Science and Engineering Ethics (2021) 27:73
https://doi.org/10.1007/s11948-021-00348-z

ORIGINAL RESEARCH/SCHOLARSHIP

Fairness as Equal Concession: Critical Remarks on Fair AI

Ryan van Nood[1] · Christopher Yeomans[1]

Received: 27 April 2021 / Accepted: 21 October 2021 / Published online: 22 November 2021

Abstract

Although existing work draws attention to a range of obstacles in realizing fair AI, the field lacks an account that emphasizes how these worries hang together in a systematic way. Furthermore, a review of the fair AI and philosophical literature demonstrates the unsuitability of 'treat like cases alike' and other intuitive notions as conceptions of fairness. That review then generates three desiderata for a replacement conception of fairness valuable to AI research: (1) It must provide a meta-theory for understanding tradeoffs, entailing that it must be flexible enough to capture diverse species of objection to decisions. (2) It must not appeal to an impartial perspective (neutral data, objective data, or final arbiter.) (3) It must foreground the way in which judgments of fairness are sensitive to context, i.e., to historical and institutional states of affairs. We argue that a conception of *fairness as appropriate concession* in the historical iteration of institutional decisions meets these three desiderata. On the basis of this definition, we organize the insights of commentators into a process-structure map of the ethical territory that we hope will bring clarity to computer scientists and ethicists analyzing Fair AI while clearing some ground for further technical and philosophical work.

Keywords Fairness · AI · History · Medical AI · Ethics

Introduction

Although existing work draws attention to a range of obstacles in realizing fair AI, the field lacks an account that emphasizes how these worries hang together in a systematic way. We endeavor to achieve this by defusing concerns about the *philosophical* definition of fairness while acknowledging how every step of design is shot

✉ Christopher Yeomans
cyeomans@purdue.edu

Ryan van Nood
rvannood@purdue.edu

[1] Department of Philosophy, Purdue University, 100 N. University Street, West Lafayette, IN 47907, USA

through with ethical sensitivity. Here, we organize the insights of commentators into a process-structure map of the ethical territory and settle upon one philosophical definition of fairness that we think best captures the array of ethical worries and which might serve as a tool for approaching AI design. We thus hope this scheme will bring clarity to computers scientists and ethicists approaching Fair AI while clearing some ground for further technical and philosophical work.

The Standard View of Fairness: Treat like Cases Alike

The obvious questions to start with are, what does fairness mean? Does the literature deploy the term in converging or equivocal ways?[1] What bearing might the history of the philosophy of fairness and justice have?[2] Etymologically, the English use of 'fair' is rooted in contexts of what is pleasing or agreeable to the eye and also to denote what is suitable, fitting.[3] Onto this aesthetic concept was later grafted a moral connotation (as in speaking someone of fair and unblemished character) that is now archaic to our contemporary, unaesthetic sense of the word. However, we can hear the aesthetic in the ethical if we remember that to each belong notions of harmony and proportion and thus activating political concepts of just portions, of portioning one her due. When formulated an idea of harmonious portioning as an imperative, the notion takes the form: treat like cases alike.

Fair AI commentators generally agree with this intuitive formulation, expressing it, e.g., as giving individuals their due (Fazelpour & Lipton, 2020), as a matter of non-discrimination (Binns, 2018), as treating similar people similarly (Dwork 2012; Friedler et al., 2016; Corbett-Davies & Goel, 2018), or as a matter of the just distribution of the decisional power afforded to AI (Floridi & Cowls, 2019). Along the same lines, anti-discrimination employment law has influenced fair AI efforts by motivating the like treatment of protected and non-protected groups (Feldman et al., 2015).

In a recent literature review, Tsamados et. al. taxonomize the extant definitions in the literature as follows:

1. Anti-classification, which refers to protected categories, such as race and gender, and their proxies not being explicitly used in decision making;
2. Classification parity, which regards a model as being fair if common measures of predictive performance, including false positive and negative rates, are equal across protected groups;
3. Calibration, which considers fairness as a measure of how well-calibrated an algorithm is between protected groups;

[1] We aim here only to systematize worries about *fairness* in particular. For a recent higher-altitude survey of ethical issues surrounding AI, see Tsamados et al., (2021).

[2] It is also worth considering the history of the philosophy of fairness together with the history of societal-level algorithmic practice in general. See, e.g., Ochigame (2020).

