AI & SOCIETY
https://doi.org/10.1007/s00146-024-01886-3

ORIGINAL PAPER

On prediction-modelers and decision-makers: why fairness requires more than a fair prediction model

Teresa Scantamburlo[1] · Joachim Baumann[2,3] · Christoph Heitz[3]

Received: 14 October 2022 / Accepted: 5 January 2024

Abstract

An implicit ambiguity in the field of prediction-based decision-making concerns the relation between the concepts of prediction and decision. Much of the literature in the field tends to blur the boundaries between the two concepts and often simply refers to 'fair prediction'. In this paper, we point out that a differentiation of these concepts is helpful when trying to implement algorithmic fairness. Even if fairness properties are related to the features of the used prediction model, what is more properly called 'fair' or 'unfair' is a decision system, not a prediction model. This is because fairness is about the consequences on human lives, created by a decision, not by a prediction. In this paper, we clarify the distinction between the concepts of prediction and decision and show the different ways in which these two elements influence the final fairness properties of a prediction-based decision system. As well as discussing this relationship both from a conceptual and a practical point of view, we propose a framework that enables a better understanding and reasoning of the conceptual logic of creating fairness in prediction-based decision-making. In our framework, we specify different roles, namely the 'prediction-modeler' and the 'decision-maker,' and the information required from each of them for being able to implement fairness of the system. Our framework allows for deriving distinct responsibilities for both roles and discussing some insights related to ethical and legal requirements. Our contribution is twofold. First, we offer a new perspective shifting the focus from an abstract concept of algorithmic fairness to the concrete context-dependent nature of algorithmic decision-making, where different actors exist, can have different goals, and may act independently. In addition, we provide a conceptual framework that can help structure prediction-based decision problems with respect to fairness issues, identify responsibilities, and implement fairness governance mechanisms in real-world scenarios.

Keywords Algorithmic fairness · Prediction-based decision systems · Responsibility · Human-in-the-loop · Prediction modeler · Decision maker

Teresa Scantamburlo, Joachim Baumann and Christoph Heitz have contributed equally to this work.

✉ Teresa Scantamburlo
teresa.scantamburlo@unive.it

✉ Joachim Baumann
baumann@ifi.uzh.ch

Christoph Heitz
heit@zhaw.ch

[1] Department of Environmental Sciences, Informatics and Statistics, European Centre for Living Technology, Ca' Foscari University, Via Torino 155, Room Z.B17, Building Z, 30172 Venezia-Mestre, Italy

[2] Department of Informatics, University of Zurich, Andreasstrasse 15, 8050 Zürich, Switzerland

[3] School of Engineering, Zurich University of Applied Sciences, Technikumstrasse 81, 8400 Winterthur, Switzerland

1 Introduction

Algorithmic fairness has become a popular topic in the research community in recent years (Barocas et al. 2019; Kearns and Roth 2019), being increasingly addressed not only from a technical angle but also from a philosophical, political, and legal perspective (Binns 2018; Barocas and Selbst 2016). Algorithmic fairness is concerned with the consequences of prediction-based decisions on individuals and groups under the perspective of social justice (Mulligan et al. 2019). Since the beginning, the debate on algorithmic fairness has been focusing on the fairness of prediction models, which represent the core of Machine Learning (ML) research (Pedreschi et al. 2008; Calders and Verwer 2010; Kamishima et al. 2012; Dwork et al. 2012; Zemel et al. 2013). Therefore, it is not surprising that the focus

Published online: 16 March 2024

of attention was put on how prediction models can create unfairness.

We argue that the prediction model as such cannot be the reason for unfairness. It is the *usage* of the prediction model in its specific context which creates unfairness. For example, the recidivism risk model of the COMPAS tool (Angwin et al. 2016) in itself does not create racial discrimination. Such discrimination only occurs when it is used by judges to make decisions based on the COMPAS risk scores. Thus, the relationship between the properties of a prediction model, such as false-positive or false-negative rates, and possible harm for a specific group of the society, such as African Americans in the case of COMPAS, rests upon an assumption of how the output of a prediction model creates actual consequences in the lives of people.

