Ethics and Information Technology (2021) 23:703–712
https://doi.org/10.1007/s10676-021-09608-9

ORIGINAL PAPER

Non-empirical problems in fair machine learning

Teresa Scantamburlo[1]

Accepted: 19 July 2021 / Published online: 5 August 2021

Abstract

The problem of fair machine learning has drawn much attention over the last few years and the bulk of offered solutions are, in principle, empirical. However, algorithmic fairness also raises important conceptual issues that would fail to be addressed if one relies entirely on empirical considerations. Herein, I will argue that the current debate has developed an empirical framework that has brought important contributions to the development of algorithmic decision-making, such as new techniques to discover and prevent discrimination, additional assessment criteria, and analyses of the interaction between fairness and predictive accuracy. However, the same framework has also suggested higher-order issues regarding the translation of fairness into metrics and quantifiable trade-offs. Although the (empirical) tools which have been developed so far are essential to address discrimination encoded in data and algorithms, their integration into society elicits key (conceptual) questions such as: What kind of assumptions and decisions underlies the empirical framework? How do the results of the empirical approach penetrate public debate? What kind of reflection and deliberation should stakeholders have over available fairness metrics? I will outline the empirical approach to fair machine learning, i.e. how the problem is framed and addressed, and suggest that there are important non-empirical issues that should be tackled. While this work will focus on the problem of algorithmic fairness, the lesson can extend to other conceptual problems in the analysis of algorithmic decision-making such as privacy and explainability.

Keywords Machine learning · Fairness · Empirical approach · Assessment of machine learning

Introduction

Since scoring and classification algorithms have been introduced to support, if not replace, human decisions in contexts as diverse as healthcare, insurance, employment and criminal justice, the problem of fairness has become a central theme in the field of machine learning.

In machine learning, the problem of fairness is addressed in the context of a policy prediction problem (Kleinberg et al., 2015): a decision about the future of a subject is made and the outcome should not be negatively affected by any sensitive attribute or feature that is considered as irrelevant for that decision. Often this situation can be modelled as a binary decision, where the subject is judged to be eligible for a specific role or a service with a certain degree of confidence.[1] Thus, once the target is fixed, e.g. "students graduating on time", "low risk insurance applicants", "high risk offenders", "vulnerable children", "successful job candidate", "low risk borrowers", a fair prediction is the one which is independent of any sensitive or irrelevant attributes. In other words, we do not want to accept or reject a student because of his/her race or social media posts.

The problem has been discussed from different angles: there are those who have provided empirical evidence of discrimination [see stories in media spotlight such as ProPublica's investigation (Angwin et al., 2016) or the MIT study on facial recognition (Natasha, 2019)], while others have addressed the problem of quantifying discrimination and proposed specific heuristics to mitigate bias in classification tasks, e.g. see (Zliobaite, 2015; Zemel et al., 2013).

The debate within the machine learning community has been productive and has led to several meaningful effects. Firstly, it raised awareness among researchers and

✉ Teresa Scantamburlo
teresa.scantamburlo@unive.it

[1] European Centre for Living Technology, Ca'Foscari University of Venice, Dorsoduro 3911, Calle Crosera, 30123 Venice, Italy

[1] Note that a decision can be categorical (e.g. "rejected" / "not rejected"), but the same outcome can be achieved by using a numerical function computing a probability value or a score for each instance and setting up a threshold for making the assignment.

practitioners undermining the alleged objectivity of algorithmic decisions (Hardt, 2014). Secondly, a machine learning approach enabled the study of fairness in a precise and quantifiable way, i.e. by studying the statistical conditions which determine the direct or indirect influence of a protected attribute on a particular decision outcome. Moreover, the statistical framework has offered valuable tools to establish boundaries and assess possible trade-offs in a view to help decision makers ponder solutions, pose constraints, and establish protocols for action.

However, by and large, the ongoing discussion has moved along empirical considerations, e.g. measuring the magnitude of discrimination or imposing constraints to existing machine learning techniques. In other words, the setup of the problem and the adequacy of proposed solutions were mostly addressed as an internal discussion that only a few people can access (e.g. those with the technical knowledge and skills), while other relevant stakeholders, such as policy makers and users, remain outside. The consequences of this prevailing attitude are several. In the first place, there was a cultural change: a long-standing issue of our society became a new attractive scientific puzzle that generated further sub-problems of a strictly computational nature with no substantial relations to the social context giving rise to the scientific effort (Powles & Nissenbaum, 2018; Selbst et al., 2019). In the second place, confining the issue of algorithmic fairness to a computational problem hinder the field of machine learning from assessing its own solutions in the light of other disciplinary perspectives and the present historical context. In short, it may cause the filed to miss relevant conceptual questions that have nothing to do with the capacity of algorithms to solve the problem of fairness in the machine learning domain. Instead, conceptual problems put the field in front of higher-order questions, such as: what kind of assumptions and decisions are implied by the algorithmic account of fairness? How do the results of the empirical approach penetrate public debate? What kind of reflection and deliberation that the different stakeholders, including citizens, may have over available metrics for fairness? What considerations would we need to keep algorithmic fairness tied to other human values (e.g. inclusion, solidarity, freedom, etc.)?