[3] OED.

4. Statistical parity, which defines fairness as an equal average probability estimate over all members of protected groups (Tsamados et. al., 2021).

All four of these definitions are easily seen to be rooted in a basic conception of ‘treat like cases alike.’ Anti-classification expresses this definition negatively, by requiring that classifications that do *not* properly distinguish like cases be suppressed from the decision-making process. Men and women are to be treated as alike, for example, which requires avoiding the usage of gender markers in training and decision-making. Both classification and statistical parity expresses the definition more positively, by providing tests for ensuring that groups that ought to be treated as alike are in fact so treated. They essentially specify what like treatment amounts to. Calibration does the same at a slightly more abstract level. In fact, we think that it is better to consider these four as operationalizations of the ‘treat like cases alike’ definition of fairness rather than definitions in their own right. (The same is true of many other taxonomies in the literature.)

Related to the distinction between background and operationalized definitions of fairness is a general issue of technical conditions. Philosophical interest in fairness in the context of AI is, in part, a response to the systemic limitations of algorithmic tools, limitations which do not necessarily constrain everyday exercises of fairness and which therefore present novel challenges to ethical design. On a naïve view of the technical conditions in which humans design for fairness, these limitations may look like sources of strength rather than liability. After all, the ability to process high quantities of data at great speed marks a contrast with the limitation of human practical reasoning. However, as Tasamados et al. rightly summarize, higher quantities of data do not translate to higher quality data—they may simply reproduce conditions of systemic unfairness that produced the data itself. Furthermore, because AI outputs express probabilistic measures, they cannot be said to identify causal relationships, and patterns identified in any quantity of data “may be the result of inherent properties of the system modelled by the data” rather than inherent in the worldly conditions we take a given model to describe (Tsamados et al., 2021). Moreover, ML systems may exhibit adversarial vulnerability, that is, they are liable to mistakes following from a mismatch between data on which a system is trained and the diversity of those to which it is applied.[4]

[4] Such mistakes, moreover, are those which a human is unlikely to make, as when imperceptible or irrelevant changes in an image provoke an ML system to erroneously label an object (Goodfellow et al., 2014). Tsamados et al. summarize certain frontiers of progress in generating artificial adversarial examples in order to make training sets more robust (Tsamados et al., 2021).

Problems with the Standard View

Despite the intuitive appeal of an apparently commonsense notion of fairness, its grounding in legal precedent,[5] and (relative) consensus in the literature, a fair AI model remains elusive. A second stage in the fair AI literature has begun to diagnose the problem: available formal definitions of fairness prove unable to achieve various criteria of success without simultaneously sacrificing others (Corbett-Davies; Friedler, 2016; Harrison et al., 2020).

This is to be expected. After all, if we recall everyday ways in which we deploy the concept of fairness, we see that it makes overlapping claims upon a decision-maker which are sometimes in tension. Evaluating a decision as fair means finding that it is fair in all respects, not just with respect, say, to its outcome. We might acknowledge the evaluative complexity involved in a fairness appraisal when we remember the diversity of kinds of objections one might raise to a decision. For example, one might protest that a given context calls for an historically-disadvantaged group to receive a disparate, reparative allocation of goods rather than an amount formally equivalent to historically-privileged groups. One might object that two identical decision results are not equally fair if one of them turns upon an arbitrary rule, which is to say that we care that the reasons for a decision are ethically meaningful rather than coincidental, arbitrary, or based on the discrimination of others. Moreover, it is perfectly familiar to protest that one is not *seeing* justly or fairly, even where there is agreement about the rules from which decisions are inferred from that data. One can also reject entirely a decision-making institution on the grounds of complicity or perpetuation of injustice. Each of these everyday dimensions of fairness have their analogous territory in our four-part map (see Sect. 4). All these and more make up perfectly familiar dimensions of fairness that cannot be suppressed in order to make theorizing formal AI constructs more convenient. Our notion of fairness for ML must therefore also be flexible enough to field this diversity of protest.

