed in terms of their *inspiration*. For instance, Kasirzadeh has observed how "most mathematical metrics of algorithmic fairness are inherently rooted in a distributive conception of justice."[74] More specifically, "theories or principles of social justice are often translated into the distribution of material (such as employment opportunities) or computational (such as predictive performance) goods across the different social groups or individuals known to be affected by algorithmic outputs."[75] In other words, when outcome-based approaches are given too much revery, we may discard the relational aspects of AI-systems. In addition, and historically speaking, machine learnings efforts arose out of researchers' attempts to realize discrimination-aware data mining or machine learning.[76] In this regard, the notion of fairness has often been closely entwined with more substantive interpretations of equality and nondiscrimination law. This often results in the identification of certain "sensitive attributes" or "protected characteristics," such as gender

[71] Doaa Abu-Elyounes, "Contextual fairness: A legal and policy analysis of algorithmic fairness" (2020) *University of Illinois Journal of Law, Technology & Policy*, 2020: 1, 5.

[72] Abu Elyounes (n 67).

[73] Ibid. See also: Agathe Balayn and Seda Gurses, "Beyond debiasing: Regulating AI and its inequalities." (EDRi, 2021).

[74] Kasirzadeh (n 47) 4.

[75] Ibid.

[76] Binns (n 46). Bettina Berendt and Sören Preibusch, "Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence" (2014) *Artificial Intelligence and Law*, 22: 175; Bettina Berendt and Sören Preibusch, "Toward accountable discrimination-aware data mining: The importance of keeping the human in the loop—and under the looking glass" (2017) *Big Data*, 5: 135; Dino Pedreschi Salvatore Ruggieri and Franco Turini, "Discrimination-aware data mining," 9.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

or ethnicity. The underlying idea would be that fairness and equality are realized as soon as the outcome of a given AI-system does not disproportionately disadvantage individuals because of their membership of a socially salient group. For instance, one could design a hiring process so the success rate of an application procedure should be (roughly) the same between men and women when individuals share the same qualifications. Even though these approaches aspire to mitigate disadvantage experienced by underrepresented groups, they may do so following a (distributive), single-axis and difference-based nondiscrimination paradigm. This could be problematic for a two-fold reason. First, intersectional theorists have convincingly demonstrated the limitations of nondiscrimination laws' single-attribute focus.[77] Following an intersectional approach, discrimination must also be evaluated considering the complexity of people's identities, whereby particular attention must be paid to the struggles and lived experiences of those who carry multiple burdens. For instance, Buolamwini and Gebru demonstrated how the misclassification rate in commercial gender classification systems is the highest for darker-skinned females.[78] Second, the relational and distributive harms generated by AI-driven applications are not only faced by socially salient groups. For instance, suppose a credit scoring algorithm links an applicant's trustworthiness to a person's keystrokes during their online file application. Suppose our goal is to achieve fair equality of opportunity or equal social standing for all. Should we not scrutinize any interference therewith, and not only when the interference is based upon people's membership of socially salient groups?[79]

Yet, in our attempt to articulate and formalize fairness, Birhane and others rightfully point out that we should be wary of overly and uncontestedly relying on white, Western ontologies to the detriment and exclusion of marginalized philosophies and systems of ethics.[80] More specifically, attention should also be paid to streams of philosophy that are grounded "in down-to-earth problems and […] strive to challenge underlying oppressive social structures and uneven power dynamics," such as Black Feminism, Critical Theory, and Care Ethics and other non-Western and feminist philosophies.[81] Hence, questions regarding fairness and justice of AI

[77] Kimberle Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics" (1989) *University of Chicago Legal Forum*, 1989: 31; Kimberle Crenshaw, "Mapping the margins: Intersectionality, identity politics, and violence against women of color" (1991) *Stanford Law Review*, 43: 1241.

[78] Joy Buolamwini and Timnit Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," 15.

[79] That is not to say that our approach in tackling the harms faced by socially oppressed and non-oppressed groups should be identical. Indeed, in our attempt to protect the interests of both groups, we may need to distinguish in the protective measures we envisage to accommodate their respective needs and struggles. Naudts (n 16). See also: Wachter (n 13).

[80] Abeba Birhane et al., "The forgotten margins of AI ethics," *2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2022), https://dl.acm.org/doi/10.1145/3531146.3533157, accessed February 2, 2023.