In this paper, I will consider the problem-solving approach to fairness from the perspective of the philosophy of science. My starting point is the simple idea that, like many other disciplines, machine learning attempts to solve empirical problems, and, when it comes to fairness, makes no exception. Nonetheless, empirical problems do not exhaust the scope of intellectual activity. According to the philosopher of science Larry Laudan, any discipline confronts with conceptual difficulties which regard the theoretical constructs of a developed theory or solution (Laudan, 1977). Conceptual problems can refer to ambiguities or inconsistencies of the premises of a theory, conflicts with other doctrines and tensions with entrenched values or beliefs, among others.

I will argue that so far the debate on algorithmic fairness has been too focused on empirical considerations and less attentive to conceptual difficulties. Starting from three claims of the empirical paradigm (see the section "Conceptual problems in fair machine learning"), I will provide some stimuli to critically reflect on specific difficulties that affect the empirical solutions and ultimately refer to the conceptual dimensions of the problem.

This paper is structured as follows. In section "The empirical and conceptual sides of a discipline", I will sketch out some key notions of Laudan's philosophy of science that would serve to frame the discussion on algorithmic fairness. In section "The empirical account of fairness", I will give an overview of how machine learning is addressing the (empirical) problem of fairness. In section "Conceptual problems in fair machine learning" I will introduce a few conceptual difficulties that relate to theoretical assumptions or implicit beliefs of the dominating empirical approach, and then conclude with some final remarks.

The empirical and conceptual sides of a discipline

The idea that science is essentially a problem-solving activity is well grounded in the history and philosophy of science.[2] However, in Larry Laudan's view of science, the idea of problem-solving is an authentic cornerstone that extends far beyond scientific problems and may well reflect any intellectual endeavor.[3]

According to Laudan, all intellectual disciplines aim at solving problems and scientific theories follow the same pragmatics: "Theories matter [...] insofar as [...] they provide adequate solutions to problems. If problems constitute the questions of science, it is theories which constitute the answers." (Laudan, 1977, p. 13). So, from this point of view, we may not see much difference between the role of a theory and that of an algorithm since both attempt to provide acceptable solutions to determined problems.

In his framework, Laudan identifies two types of problems: the class of empirical problems and the class of conceptual problems. The former is probably the most intuitive

[2] In addition to Larry Laudan, even Karl Popper acknowledged that: "The activity of understanding is essentially the same as that of all problem solving." (Popper, 1972, p. 166).

[3] Indeed, he does not believe "that "scientific" problems are fundamentally different from other kinds of problems (though they often are different in degree)." (Laudan, 1977, p. 13).

one since it deals with the "first order questions" of a discipline, which spring from basic observations and arise in a certain context of inquiry. For example, questions like "why do people learn to process spoken language more easily than they do with writing?" or "how do communication disorders interfere with learning skills?" can constitute genuine empirical problems in linguistics. Likewise, if we turn to the field of machine learning, empirical problems can originate from questions such as: "what can be inferred from a set of observations?"; "what class of functions best generalize beyond a given set of examples?"; "how many examples are needed to learn to classify?". Note, however, that empirical problems must not be considered as pieces of unambiguous data found in the real world.[4] On the contrary, they are always perceived through the lens of the concepts and abstractions of the field where they arise (Laudan, 1977). Thus, for example, it is because of statistical assumptions and the inductive nature of the problem at stake that the relation between the size of the hypothesis space and the number of observed examples is perceived as problematic in the field of machine learning.

The class of conceptual problems, on the other hand, is less apparent than that of empirical problems, but its role is anything but marginal. Conceptual problems are higher order questions that concern the "adequacy of solutions to empirical problems." (Laudan, 1977, p. 15). Note that the distinction between empirical and conceptual problems does not relate to the naive polarization between abstract *versus* verifiable observations. The key point is the locus of intervention when solving a problem and assessing its solution. When a researcher acts within the field of study where the problem arose, his/her intervention is empirical (first-order problems), regardless of the methods used in that field (be those argumentation or experiments). When he/she moves outside the contours of "his/her" field, the research work rises up to a conceptual level, which does not necessary imply that consideration will be more abstract. Indeed, this move can create new problems (second-order problems) which need interactions with different approaches and points of view.

Conceptual problems can have an internal or an external source. For example, a theory can suffer from terminological ambiguity or circularity (e.g. Faraday's model of electrical interaction employed the same concept of action-at-distance that he was actually supposed to eliminate). But a theory can also conflict with other doctrines or extra-scientific beliefs. Note that a theory, when confronted with external elements, can also deal with a body of knowledge which does not obey the conventional canons of empirical or formal sciences such as ethics, metaphysics or worldviews. Examples of such tensions include the mismatch between Newtonian ontology and the philosophical notions of "substance" and "properties" in the 18th century[5] or the contrast between Darwin's theory of evolution and religious views or other social practices like altruism and love.

Laudan's distinction between empirical and conceptual problems is part of a framework whose goals go beyond the purpose of this paper. However, the underlying intuition invites us to open the assessment of a problem solution to a broader set of considerations drawing on different disciplinary perspectives and more informal ways of thinking.[6] If applied to the problem of algorithmic fairness, this translates into concrete questions such as: "how do fairness metrics relate to the various conceptions of justice?"; "what kind of assumptions do they presuppose?"; "what type of reasoning and decisions do they solicit?".