This aspect is often neglected in the algorithmic fairness literature. Assumptions on this relationship are often left implicit, and a fixed relationship between the prediction outcome and the impact on lives is taken for granted (Chouldechova 2017; Kusner et al. 2017). While assuming such a fixed relationship is convenient for studying the impact of the prediction model's features on the resulting fairness, it also ignores a central part of almost all implementations of influential prediction-based systems, which is the part of the actual decision-making: Only insofar as the output of a prediction model changes the course of the world, it can create unfairness. And *how* a prediction changes the course of the world depends strongly on how the prediction is actually used.

As a prototypical case of how prediction models are implemented in real-world applications, we focus our discussion on *prediction-based decision systems*, where the outcome of ML prediction algorithms is used to make decisions affecting human subjects.[1] We imagine a (human or automated) decision-maker who is taking decisions on people or for people, while this decision is informed by a prediction of some features of these persons. This is the typical scenario for many of the discussed cases of algorithmic fairness, such as a bank taking loan decisions based on repayment prediction, an enterprise taking hiring decisions based on job performance prediction, or a university taking admission decisions based on academic performance prediction.[2]

In such prediction-based decision systems, we may distinguish two conceptually different functions: first, we have the function of *prediction*, performed by a prediction model which is fed with individual data of a person, and whose output is some form of prediction of a target variable attributed to this person, which is not known to the decision-maker at the time of decision-making. This prediction might come in the form of a score, a probability, or a point prediction. Second, we have the function of *decision* which is informed by the prediction, but in nearly all cases also influenced by additional parameters. For example, for a loan decision of a bank, not only the repayment probability but also the interest rate and the bank's business strategy may be decisive parameters. This idea has been studied in so-called cost-sensitive learning problems (Elkan 2001). However, it remains unclear, how the cost-sensitive approach changes once the additional requirement of fairness is introduced and how the concepts of prediction and decision interact in this process.

For studying the interaction of prediction and decision, we introduce a framework allowing us to distinguish the tasks and responsibilities of two different roles: The role of the 'prediction-modeler,' and the role of the 'decision-maker.' Following decision-theoretic concepts, we may think of two different agents, one being responsible for the prediction model and the other one being responsible for the decision-making. Our motivation for distinguishing these roles is not only fed by the theoretical analysis of how predictions are converted into (un)fair treatment as discussed above, but also by the observation that in practice these two roles are often split organizationally and covered by different people, different departments, or even different companies.[3] Under a perspective of responsibility, the decision-maker is responsible for the decisions, and hence their consequences. However, as we will discuss below, the prediction-modelers also have their area of responsibility. They are responsible for creating the basis for a good decision, which consists in (a) delivering a meaningful and robust prediction (e.g., think of transparency and safety requirements in High-Level Expert Group on Artificial Intelligence 2019), and (b) delivering all information needed for the decision-maker to care for fairness and other relevant ethical requirements (see accountability and fairness requirements in High-Level Expert Group on Artificial Intelligence 2019 and the obligations requested by European Commission 2021a).

These two roles have different tasks and often conflicting goals. On one hand, the prediction-modeler strives for prediction performance such as accuracy. This may be problematic when using ML models for consequential decision-making. For example, Athey (2019) argues that standard ML prediction algorithms, optimized for accuracy, are not sufficient to take decisions in complex settings as there are often

[1] Such systems may be implemented either in the form of Automated Decision-Making (ADM) systems or in the form of a combination of a prediction system with a human decision-maker.

[2] Note that other scenarios exist such as recommender systems where predictions are communicated to people who are taking decisions on themselves. In such cases, the findings of this paper are not directly applicable but they may inspire future work.

[3] Of course, in an integrated and fully automated data-based decision system, both agents may be combined into one function, but we think that it is conceptually useful to distinguish the two functions.

other relevant factors that are not represented in how well a model fits the training data. It also has shown that the fact that prediction-modelers usually have little specific knowledge of the domain in which an algorithm is applied may be problematic in consequential decision-making (Athey 2017). Similarly, Cabitza et al. (2021) argue that optimizing for accuracy is imperfect and that a larger spectrum of metrics and information should be considered to assess a system's performance. On the other hand, the decision-maker aims to optimize their benefit resulting from the decision-making (e.g., considering business-related goals). These observations clearly show that the goal of a prediction-modeler needs not to be consistent with the final goal of the decision system. The framework we propose addresses this tension. Specifying the prediction-modeler and the decision-maker as two different roles allows for separate performance measures. Furthermore, it allows a separation of the responsibilities of the actors which may also account for their domain-specific competencies.