This gloss on fairness serves as a reminder of how dynamic the actual world behaves in contrast to the algorithmic abstractions that involve some degree of idealizing or airbrushing of contextual complexity.[6] Political philosophy can help us

[5] It must be noted that challenges to the fixation on definitions invoking legal precedents in anti-discrimination law exist. Although anti-discrimination law may more or less neatly map onto quantitative measures of fairness (in whatever way they are contrived), that fixation may cover over other more robust demands for social justice, such as those that would target structural conditions (Hoffmann 2019). Fairness approaches that reduce to risk assessments based upon historical data may fatalistically encourage the carceral state in ways that attention to welfare provision might not (Ochigame 2020). For a general treatment of the relation between EU non-discrimination law and AI fairness, see (Wachter et. al., 2020).

[6] Fazelpour and Lipton (2020) invoke the ideal/non-ideal theory distinction from political philosophy to diagnose the temptation to artificially limit the actual scope of fairness. Whereas ideal theory imagines a perfect world and seeks to solve discrepancies between it and the actual world from that ideal standard, non-ideal theory orients itself from a description of the actual world and the manifold web of causes generating a given injustice, thus situating itself in a position to ameliorate an injustice while keeping track of diffuse burdens of responsibility (on account of that attention to material conditions). By limiting fairness definitions to parity outcomes, aspiringly fair AI systems instantiate localized expressions of naïve ideal theorizing, thereby passing off degenerate definitions of fairness as the complex and

bring into focus ethical aspects of states of affairs to which we might otherwise be blind.[7] Rather than fixating on regions of mathematical abstraction as the domain of fairness (such as class parity outcomes or classification process principles), fairness evaluations can only actually be made against the background of the institutional states of affairs that condition both an individual's prospects for flourishing and the variables about those individuals captured in a data set.[8]

In the place of ideal theorizing that is abstracted from the everyday, variegated territories of concern, we intend our map as a guide to attention for the rough ground of the actual, non-ideal world. Given these reminders about what we ordinarily mean by 'fairness' and the kinds of grounds from which objections might issue and to which fair AI must be responsive, we can ask: how might philosophy reply to the fair AI conversation in order to provide theoretical support for an action-guiding notion of fairness in AI?

It should be noted that 'fairness' is not one of the standard technical terms in philosophy such as 'knowledge' or 'justice.' Nonetheless, some philosophers have sought to explain the nature of justice in terms of fairness. In those discussions the conception of fairness as treating like cases alike receives some discussion. Here, we briefly review those discussions in the cases of Aristotle and John Rawls.

Aristotle The relevant discussion takes place in Book V of Aristotle's *Nicomachean Ethics*. The specific virtue of justice is held to be a kind of fairness understood as not "overreaching" with respect to social benefits (NE v.1).[9] Aristotle also puts this as being a mean between doing injustice (depriving others of such benefits) or suffering injustice (being deprived of benefits). Aristotle then tries to specify what such a mean would amount to in different situations, each generating a different species of fairness (NE v.3–5). In the distribution of common assets (e.g., honor or wealth), treating like cases alike amounts to distributing them in proportion to desert grounded in character (geometrical proportion). In the correction of injuries, fairness simply returns the parties to their original, equal standing (numerical proportion). Finally, in the exchange of goods, the norm of fairness is the equality of value, e.g., between apples and shoes (reciprocal proportion).

Footnote 6 (continued)

internally diverse everyday notion described above, as fair in general. Fazelpour and Lipton note industry AI products hastily certifying themselves as fair on account of controlling for demographic parity, placing some blame on fair AI literature making it possible: "In many papers, these fairness-inspired parity metrics are described as *definitions of fairness* and the resulting algorithms that satisfy the parities are claimed axiomatically to be *fair*" (Fazelpour and Lipton, 2020 9).

[7] Rueben Binns explores various conversations from the history of political philosophy to try on different lenses for capturing what would make certain states of affairs upon which AI systems might bear fair or not, such as how classifiers relate to an individual's responsibility, culpability, or desert for them (Binns, 2018). This is a valuable exercise in using the history of philosophy to see more clearly. Our project complements such efforts while actually settling upon a specific theoretical tool, namely, fairness as equal concession.

[8] We adapt this point from non-deal theorists such as Elizabeth Anderson and Chris McMahon (See: Anderson, 2013; McMahon, 2016).

[9] References to the *Nicomachean Ethics* are to book and chapter numbers. See Aristotle (1984).