Research moving in this direction already exists. For example, Selbst et al. (2019) pointed out distinct "traps" that occur when failing to acknowledge the wider social context of fair decisions, while Heidari et al. (2019) proposed a conceptual mapping between existing definitions of algorithmic fairness and available notions of equality of opportunity in political philosophy. However, since most of these contributions rest on approaches and languages that are distant from those commonly used by the machine learning community, one may consider them as marginal for making progress in algorithmic fairness. On the other hand, people who are not familiar with such languages and approaches could misunderstand the very contribution of fair machine learning, and therefore under- or over-estimate the solutions offered by the field. Laudan's lesson is that we can fill these cultural gaps by keeping the empirical and the conceptual problems united, an idea that is rooted in a broad sense of rationality.[7]

[4] Laudan stresses that empirical problems do not necessarily refer to real world facts or unambiguous evidence. The history of science has plenty of anecdotes of "fake" stories that were treated as they were real problems, and examples of known phenomena that were not considered as a problem at all. Examples and more details can be found in (Laudan, 1977, pp. 15–17).

[5] In the 18th century Newtonians confronted with the conceptual problem of reconciling the language of "substance" and "properties" with the ideas of "bodies" and "forces": "can bodies exert force at points far removed from the bodies themselves? what substance carries the attractive force of the sun through 90 million miles of empty space so that the earth is pulled towards it?" (Laudan, 1977, p. 61).

[6] These would include also considerations which do not obey a specific method of inquiry (e.g. experimental method, philosophical argumentation, archival study, etc.), but arise out of informal reflections and practical reasoning. Examples of these types of knowledge are intuitions about social problems and practical rules learned from repeated experiences.

[7] Laudan's conception of science extends the contours of rationality. While several philosophers and sociologists of science undermined the role of worldview difficulties – making them "pseudo-problems" or a sign of "irrationality" of a field – Laudan considers them as legitimate factors in the rational development of science. Motivated by the fact that scientific theories cannot exhaust the domain of rational beliefs, he suggests that worldview difficulties challenge both sides of

In the end, my proposal is to take Laudan's view of science as an encouragement to expand the assessment of fair machine learning and "cast our nets of appraisal sufficiently widely" so as to "include all the cognitively relevant factors" (Laudan, 1977, p. 128) which are present in our time. The next sections can be therefore interpreted as an exercise of this broad appraisal and an attempt to address non-trivial conceptual difficulties in fair machine learning.

The empirical account of fairness

Before undertaking a discussion of conceptual issues that affect algorithmic fairness, I want to outline the empirical traits of the problem. In general, the goal of a fair algorithm is to avoid unjustified discrimination.[8] This occurs, for instance, when two individuals, who differ only in a sensitive attribute (say gender), get different outcomes: one is hired and the other is not hired. If undetected, biased decisions can in fact amplify and systematize existing social inequities.

In the field of machine learning the problem of fairness is statistical in nature: given a set of individuals described by a series of legitimate and protected attributes the problem boils down to quantifying the degree of independence between the outcome (i.e. the target variable) and the protected features. The problem can be easily stated in information-theoretic terms. For example, if $P(X)$ is the probability of "being a woman" and $P(Y)$ is the probability of "being hired," the objective is to measure the mutual information between the two variables, $I(X, Y)$, where $I(X, Y) = 0 \iff X$ and Y are independent, i.e. $P(X, Y) = P(X)P(Y)$. Therefore, the lower the mutual information $I(X, Y)$ the greater the independence between the two variables and the probability of fair classification. Note that unfairness rarely derives from a deliberate discriminatory design choice, it rather connects to data, and not to the algorithm itself Menon and Williamson (2018). Thus, in the field of machine learning scholars are more familiar with "indirect discrimination" or "statistical discrimination" Zliobaite (2015).

There exist many ways to formalize the notion of fairness. If we consider a simple binary decision where $Y \in \{0, 1\}$ is a binary, target variable, Z is the set of legitimate features, $X \in \{0, 1\}$ is a binary variable representing a protected attribute, and $\hat{Y} = f(X, Z)$ is the classifier[9] we want to build, three popular fairness criteria are:

- *Statistical parity* (Dwork et al., 2012): the probability of receiving a positive/negative classification should be the same in any group (or the proportion of positive/negative predictions should be the same for each group). The classification should be independent of the protected attribute.

  $$P(\hat{Y} = 1|X = 1) = P(\hat{Y} = 1|X = 0)$$

- *Calibration* (Kleinberg et al., 2017): the probability of being a positive instance conditioned on the received classification should be the same in any group, i.e. groups with the same probability of being classified as a positive instance should be in fact a positive instance.

  $$P(Y = 1|\hat{Y} = 1, X = 1) = P(Y = 1|\hat{Y} = 1, X = 0)$$

- *Equalized odds and equality of opportunity* (Hardt et al., 2016): the probability of true positive and false positive is equal across groups. When this criterion relates to the advantaged outcome (e.g. $X = 1$ ="being hired"), it is called "equality of opportunity":

  $$P(\hat{Y} = 1|Y = 1, X = 1) = P(\hat{Y} = 1|Y = 1, X = 0)$$

Note that these fairness metrics, in the end, pose constraints on how the classifier performs on different groups. For example, calibration enforces equal precision,[10] while equality of opportunity requires an equal, true positive rate [(for a nice parallel between fairness and performance metrics see Berk et al. (2018)]. Alternatives to isolating the effect of sensitive attributes commit to different principles. In this regard, Dwork et al. (2012) puts forward a theoretical framework which rests on the assumption that "any two individuals who are similar with respect to a particular task should be classified similarly" (Dwork et al., 2012, p. 1). Additionally, Jia et al. (2018) proposes a method developed in the context of domain adaptation to make predictions that are "right for the right reason," [(e.g. "gender of a subject in an image should not be based on the background." (Jia et al., 2018, p. 1)].