[81] Ibid., 949–50.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

systems must be informed by the lived experiences of those they affect, rather than rendered into a purely abstract theoretical exercise of reflection or technological incorporation.

4.4.2 *Disadvantages of Abstraction*

If fairness is constructed toward the realization of a given outcome by design, they run the risk of oversimplifying the demands of fairness as found within theories of justice or the law. Fairness should not be turned into a simplified procedural notion the realization of which can be achieved solely via the technological procedures that underlie decision-making systems. While fairness can be used to specify the technical components underlying a decision-making process and their impact, it could also offer broader guidance regarding the procedural, substantive, and contextual questions that surround their deployment. Suppose a system must be rendered explicable. Though technology can help us in doing so, individual mechanisms of redress via personal interaction may enable people to better understand the concrete impact AI has had on their life. Moreover, when fairness is seen as a technical notion that governs the functioning of one individual or isolated AI-system only, the evaluation of their functioning may become decontextualized from the social environment in which these systems are embedded and from which they draw, as well as their interconnection with other AI-applications.[82] Taking a relational perspective as a normative point of departure, the wider social structures in which these systems are developed, embedded, and deployed, become an essential component for their overall evaluation. For example, fairness metrics are often seen as a strategy to counter systemic bias within data sets.[83] Large datasets, such as CommonCrawl, used for training high-profile AI applications are built from information mined from the world wide web. Once incorporated into technology, subtle forms of racism and sexism, as well as more overt toxic and hateful opinions shared by people on bulletin boards and fora, risk being further normalized by these systems. As Birhane correctly notes: "Although datasets are often part of the problem, this commonly held belief relegates deeply rooted societal and historical injustices, nuanced power asymmetries, and structural inequalities to mere datasets. The implication is that if one can 'fix' a certain dataset, the deeper problems disappear."[84] Computational approaches might wrongfully assume complex (social) issues can be formulated in terms of problem/solution. Yet this, she believes, paints an overly simplistic picture of the matter at hand: "Not only are subjects of study that do not lend themselves

[82] See, for instance: Andrew D. Selbst et al., "Fairness and abstraction in sociotechnical systems," *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2019), https://doi.org/10.1145/3287560.3287598, accessed February 2, 2023.
[83] Balayn and Gurses (n 73).
[84] Birhane (n 48) 6.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

to this formulation discarded, but also, this tradition rests on a misconception that injustice, ethics, and bias are relatively static things that we can solve once and for all."[85] As AI systems operate under background conditions of structural injustice, efforts to render AI fairer are fruitless if not accompanied by genuine efforts to dismantle existing social and representational injustice.[86] Fairness thus requires us to view the bigger picture, where people's relationships and codependencies become part of the discussion. Such efforts should equally extend to the labor conditions that make the development and deployment of AI systems possible. For instance, in early January 2023, reports emerged how OpenAI, the company behind ChatGPT, outsourced the labeling of data as harmful to Kenyan data workers as part of their efforts to reduce users' exposure to toxic-generated content. For little money, data workers have to expose themselves to sexually graphic, violent, and hateful imagery under taxing labor conditions.[87] This begs the question: can we truly call a system fair once it has been rid of its internal biases knowing this was achieved through exploitative labor structures, which rather than the exception, appear to be standard practice?[88]

Finally, one should be careful as to which actors are given the discretionary authority to decide how fairness should be given shape alongside the AI value-chain. For example, the EU AI Act, which governs the use of (high-risk) AI systems, affords considerable power to the providers of those systems as well as (opaque) standardization bodies.[89] Without the public at large, including civil society and academia, having access to meaningful procedural mechanisms, such as the ability to contest, control, or exert influence over the normative assumptions and technical metrics that will be incorporated into AI-systems, the power to choose and define what is fair will be predominantly decided upon by industry actors. This discretion may, in the

[85] Ibid.

[86] See also: Annette Zimmermann and Chad Lee-Stronach, "Proceed with caution" (2021) *Canadian Journal of Philosophy*, 1.

[87] Billy Perrigo, "The $2 per hour workers who made ChatGPT safer" (2023) *Time*, January 18, https://time.com/6247678/openai-chatgpt-kenya-workers/, accessed July 5, 2023.