As Aristotle's own discussion already shows, fairness understood in this way is tremendously sensitive to context, and if identified with object-level norms it must be cashed out into a multitude of species. It is far from obvious that the three notions of proportion Aristotle himself discusses are exhaustive. But there is a further difficulty in that each of these proportionalities requires some neutral standpoint to specify the relevant metric of exchange, e.g., quality of character for distribution or equality of value for exchange. In the modern period, few philosophers have been as confident as Aristotle in their ability to identity the relevant standpoint and who occupies it.

But perhaps more importantly, 'treating like cases alike' seems at best like a good summary of three different cases in which the reasoning involved is quite different. On Aristotle's account, each different kind of good is correlated in a different way with a different feature of the agents who are candidate recipients of the good in question; both 'like' and 'treat' seem to vary in meaning in each of the three cases. 'Treating like cases alike' does not seem to get to the normative heart of the matter or provide any independent standard of fairness. (When we come to the next section, we will try to make a virtue of this feature of fairness.) If we have the relevant criteria and are confident in our ability to judge, the invocation to treat like cases alike doesn't add anything to our decision-making process. It only functions as an independent principle when we don't have the relevant criteria, or are not able to judge with certainty, or are for some other reason unable to produce the morally prescribed state of affairs by our own actions. Even in such cases, it just amounts to the principle that unequal treatment must be justified (see Strauss, 2002). In this vein it is worth taking up Rawls view of justice as fairness, which ramifies into two principles, namely, one of equality, and one specifying a justification for departures from equality.

Rawls Rawls understood justice as fairness, where the meaning of fairness was cashed out in two principles (Rawls, 1971). The first principle mandates basic equality, and the second licenses deviation from equality under certain circumstances—namely equality of opportunity and when the deviation is to the benefit of everyone (particularly the worst off). Since Rawls' theory concerns what he calls the "basic structure" of society rather than any individual decision, it is not of direct application to issues in AI decision making. But two things can contribute to our discussion. First, a recognition that, like Aristotle, different goods are apportioned to people in different ways. *Political goods* are apportioned in strict equality by the first principle, and *economic* goods are apportioned according to benefit by the second principle. And furthermore, the apportionment of economic benefits has nothing to do with desert (only with the proper distribution of benefits), and so the deviation from equality can be justified as fair in a way that does not involve treating like cases alike. Second, a specific feature of the view long brought into relief by feminist and religious critics deserves mention. According to Rawls, the correct thought experiment for determining which goods ought to be apportioned in which ways is to retreat behind a veil of ignorance—i.e., to forget which actual position one holds in society—so as to put oneself in the position of anyone. But this seems to assume that underneath our, e.g., gender or religious identity, there is a set of basic wants or values. Not only is this in itself questionable, but even if this is the case, we may prioritize them in different ways and view them from different perspectives. If this is potentially the case with the thought experiment of a choice of the basic structure of society, it is *a fortiori* the case in the more specific and local choices to be made by AI reasoning.

We leave aside a consideration of whether this invalidates Rawls's view as an issue in political philosophy—for our purposes here most of that debate is not germane. What we do think is that the consideration of both Aristotle and Rawls shows is that any substantive conception of fairness will have to move beyond the invocation to 'treat like cases alike', *in order to do justice to the different perspectives involved in decisions that are potentially fair or unfair*.

The considerations advanced in this section generate the following desiderata for any conception of fairness valuable to AI research:

1. It must provide a meta-theory for understanding tradeoffs, entailing that it must be flexible enough to capture diverse species of objection.
2. It must not appeal to an impartial perspective (neutral data, objective data, or final arbiter.)
3. It must foreground the way in which judgments of fairness are sensitive to context, i.e., to historical and institutional states of affairs.

Fairness as Appropriate Concession

We get pretty close to a view serviceable as guidance for fairness in AI reasoning from a contemporary philosopher in the Rawlsian tradition, Christopher McMahon.[10] On this theory, fairness is understood as a norm of **reciprocal concern** in **cooperative arrangements**, and amounts to **appropriate concession**. That is, in every cooperative arrangement, participants would like to see the cooperation arranged differently, and in any arrangement of sufficient complexity and number of participants, some or all of the participants will have to concede their preferences in favor of those of others. These preferences and concessions are distributed in non-uniform ways, which pattern looks different from the non-uniform perspectives of the participants, and the best way to understand a fair arrangement is one in which the concessions specifically are appropriately distributed. Four important implications of this definition can be emphasized. First, appropriate concession is equal concession (29). The thought here is that we have all given up enough when the amount that each has given up seems about the same *from their own perspective* as the other sees him or herself as having given up *from their perspective* (163–4). Second, the value of fairness is promoted primarily negatively, by the elimination of perceived unfairness. As McMahon puts it, "fairness is promoted by eliminating disparities of concession—by eliminating unfairness. This means that in an important respect, unfairness is the central concept in this part of the morality of reciprocal concern. Fairness is the absence of