Other studies have highlighted different aspects of the problem. For example, Menon and Williamson (2018) analyzes the optimal trade-off between fairness and predictive

Footnote 7 (continued)

the conflict (scientific and extra-scientific) testing the quality of their underlying assumptions and their ability to provide satisfactory solutions.

[8] A broad perspective on the problem of discrimination in data analysis, including its legal and cultural underpinnings, is provided (Romei & Ruggieri, 2014).

[9] In short, a classifier is a function estimating the value of the target variable, which can be a class label (e.g. "high risk") or a numerical value (e.g. the probability to give back money).

[10] In information retrieval and machine learning, the notion of precision is a popular performance metric which represents the ratio between true positive instances and instances classified as positive. For a basic introduction see https://en.wikipedia.org/wiki/Precision_and_recall.

accuracy that is inherent in learning algorithms (with fairness constraints). Other studies revealed impossibility results (Chouldechova, 2017; Kleinberg et al., 2017) proving that some fairness metrics are not compatible with one another. This condition is also one of the reasons behind the dispute surrounding COMPAS,[11] a famous predictive tool which turned out to be well-calibrated, but unsatisfactory with respect to the error rate balance (e.g., false positive rates were much higher for Afro-Americans compared to others). Overall, empirical evidence has made it clear that when we approach fairness in algorithmic decision making we need to take important choices, for instance, as to whether maximizing accuracy or fairness, and, as for the latter, what metric would make more sense in the context of application.

Conceptual problems in fair machine learning

In the field of machine learning the empirical approach to fairness is a natural option. The problem is tackled through the lens of machine learning constructs and concepts, and making use of optimization constraints, statistical assumptions and performance metrics, among others. This well reflects the idea that algorithmic fairness can be considered as a problem of the field, on par with over-fitting or lack of data. However, as we said before, a scientific investigation may also generate conceptual, i.e. non-empirical, problems, a few of which are addressed in this section. In particular, I will discuss three broad conceptual difficulties which connect to the empirical account of fairness. Note that my aim is not to solve these difficulties but to propose an exercise for reflection that could stimulate a broader appraisal of empirical solutions.

Problem 1: Can fairness be engineered?

A first, quick answer would be 'yes' since many scholars have suggested that fairness can be tested and measured (Zliobaite, 2015). Notwithstanding, framing a complex notion in terms of quantities and formal relations can be highly problematic. Indeed, when we define a fairness criterion we impose limits to the general understanding of the concept. This is a natural consequence of any formal definition, which is supposed to retrieve (literally, "to pull away") only the elements that are necessary and useful for the investigation. Accordingly, engineering fairness implies the selection of what counts most, where to direct attention, and, as a consequence, introduces some level of simplification.

For example, Barocas et al. (2017) suggests that the machine learning approach tends to focus on harms which derive from unequal allocations of goods and opportunities (e.g. when certain groups are repeatedly excluded from getting highly qualified jobs), while it ignores those harms connected to unjust identity representations, such as search engine results displaying racist or sexist prejudice (Snow, 2018).

Other simplifications arise when we consider two different concepts of discrimination: the comparative and non-comparative ones (Hellman, 2016). The comparative notion of discrimination determines if an outcome is fair by making reference to other individuals (for example, when a subject is treated worse than another one). Many definitions of fairness, like the criteria introduced in the previous section, build upon the same intuition, i.e. they are based on the probability of a certain outcome with respect to different groups of people. However, there also exists a non-comparative view of discrimination where justice consists in treating "each individual as she is entitled to be treated." (Hellman, 2016).

A non-comparative conception of discrimination is independent of other individuals. Instead, it needs some standards that specify which rights people are entitled to and which criteria are relevant for a certain decision. Note that this last conception of justice would be hardly applicable in the context of algorithmic decision making. Indeed, the majority of today's machine learning applications mask the relevant criteria for making a decision, and a fundamental reason is that our best performing models do not need to specify a decision rule. The criteria to take decisions are often implicit, encoded into the masses of input-output relations available in the wild (Halevy et al., 2009) and used to train the systems.

These observations do not only relate to the fact, already stressed in the literature (Heidari et al., 2019), that fairness criteria make assumptions about the notion of the equality and justice they implement. They also suggest that engineering fairness implies the translation of a complex notion into a static, as well as partial, definition that might be then applied to a variety of decisions with different implications (e.g. advertisements, media treatments, school admissions, etc.). As we establish measurements and formalization, we enforce a certain (abstract) meaning and, with it, a mechanical rule for action determining what is fair or unfair. While formal definitions are essential to reduce subjective interpretations and make fairness an operational requirement, ambiguities and inconsistencies may not disappear. For example, one may ask: "Who is responsible for setting such

[11] On the one hand, Angwin et al. (2016) showed that COMPAS generated unbalances between error rates based on ethnicity, i.e. the algorithm wrongly flagged black defendants as high risk at almost twice the rate as white defendants and mislabelled the latter as low risk more often than black defendants. On the other hand, Flores et al. (2016) showed that COMPAS' performance is consistent with the calibration criterion.

standards?" or "How may one decide which measures are appropriate and which are not?".