[88] Milagros Miceli and Julian Posada. 2022. The Data-Production Dispositif. Proc. ACM Hum.-Comput. Interact. 6, CSCW2, Article 460 (November 2022), 37 pages. https://doi.org/10.1145/3555561

[89] See among others, Article 16 (Obligations of Providers of High-Risk AI Systems), as well as Article 40 (Harmonised Standards and Standardisation Deliverables), read in conjunction with Section 2 (Requirements for High-Risk Systems) of the AI Act. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)Text with EEA relevance. See also: Nathalie A Smuha et al., "How the EU can achieve legally trustworthy AI: A response to the European Commission's Proposal for an Artificial Intelligence Act" (August 5, 2021) <https://papers.ssrn.com/abstract=3899991> accessed July 21, 2023; Johann Laux, Sandra Wachter, and Brent Mittelstadt, 'Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act', *Computer Law & Security Review* 53 (1 July 2024): 105957, https://doi.org/10.1016/j.clsr.2024.105957.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

words of Barocas, lead to situations "in which the work done by socially conscious computer scientists working in the service of traditional civil rights goals, which was really meant to be empowering, suddenly becomes something that potentially fits in quite nicely with the existing interests of companies."[90] In other words, it could give those in control of AI the ability to pursue economic interests under the veneer of fairness.[91] In this regard, Sax has argued how the regulation of AI, and the choices made therein, may not only draw inspiration from liberal and deliberative approaches to democracy, but could also consider a more agonistic perspective. While the former try to look for rational consensus among political and ideological conflict through rational and procedural means, agonism questions the ability to solve such conflicts: "from an agonistic perspective, pluralism should be respected and promoted not by designing procedures that help generate consensus, but by always and continuously accommodating spaces and means for the contestation of consensus(-like) positions, actors, and procedures."[92]

4.5 CONCLUSION

The notion of fairness is deep and complex. This chapter could only scratch the surface. This chapter demonstrated how a purely procedural conceptualization of fairness completely detached from the political and normative ideals a society wishes to achieve, is difficult to maintain. In this regard, the moral aspirations a society may have regarding the responsible design and development of AI-systems, and the values AI-developers should respect and incorporate, should be clearly articulated first. When we have succeeded in doing so, we can then start investigating how we could best translate those ideals into procedural principles, policies, and concrete rules that can facilitate the realization of those goals.[93] In this context, we argued that as part of this articulation process, we should not only be focused on

[90] Solon Barocas, "Machine learning is a co-opting machine" (*Public Books*, June 18, 2019), www.publicbooks.org/machine-learning-is-a-co-opting-machine/, accessed February 15, 2023.

[91] Ben Wagner, "Ethics as an escape from regulation. From 'ethics-washing' to ethics-shopping?" in Emre Bayamlioglu et al. (eds), *BEING PROFILED* (Amsterdam University Press, 2019), www.degruyter.com/view/books/9789048550180/9789048550180-016/9789048550180-016.xml, accessed August 26, 2020; Luciano Floridi, "Translating principles into practices of digital ethics: Five risks of being unethical" (2019) *Philosophy & Technology*, 32: 185; Elettra Bietti, "From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, 2020), https://doi.org/10.1145/3351095.3372860.

[92] Marijn Sax, "Algorithmic news diversity and democratic theory: Adding agonism to the mix" (2022) *Digital Journalism*, 10: 1650, 1651. In it, the author draws on the work of political theorist Chantal Mouffe.

[93] See also: Wibren van der Burg, "The morality of aspiration: A neglected dimension of law and morality" in Willem Witteveen J. and Wibren van der Burg (eds), *Rediscovering Fuller: Essays on Implicit Law and Institutional Design* (Amsterdam University Press, 2009).

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006

how AI-systems interfere with the distributive shares or outcomes people hold. In addition, we should also pay attention to the relational dynamics AI systems impose and their interference into social processes, structures, and relationships. Moreover, in so doing, we should be informed by the lived experiences of the people that those AI systems threaten to affect the most.

Seeking fairness is an exercise that cannot be performed within, or as part of, the design phase only. Technology may assist in mitigating the societal risks AI systems threaten to impose, but it is not a panacea thereto. The realization of fair AI requires a holistic response; one that incorporates the knowledge of various disciplines, including computer and social sciences, political philosophy, ethics, and the law, and where value-laden decisions are meaningfully informed and open to contestation by a plurality of voices and experiences.

Downloaded from https://www.cambridge.org/core. UVA Universiteit van Amsterdam, on 24 Jun 2025 at 15:24:59, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009367783.006