[10] McMahon (2016).

unfairness" (38–9). Third, the promotion of fairness is an iterable but also interminable process. On McMahon's view, because of both the plurality of perspectives involved and the arising of new circumstances, judgments of the fairness of an arrangement are always subject to revision and the need to make such judgments about new arrangements (152–3). (As we will see below, there is a further aspect to iteration not mentioned by McMahon that is central for AI decision making.) Finally, reasonable disagreement about fairness is always possible and thus ineliminable: "even if bias is eliminated, different people may, because of their different judgmental histories as cooperators, be able to competently reach conflicting conclusions about what fairness requires in a given case" (162–3).

The primary advantage of this definition over 'treat like cases alike' is that the notion of equal concession can give an account of the way that fairness in AI requires trade-offs (our first desideratum, above). Right now, these are sometimes understood as trade-offs inherent within different definitions (Harrison et al., 2020) or formulations (Friedler et al., 2016) of fairness. But McMahon's conception allows us to conceive of fairness as the meta-level balance of tradeoffs between these object-level values. For example, Floridi and Cowls identify the four traditional biomedical principles of beneficence, non-maleficence, justice and autonomy while adding explicability as a fifth AI-relevant principle (2019). A fair decision will operationalize the proper trade-off between those values in the circumstances. Furthermore, these trade-offs are *concessions* because these object-level values are generally held to be important to different degrees by the different individuals or groups, often because of uneven privileges of access to the good in question (see our examples in Sect. 4). Talking of trade-offs as concessions brings perspectival variation into view and foregrounds the essentially *social dimensions* of the trade-offs. (This goes some ways towards satisfying our second desiderata above, namely that a conception of fairness not make appeal to an impartial point of view.) On this account, fairness is the virtue of the judgment as to which trade-offs to make. That means judgments about who gets how much of what matters to them or in what form.

However, there are still two difficulties with McMahon's view. The first goes to the heart of the theory on its own terms, whereas the second primarily concerns its application to AI.

First, though McMahon's view is historical in one sense—he acknowledges that conception of fairness change over time—it is not historical in a more important sense—that judgments of fairness are *responsive to immediate past patterns of concession*. But the latter is the key to the former, because the judgments of unfairness that are the triggers for the learning experiences that change our evaluations of both institutions and conceptions of fairness are keyed in part to what we think people have conceded over time. In the example of AI decision-making in criminal justice contexts (e.g., the COMPAS program), part of what is generating the judgment of unfairness is the historical experience that African-Americans have conceded too much in past iterations of sentencing and probation decisions (Harrison et al., 2020 398). Rather than reversing this disparity of concession in the new iteration, the data sets used to train the algorithms seem to project it into the future as though it were a neutral baseline.

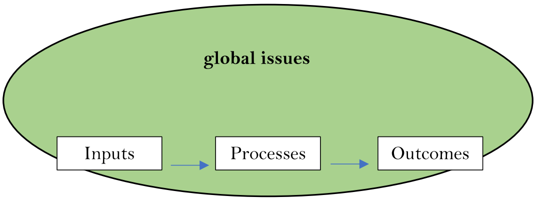
McMahon primarily sees the challenge that generates new conceptions of fairness to lie in the need to extrapolate from known contexts of cooperation to novel contexts of cooperation (157–8 & 168–9). Both generally and as a matter for AI contexts, it is essential that we recognize that fairness judgments are inherently tied to immediately past iterations of the institutional arrangements and the patterns of concession they embodied. This helps to satisfy the third desideratum, i.e., that the context sensitivity of fairness be recognized.

The second difficulty with McMahon's view is that it is still too closely tied to a Rawlsian conception of cooperative enterprises. Though we agree with Rawlsians who push back against the notion that the economy or the state or even the criminal justice system is a pure power relation (Max Weber was right that any stable power relation requires legitimation and is thus more than mere power), the setting of reciprocal concern in a cooperative enterprise raises problems that distract from the central issues of fairness.