In addition, if we consider the variety of definitions and metrics proposed in machine learning research (Verma and Rubin, 2018), one may ask if there is a legal correlate or doctrine for each. However, even if such legal background were available, the technical criteria could nonetheless fail to meet the policy requirements they sought to satisfy. In particular, Lipton et al. (2018) showed that algorithmic solutions which rest upon the doctrine of disparate impact,[12] under certain circumstances, can incidentally create situations of treatment disparity, where, for example, people within the same group are treated differently on the basis of irrelevant features (Lipton et al., 2018). Finally, in case of multiple criteria associated with the same legal concepts, there still might be questions on how and why they technically differ.

Social sciences can provide some useful tools to explore the theoretical constructs of fairness metrics and the choices which motivated them. For example, drawing on quantitative social science, Jacobs and Wallach (2021) introduces the language of measurement modelling to clarify possible mismatch between theoretical abstractions, which cannot be directly measured, such as fairness and their operationalizations through observed data.[13]

Problem 2: How to determine acceptable trade-offs between fairness metrics and other relevant variables?

Most of the studies conducted so far have suggested the existence of possible trade-offs regarding either the fairness criteria that can be simultaneously satisfied or the balance one may achieve between accuracy and fairness (given a specific fairness notion). However, there is no mathematical answer to the choice of an "acceptable" or "desirable" balance. Once we have quantified the variables of interest, be those for accuracy or fairness, the empirical framework can help us identify the "best" possible compromises. But the discussion on whether these are acceptable is left to human deliberation. Indeed, machine learning scientists have acknowledged that algorithms "are agnostic about what the target outcome for fairness should be." (Berk et al., 2018, p. 23) For example, a human decision maker may address questions such as: "what fairness criteria should be prioritized?" and "to what extent may a loss of accuracy be acceptable?". At the same time, answering these and similar questions requires moral and social skills that are as important as technical ones.

Discussing social issues is not just a matter of expressing preferences or a decision policy that should be encoded into an algorithm. It concerns the exercise of moral judgment and a careful reflection on what serves human well-being and common good in the particular context of algorithm's application. For example, in the domain of criminal justice an often-quoted example of social dilemma is the decision as to whether to reduce the crime rate or the jail population[14]. While such a decision is part of a wider political discussion, it does not fit the layout of an optimization problem. The discussion of social issues, such as justice, is articulated and needs time to mature. For example, understating the value (i.e., the "weight") of mass-incarceration in a policy prediction problem (e.g. bail decisions) can be hardly isolated from broader considerations, such as the quality of prisoners' life, the purposes of punishment, the rehabilitation of the offender, etc. Moreover, the discussion can involve a variety of actors, e.g. government, judges, citizens, correctional agencies, etc. each of whom has a particular viewpoint of the problem and reflects the needs and expectations of people affected by algorithmic decision-making.

Understanding the social context where fairness issues arise is also a matter of discretion and practical wisdom. The application of abstract concepts and principles in particular cases profits not only from a technical understating of those notions, but also from the practice of virtues and responsible action. In ancient philosophy, virtues are attitudes which come about as a result of habit and whose practice can guide individual and group decision-making even in modern organizations [for an example of their application to a big tech company see Neubert and Montañez (2020)]. Although their number varies among philosophers and different accounts exist, classical tradition identified some virtues that can still offer valuable guidance. These include, for example, prudence, a disposition which deals with human

[12] In United Sates, there are two main interpretations of discrimination law: disparate treatment and disparate impact. In short, the former is concerned with intentional discrimination and the latter with apparent neutral decisions which negatively affect members of a protected class.

[13] "Articulating the distinction between constructs and their operationalizations allows us to make assumptions explicit, identify the source of existing fairness-related harms, and characterize and even remedy potential harms. The language of measurement allows us to negotiate and evaluate our constructs and operationalizations thereof, providing a common framework to unite and clarify existing conflicts in the fairness, accountability, and transparency in machine learning community." (Jacobs & Wallach, 2021, p. 2)

[14] Indeed, one of the reasons put forward to motivate the use of machine learning and statistical tool in recidivism prediction is the need to reduce jail population. For example, some policy simulations have shown that a machine learning model can bring important welfare gains in bail decisions: "either crime can be reduced by up to 24.8% with no change in jailing rates, or jail populations can be reduced by 42.0% with no increase in crime rates" (Kleinberg et al., 2017).

foresight and consideration of long-term goals, and equity, a virtue that closely relates to the notion of fairness.

Equity, like fairness, contributes to the realization of justice, but its meaning extends beyond the idea of impartiality and connects to the ability of making tailor-made, particularized judgments (Shiner, 1994). This virtue counterbalances the effects of laws and, in particular, the deficiencies caused by its mechanical application. It demands the use of judgment in adapting the rules to particular situations where the subject matter is true for the most part.[15] This flexibility would be valuable even in the context of fair machine learning, where compromises can vary depending on the domain application - for example, a loss of accuracy could be acceptable in a fraud detection application but it may not be tolerable in a diagnostic tool for cancer. Moreover, the exercise of virtues, like equity, allows to think out of the box and discover hidden tensions, as well as identify which are morally salient for the case at stake.