We thus propose to modify McMahon's view as follows. First, fairness should be understood as historical in terms of different iterations of similar choices and the past concessions made by each party. And second, we will speak of institutions rather than cooperative enterprises. But within those contexts, we nonetheless think that fairness understood as appropriate concession can do a better job of orienting judgements of fairness with respect to AI than can the conception of fairness as treating like cases alike. Rather, we might say that the notion of appropriate concession helps us evaluate how alike cases actually are by projecting in the contexts of concern found on our map.

A Map of AI Fairness Issues

Floridi and Taddeo 2016 identify three axes of conern in data ethics: (1) the ethics of data, concerned with, e.g. privacy and its risks; (2) the ethics of algorithms, concerned primarily with ethical design and auditing of ML; and (3) the ethics of practices, which concerns professional ethics and policy pertaining to responsible innovation. These dimensions of concern are obviously intertwined, such that, for example, worries about data privacy or auditing of algorithmic processes and professional responsibility cannot be pulled apart (Floridi and Taddeo 2016 4). We organize the ethics of data and algorithms into a three-part process map set against a background category implicated at each step. We thus picture fair AI under this general scheme:: In addition to privacy issues, the level of data or **inputs** concerns the provenance of and value-judgments informing the construction of data sets; the category of **processes** take the ethical design and auditing of algorithms within its scope; **outcomes** concerns the formalized criteria of "fairness" such as parity; the background category of **global issues** involves questions such as the place of algorithmic solutions ought to have in society, broader worries of policy and philosophy arising from the societal-institutional background that conditions what is decided on behalf of those affected by AI decision-making.



The foundation of any would-be fair AI system is its **inputs**, which present the most recalcitrant and oft-noted obstacles of any region on our map. Although it is trivially true that the quality of one's raw materials conditions the quality of one's results, acknowledging this point rigorously through an algorithm's design no straightforward matter. What makes for these difficulties?

Contrary to the naïve view that data is an impartial reflection of an objective world (and hence free of bias), commentators stress that data are prone to reiterating structural unfairness on account of the conditions surrounding data collection, the societal privileges and blindspots reflected therein (Kitchin, 2014; Crigger & Khoury, 2019; Middlestadt and Floridi 2016). Data are partial and do not straightforwardly represent the fuller context in which they are embedded and on which fairness considerations so often hinge (Binns, 2018 79; Middelstadt and Floridi 2016 477). Data are always partial in the sense of being incomplete and in the sense of freighting value distinctions rather than an imaginary, value-neutral picture of things.

Understanding a distribution of concessions demands not only understanding the historical conditions that make a classifier meaningful but also the conditions of data collection inflecting that meaning in fairness-ramifying ways. For example, if we develop an ML tool for skin cancer diagnosis that is trained on a data set which includes only a marginal number of samples from non-white skin in the training data, it makes sense to understand the unfairness involved in terms of unequal concession. When it comes to being subject to any such diagnostic tool, every patient concedes that their individual case is imperfectly represented in an AI system trained to make generalizations gleaned from training data. In this case, however, non-white patients would be conceding much more of that fundamental reliability than white patients, which would be to compound disparate concessions given the disparate treatment, access, and trust in the American healthcare system experienced by, for example, African-American patients. We can see here how deploying a notion fairness as equal concession—as with understanding any ethical situation in terms of a given theoretical construct—requires ethical and social sensitivities beyond the bare definition in order to recognize what is being conceded and why. That is to say, the notion must be projected into the contexts of data in the relevant ways, work which no formula can do for us.

On account of their complexity and ability to generate new decision-making rules, the **processes** of ML systems create unique challenges for their ethical oversight. And the point just made regarding ethical sensitivities in the context of handling data goes also in this context. For if there is no impartial standpoint but rather the interested perspectives of those affected, then evaluating ML behavior can only be done after the fact by a human observer attending to that behavior's contextual ramifications. It is possible, for example, for ML systems to pick up on arbitrary patterns that harm the groups they were designed to protect (Fazelpour & Lipton, 2020; Middelstadt, 2016). Just as in everyday ethical life it is taken as given that the grounds for a decision made on one's behalf are available in order to understand, accept, or challenge them, so too is the availability of reasons behind life-altering judgements enshrined in law. The US Equal Credit Opportunity Act requires a statement of reasons for the denial of a loan and the EU's GDPR requires those responsible for certain decision-making systems to provide "meaningful information about the logic" driving their outputs (qtd. in Binns et al., 2018). On account of the unique difficulty of articulating that logic in the context of AI, together with the magnification of decision-making power by the societal minority tasked with designing and implementing it, ethics initiatives worldwide tend emphasize the necessity for these systems to be explicable and transparent in order to understand the systems themselves and the distributions of human responsibility for them (Floridi & Cowls, 2019).

Aside from articulating how processes might unblind protected classes or create new vulnerabilities in the ways ML rules are adopted, fairness as equal concession might help computer scientists bring home how transparency is conceded disparately between them and users. Designers concede some degree of autonomy to their ML creations, taking for granted a measure of opacity between a ML system's initial rules and those it might devise, but users whose lives it actually affects are likely to be several steps removed from both the actual processes it adopts and a capacity to evaluate them. Great care is thus necessary to ensure that this disparity of concession between the architects of great computing power and those who would endorse its role in their lives as reasonable resolves in the favor of those users whose concessions mark their vulnerability.

As noted above, initial discussions within the literature on fairness in AI focused on the scope of **outcomes** and satisfaction of criteria such as group parity (cf. Fazelpour & Lipton, 2020). Such artificially narrow definitions of fairness are only plausible on the naïve assumption that data actually reflects the individuals they are about in ways that are evenly distributed across protected classes (Friedler et al., 2016). In a systemically unjust world, furthermore, it is not obvious that formal group parity would in every case be fair, even given ideal data. Harrison et al. find that in the context of the much-discussed COMPAS scenario, surveyed participants rated outcomes as biased that were equalized at the expense of false positive rates that disadvantaged African-Americans (2020 398). This suggests intuitions about fairness are more complex than the scope of formal group parity would allow. In many cases, systemic injustice may demand policies of affirmative action in order to right historic wrongs (Anderson, 2013; Binns, 2018). Here, as in every region of analysis on our map, much wider backgrounds of relevance are pertinent to recognizing what will violate everyday intuitions about fairness. Fairness as equal concession can help us evaluate and support the intuition that equalizing outcomes that favor rather than disadvantage African-Americans because it attunes us to the history of unfair institutional decision-making that might be rectified. So while concessions between public safety and personal autonomy must always be carefully balanced, they take place against an historical institutional background in which those concessions occur unevenly, on account of, for example, unfair policing practices.

The domain of **global issues**, then, constitutes not only the background against which data might count as relevant, reliable, and in any sense fair and whether we would accept certain decision rules and outcomes as acceptable, but also whether ML even belongs in certain contexts. For example, some raise questions regarding whether influencing behavior through targeted advertising or through personalized pricing *could ever* count as fair treatment (Middelstadt, 2016 9–10). In the context of COMPAS, Green points out that since African-Americans are disproportionately subject to criminogenic circumstances, even perfectly accurate risk assessments in criminal sentencing could only disproportionately reproduce the outcome of incarceration, itself a criminogenic factor (Green, 2018). Others argue that risk assessments as such are only quasi-scientific predictors of outcomes, and that these data and AI systems would be better put to use as diagnostic tools for evaluating unfair societal conditions (Barabas et al., 2018). Fairness as equality of concession demands acknowledgement of the ways in which concessors might give voice to their situation in society by declaring the very deployment of a given AI tool to reflect and reiterate inequality.

These issues are matters of professional responsibility to which any computer scientist is answerable, but settling them requires sustained attention and deliberation to the kinds of sociological, historical, and philosophical considerations adduced in every region of our map. Global issues are thus implicated at every level of AI architecture. Justifying as fair the inclusion, exclusion, and treatment of a given datum, for example, plainly depends upon one's willingness to appreciate these aspects of the world itself. And it may be the case that machine judges will never be acceptable replacements for human ones when it comes to tools such as COMPAS, as some intuit (Harrison et al., 2020 397), just as the American Medical Association officially adopts the term 'Augmented Intelligence' (AuI) in order to emphasize AI's role as an enhancement of human intelligence in these sensitive contexts rather than its replacement (American Medical Association, 2018).