Human and social deliberation is a complex and, often, arduous effort. Unfortunately, popular digital technologies, in particular those enabled by AI and ML, rely upon a simplified view of rationality,[16] where decisions are reduced to a value-maximizing equation. The influence of this intellectual framework within the field of artificial intelligence and machine learning has somewhat limited the knowledge and adoption of different ways of thinking and approaching decisions. As reported by some organizations of engineers (IEEE, 2020), it is crucial to go beyond existing practices and to develop additional skills that can help artificial intelligence developers engage with and address ethical challenges. One way to move in this direction is to allocate more time for generative discussion and debate within and outside organizations (IEEE, 2020) as well as increase the interactions with stakeholders.[17] Further measures can include changes in education and vocational training of AI practitioners to support them in the development of ethical skills [see for example Grosz et al., (2019)].

Problem 3: Can algorithmic fairness be fixed?

In the last few years the problem of fairness has become a top priority. Documents put forward by governments and organizations recommend fairness as a fundamental requirement while acknowledging the role of technical and organizational measures. For example, the European Guidelines for Trustworthy AI include the principle of fairness among the four "moral imperatives" that should inspire the design and deployment of artificial intelligence systems (HLEG-AI, 2019). Often, this effort translates into a by-design approach which led the community to develop a large pool of methods, ranging from algorithmic remedies to *ad hoc* testing and audit processes. Note that most of these resources consist of empirical solutions, i.e. they solve a well-structured definition of the problem in a specific scientific domain or culture. However, this abundance raises conceptual issues that, more or less implicitly, relate to the temptation of technological solutionism (Morozov, 2013).

The possibility to encapsulate fairness into a mathematical formula or a procedure might hide the illusion that we can isolate values and purse them separately. The market of tech solutions has, in fact, increased in the last few years with more and more organizations embracing the challenge of algorithmic fairness.[18] Unfortunately, this has also contributed to a poor understating of the problem, and fragmented into a multitude of partial definitions and methods that can operate independently or with few interrelations.

A conceptual difficulty which is inherent in the empirical approach to fairness is the limited consideration of how fairness issues connect to other human values. Contemporary philosophy has presented fairness as an essential condition to build a society of free and equal citizens (Wenar, 2017) and in democratic societies we see it at work in the pursuit of justice, inclusion, and solidarity. Additionally, fairness

[15] For Aristotle, the subject matter justifies the precision and, hence, the flexibility, that should be demanded to human judgment. In the case of practical knowledge, judgment cannot be as precise as it is in theoretical sciences, like math. Indeed: "noble and just actions, which political science investigates, exhibit much variety and fluctuation" (Aristotle, 2009, p. 1094b15). This flexibility is also described through a metaphor which links equitable judgment to a mason's rule made of lead found on the island of Lesbos and used to reproduce irregular curves: "For when the thing is indefinite the rule also is indefinite, like the leaden rule used in making the Lesbian moulding; the rule adapts itself to the shape of the stone and is not rigid, and so too the decree is adapted to the facts." (Aristotle, 2009, p. 1137b30). So, like the leaden rule, equitable judgment reaches the goal of justice adapting a general law to a particular situation. For example, the mechanical application of law forbidding stealing would fail to do justice in the case of a man taking without permission a neighbor's garden water hose to fight a fire in his own house (Shiner, 1994).

[16] Usually artificial intelligence and machine learning are based on the concept of a rational agent defined as an individual that acts to maximize expected utility (Russell, 2019).

[17] Similar examples include initiatives promoting dialogue with citizens and advancing public understanding of AI technologies such as the *Partnership on AI* (https://www.partnershiponai.org), the *Tactical tech* (https://tacticaltech.org), and the *AI4EU Observatory* (https://www.ai4eu.eu/ai4eu-observatory). Others try to encourage mutual understating among stakeholders and the dialogue with under-represented groups, such as the so-called "diverse voices" methodology developed by the tech Policy Lab of Washington University.

[18] There are several organizations which are thriving on the design of fairness interventions. These can include for profit initiative such as *Unbiased* (https://unbiased.cc/) and research projects such as *Aequitas*, an open source toolkit developed by the University of Chicago (http://aequitas.dssg.io/).

operates in concert with other democratic principles, such as transparency and accountability, to build trust within communities. It is rare, if not impossible, to implement these principles in isolation. One influences the other. We cannot guarantee, for instance, fair elections, if we do not sustain freedom of speech and association along with open and transparent communication. Unfortunately, past elections in Europe and United States have given us renewed evidence on how micro-targeted advertising can erode all these principles and place our democracy under strain (Editorial, 2018). Thus, implementing fairness safeguards without an overall assessment of connected dimensions might not be sufficient to fix the problem.

This interconnectedness also has practical consequences on the attempts to implement fairness. As seen in the previous sections, most empirical solutions to algorithmic fairness act on a set of elements identified as a part of the problem definition. Usually these include: personal data, proxy variables, asymmetries in data sets, error rates disparities, etc. However, there are other actions which, even if they fall under the umbrella of other ethical requirements, would fit for fairness as well. For example, the General Data Protection Regulation recommends the adoption of transparency safeguards to avoid unfair processing. Specifically, articles 12–14 require anyone using an algorithm for automatic decision-making to inform data subjects of the existence of this processing and provide information about its purpose and logic, as well as the significance and envisaged consequences .[19] Indeed, transparent communication would help data subjects make use of their "right not to be subject to a decision based solely on automated processing" (see art. 22 GDPR) and ask for human intervention, if necessary. In a similar spirit, the OECD AI Principle of "Human-centred values and fairness"[20] acknowledges other important practices such as impact assessments on human rights, human oversight (i.e., a "human in the loop"), and codes of ethical conduct.

In the end, considering fairness within a web of interrelated values which contribute to the flourishing of human society raises the question as to whether an ultimate solution really exists. If the notion of fairness has many facets and relates to an indivisible set of human rights, including freedom and human dignity, the remedy will look more like a process rather than an empirical solution. If we look at fairness in the broader context of policy problems, Horst Rittel and Melvin Webber would say that it is a wicked problem, originating from the obstinate attempt to follow the same method of science (Rittel and Webber, 1973). Wicked problems have no clear traits, no exhaustive formulation, and, for this reason, they are never solved. "At best they are only re-solved - over and over again." Rittel and Webber (1973). The characteristic of a wicked problem is to have no stopping rule, meaning that the problem solver never knows when he has terminated his job. The social nature of fairness problems render any solution limited in time and scope since any progress made toward a solution opens new issues in need of further investigation. As Rittel and Webber (1973) points out, the decision to release a certain solution to a social problem is determined by reasons which are external to the problem, such as lack of resources or time.

While Rittel and Webber (1973) points to the intrinsic limits of human solutions to social problems, I would like to emphasize the positive aspect. Specifically, I suggest to shift the focus from solutions to processes. If solutions are inherently incomplete, the process to achieve them is the key to broaden our knowledge of the problem and keep the search alive. In other words, the goal is the process, not the solution. With regards to fairness, this would mean, for example, that the effort made to gather and listen to the needs of stakeholders and representatives of minorities is even more important than the resulting outcome. It is the process that creates the necessary culture and attitudes to move the search forward and shape an entire field. A process can create more value than a solution.

Concluding remarks

Herein, I gave an overview of the problem of fairness in the context of machine learning. Based on Laudan's account of science I suggested two different levels of inquiry based on the distinction between empirical and conceptual problems. After outlining the empirical account of fair machine learning I discussed a few conceptual difficulties surrounding three problems: (1) the definition of fairness in formal terms; (2) the role of human deliberation in discussing conflicts and balances between fairness metrics and other variables; and (3) the expectation of definitive solutions.

Note that what I have discussed so far on fairness may also hold for other problem domains which have raised great attention within machine learning community, such as privacy and transparency. Even in these fields empirical solutions have being thriving and suggesting similar conceptual issues addressed in this paper. For example, Gürses (2014) acknowledged that engineering privacy could be an ideal that misleadingly suggests we can engineer social and legal concepts. In the field of explainable AI scholars are

[19] Further details on the implications of GDPR for algorithmic fairness can be found in the letter of the European Data Protection Board to the European Parliament: https://edpb.europa.eu/sites/edpb/files/files/file1/edpb_letter_out2020_0004_intveldalgorithms_en.pdf.

[20] The Organisation for Economic and Cooperative Development published a set of principles promoting a human-centric approach to AI in spring 2019. These principles are available online: https://oecd.ai/ai-principles.

increasingly aware of the fact that explainability is a kaleidoscopic problem and in some contexts powerful visualizations tools are not enough to explain a prediction (Kuang, 2017).

Ultimately, this analysis has suggested that the assessment of a field cannot be measured just by looking at the empirical side of a discipline. The broad account of scientific progress in Larry Laudan's philosophy of science suggests that the appraisal of machine learning is a multi-factorial affair that involves distinct levels of answer, the confrontation with different disciplines and the capacity to understand the limits of a problem solution.

In recent times, the philosophy of technology has clearly expressed the idea that technical artifacts are a powerful vehicle of ideologies and moral values (van de Poel & Royakkers, 2011). Technical artifacts do not only fulfill a specific task, they also shape the actions and experiences of their users. For example, assessment platforms, like those used for recruiting, mediate the relation between candidates and the employer. On one hand, they influence the presentation of candidates through ranking and scoring, and on the other, give incentives to meet certain standards of qualities and success. Acknowledging the role of technology as a moral mediator is an essential step to train engineering and promote responsible design. In this regard, the analysis of conceptual difficulties adds further insights. If it is true that technical artifacts are the "bearer of moral values," van de Poel and Royakkers (2011) then it is also the case that we need to go beyond them when addressing the social problems that they raise. Conceptual issues help us to realize that redesigning the technological mediation is only part of the solution, which lies in the road ahead.

Funding Open access funding provided by Università Ca' Foscari Venezia within the CRUI-CARE Agreement. The research was supported by the project A European AI On Demand Platform and Ecosystem (AI4EU) H2020-ICT-26 under (Grant 825619).

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest Not applicable.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. Technical report, ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aristotle. (2009). The nicomachean ethics (D. Ross and L. Brown, Trans.). Oxford University Press.

Barocas S. Crawford K., Shapiro, A., & Wallach, H. (2017). The problem with bias: Allocative versus representational harms in machine learning. In *9th annual conference of the special interest group for computing, information and society. Philadelphia*. http://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods and Research, 50,* 3–44.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047.

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, ITCS '12* (pp. 214–226). Association for Computing Machinery. https://doi.org/10.1145/2090236.2090255.

Editorial, O. (2018). *Democracy dies without transparency and fairness*. The Guardian.