We might say that in certain medical and legal contexts, we intuitively concede to the imperfect partialities of a particular human being's judgement, partial though it only can be. In many of the above examples, *intuitive correctives to machine results are less accurately described as achievements of impartiality but rather as socio-historically sensitive instances of partiality, being partial to the right groups in the right ways and in the right contexts.* This necessary partiality is precisely why it is so important that there be a distribution of the decisional power afforded to AI (Floridi & Cowls, 2019). But the correctives that ethicists and computer scientists are able to instantiate in an AI system are positive outworkings of a more primitive fact, namely, that while an AI system can conceivably fair, it is not really conceivably *ever* impartial, but rather partial in ways better or worse when evaluating against the global issues manifest in every given datum, issues on which an AI system takes a stand in every discrimination it makes. So although human judgement is necessarily always partial, it is also ameliorable, projecting itself from its present understanding toward unattained but achievable states of understanding the fair relation between individual human lives and the institutional backgrounds in which they are embedded, on behalf of both of which that judgement grounds machine transactions. The endeavor for ever more detailed, truer, harmonious, and fair evaluative visions must be context-sensitive and dynamic—and it has no end point. The phrase 'equality of concession' does a lot of work in our account. Determining the sense of it is the task in each given case, and two cases might invite rather different senses. The usefulness of our account, we hope, might lie in that it furnishes a criterion for our honesty about whether the descriptions given of a fair-aspiring algorithmic system are as rigorously attuned to its objects and persons as required by those who conceding the most.

Authors' contributions Equal.

Funding Work on this project was supported by US National Science Foundation Grant 1939728 FAI: Identifying, Measuring, and Mitigating Fairness Issues in AI.

Availability of Data and Material All material is publicly available.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declared that they have no conflict of interest.

References

American Medical Association. (2018). *AMA passes first policy recommendations on augmented intelligence*. 2018. Accessed at www. ama-assn. org/ama-passes-first-policy-recommendations-augmented-intelligence.

Anderson, E. (2013). *The imperative of integration*. Princeton University Press.

Aristotle, J. B. (1984). *The complete works of Aristotle: The revised Oxford translation*. Princeton University Press.

Barabas, C., Virza, M., Dinakar, K., Ito, J., & Zittrain, J. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on fairness, accountability and transparency* (pp. 62–76). Association for Computing Machinery.

Binns, R. (2018). What can political philosophy teach us about algorithmic fairness? *IEEE Security & Privacy, 16*(3), 73–80.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023

Crigger, E., & Khoury, C. (2019). Making policy on augmented intelligence in health care. *AMA Journal of Ethics, 21*(2), 188–191.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214–226). Association for Computing Machinery.

Fazelpour, S., & Lipton, Z. C. (2020). Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 57–63). Association for Computing Machinery.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *ACM SIGKDD international conference on knowledge discovery and data mining*. Association for Computing Machinery.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572

Green, B. (2018). 'Fair' risk assessments: A precarious approach for criminal justice reform. In *5th Workshop on fairness, accountability, and transparency in machine learning* (FAT/ML 2018).

Harrison, G., Hanson, J., Jacinto, C., Ramirez, J., & Ur, B. (2020). An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 392–402). Association for Computing Machinery.

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

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data & Society, 1*(1), 2053951714528481.

McMahon, C. (2016). *Reasonableness and fairness: A historical theory*. Cambridge University Press.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 2053951716679679.

Mittelstadt, B. D., & Floridi, L. (2016). The ethics of big data: Current and foreseeable issues in biomedical contexts. *Science and Engineering Ethics, 22*(2), 303–341.

Ochigame, R. (2020). The long history of algorithmic fairness. In *Phenomenal World*. Retrieved December 11, 2020 from https://phenomenalworld.org/analysis/long-history-algorithmic-fairness

Rawls, J. (1971). *A theory of justice*. Belknap Press of Harvard University Press.

Strauss, D. A. (2002). Must Like Cases Be Treated Alike? U of Chicago, Public Law Research Paper No. 24. http://dx.doi.org/https://doi.org/10.2139/ssrn.312180

Tsamados, A., Aggarawal, N., Cowls, J., Morely, J., Roberts, H., Taddeo, M., & Floridi, L. (2021). The ethics of algorithms: Key problems and solutions. *AI & Society*. https://doi.org/10.1007/s00146-021-01154-8

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. arXiv preprint arXiv:2005.05906