Flores, A., Bechtel, K., & Lowenkam, C. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals, and it's biased against blacks. *Federal Probation, 80,* 38.

Grosz, B. J., Grant, D. G., Vredenburgh, K., Behrends, J., Hu, L., Simmons, A., & Waldo, J. (2019). Embedded ethics: Integrating ethics across cs education. *Communications of the ACM, 62*(8), 54–61. https://doi.org/10.1145/3330794.

Gürses, S. (2014). Can you engineer privacy? *Communications of the ACM, 57*(8), 20–23. https://doi.org/10.1145/2633029.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems, 24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36.

Hardt, M. (2014). How big data is unfair. understanding unintended sources of unfairness in data driven decision making. Medium. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, R. Garnett (Eds.) Advances in neural information processing systems (vol. 29, pp. 3315–3323). Curran Associates, Inc. http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

Heidari, H., Loi M., Gummadi, K., & Krause, A. (2019). A moral framework for understanding fair ml through economic models of equality of opportunity. In Proceedings of the conference on fairness, accountability, and transparency, FAT* '19 (pp. 181–190). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287584

Hellman, D. (2016). Two concepts of discrimination. *Virginia Law Review, 102*(4), 895–952.

HLEG-AI. (2019). Ethics guidelines for trustworthy ai. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

IEEE. (2020). Addressing ethical dilemmas in ai: Listening to engineers. https://standards.ieee.org/initiatives/artificial-intelligence-systems/ethical-dilemmas-ai-report.html

Jacobs A. Z., & Wallach H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery* (pp 375–385). New York. https://doi.org/10.1145/3442188.3445901.

Jia, S., Lansdall-Welfare, T., & Cristianini, N. (2018). Right for the right reason: Training agnostic networks. In *Advances in intelligent data analysis XVII - 17th international symposium, IDA 2018, 's-Hertogenbosch, The Netherlands, October 24-26, 2018, Proceedings*, pp. 164–174. https://doi.org/10.1007/978-3-030-01768-2_14.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions*. *The Quarterly Journal of Economics, 133*(1), 237–293. https://doi.org/10.1093/qje/qjx032.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491–95.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent tradeoffs in the fair determination of risk scores. In C. Papadimitriou (ed.) *8th Innovations in Theoretical Computer Science Conference (ITCS 2017), Leibniz International Proceedings in Informatics (LIPIcs)* (vol. 67, pp. 43:1–43:23).

Kuang, C. (2017). Can a.i. be taught to explain itself? The New York Time Magazine. https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

Laudan, L. (1977). *Progress and its problems. Towards a theory of scientific growth*. University of California Press.

Lipton, Z., McAuley, J., & Chouldechova, A. (2018). Does mitigating ml's impact disparity require treatment disparity? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.) *Advances in neural information processing systems* (vol. 31, pp. 8125–8135). Curran Associates, Inc. http://papers.nips.cc/paper/8035-does-mitigating-mls-impact-disparity-require-treatment-disparity.pdf.

Menon, A., & Williamson, R. (2018). The cost of fairness in binary classification. In S.A. Friedler, C. Wilson (Eds.) Proceedings of the 1st conference on fairness, accountability and transparency, *Proceedings of Machine Learning Research*, (vol. 81, pp. 107–118). PMLR. http://proceedings.mlr.press/v81/menon18a.html.

Morozov, E. (2013). *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

Natasha, S. (2019). Amazon is pushing facial technology that a study says could be biased. New York Times. https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html.

Neubert, M. J., & Montañez, G. D. (2020). Virtue as a framework for the design and use of artificial intelligence. *Business Horizons, 63*(2), 195–204. https://doi.org/10.1016/j.bushor.2019.11.001.

van de Poel, I., & Royakkers, L. (2011). *Ethics, technology, and engineering: An introduction*. Wiley-Blackwell.

Popper, K. (1972). *Objective knowledge: An evolutionary approach*. Oxford University Press.

Powles, J., & Nissenbaum, H. (2018). The seductive diversion of 'solving' bias in artificial intelligence. Medium. https://onezero.medium.com/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53.

Rittel, H., & Webber, M. (1973). Dilemmas in a general theory of planning. *Policy Sciences, 4,* 155–169.

Romei, A., & Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review, 29*(5), 582–638. https://doi.org/10.1017/S0269888913000039.

Russell, S. (2019). *Human compatible: AI and the problem of control*. Allen Lane.

Selbst, A., Boyd, D., Friedler, S., Venkatasubramanian S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the conference on fairness, accountability, and transparency, FAT* '19 (pp. 59–68). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598.

Shiner, R. A. (1994). Aristotle's theory of equity. *Loyola of Los Angeles Law Review, 27*(4), 1245–1264.

Snow, J. (2018). Bias already exists in search engine results, and it's only going to get worse. MIT Technology Review. https://www.technologyreview.com/s/610275/meet-the-woman-who-searches-out-search-engines-bias-against-women-and-minorities/.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *Proceedings of the international workshop on software fairness, fairWare* (vol. 18, pp. 1–7). Association for Computing Machinery. https://doi.org/10.1145/3194770.3194776.

Wenar, L. (2017). John rawls. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (2017th ed.). Stanford University.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In: *Proceedings of the 30th international Conference on Machine Learning, ICML'13* (vol. 28, pp. 325–333).

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. CoRR. arXiv:abs/1511.00148.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.