Our framework starts from a decision-theoretic analysis, thus connecting to existing literature that conceptualizes fairness as a decision-theoretic problem (see for example Petersen 1976; Sawyer et al. 1976). However, instead of formulating fairness in terms of utility statements (e.g., see Corbett-Davies et al. (2017) for a similar approach in the contemporary debate), we encode fairness as constraints of a decision problem. This paper focuses on group fairness as the most established and most commonly used fairness category. This type of fairness intends to avoid systematic disadvantages of algorithmic decisions with respect to a sensitive attribute (such as gender, age, or race) (Binns 2020; Barocas et al. 2019). There are also other types of fairness [for example, counterfactual fairness (Kusner et al. 2017), individual fairness (Dwork et al. 2012), or procedural fairness definitions (Grgić-Hlača et al. 2018)] but these are not covered in the present paper.

This paper is structured as follows: In Sect. 2, we give a short review on the ML literature with respect to prediction-based decision-making, with a particular emphasis on group fairness metrics. In Sect. 3, we comment on the relationship between prediction and decision, which we then articulate and formalize in Sect. 4. In Sect. 5, we discuss some insights derived from our framework.

2 Fairness in the machine learning literature

2.1 Fairness of prediction-based decision-making systems

Prediction-based decision-making systems are increasingly used to assist (or replace) humans in making consequential decisions. Algorithms are used to inform or automatically take decisions in lending (Hardt et al. 2016; Fuster et al. 2017; Liu et al. 2018), pretrial detention (Angwin et al. 2016; Dieterich et al. 2016; Chouldechova 2017; Berk et al. 2021; Baumann et al. 2022) college admission (Kleinberg et al. 2018), hiring (Miller 2015a, b; Li et al. 2020; Raghavan et al. 2020), insurance (Baumann and Loi 2023), and many other fields. Recently, there has been a growing interest in the ethical implications of such prediction-based decisions, both from society and policymakers (European Commission 2021b). This has motivated the study of fairness in the field of ML, which has led to a newly formed community.[4]

Various factors can lead to algorithmic unfairness, such as a biased dataset, a systematic measurement error, the selection of a specific evaluation metric, or the taken modeling choices (Mitchell et al. 2021). Pursuing the goal of alleviating issues of algorithmic unfairness, researchers have proposed a plethora of fairness definitions (Narayanan 2018; Verma and Rubin 2018). In this paper, we focus on group fairness, a definition that has been of particular interest in the literature on fair ML (Pessach and Shmueli 2020). We now introduce the most common group fairness metrics before we describe how they can be ensured.

2.2 Measuring group fairness

We use A to denote the *sensitive attribute* (sometimes also referred to as *protected attribute*). Following related work, we consider binary group membership $A = \{0, 1\}$, but our arguments generalize multi-group situations. $\mathbf{X}$ denotes the observable attributes that are used for prediction,[5] while Y denotes the unknown but decision-relevant target variable. For the sake of simplicity, we assume Y to be a binary variable. We assume that there is *prediction function f* that maps instances of $\mathbf{X}$ to a prediction $\hat{Y} = f(\mathbf{X})$. The *decision function* is a (possibly group-specific) function $d(\hat{Y})$ or $d(\hat{Y}, A)$ that transforms the prediction $\hat{Y}$ into a decision D.

According to Barocas et al. (2019) and Kearns and Roth (2019), most of the existing group fairness criteria fall into one of three categories: independence, separation, or sufficiency. Due to ambiguities regarding the notion of a 'fair prediction' vs. that of a 'fair decision,' different notations are being used for the same fairness criterion. Those who apply the criteria to prediction models usually refer to the

[4] See for example: https://facctconference.org/ and special tracks in major conferences in the field (ICML, NeurIPS, ECML, etc.).

[5] Notice that $\mathbf{X}$ may or may not contain A. Not using the sensitive attribute as an input for the ML algorithm refers to a somewhat naive concept of fairness called *fairness through unawareness* (Grgić-Hlača et al. 2016), which does not effectively avoid disparate impact in case of redundant encodings (meaning that the sensitive attribute can be predicted by the remaining observable attributes, which is a likely scenario in the age of big data) (Pedreschi et al. 2008).

prediction $\hat{Y}$ (sometimes also expressed as a score, usually denoted by S or R), while those applying it to decision algorithms refer to the decision D for the same criteria (Hardt et al. 2016; Verma and Rubin 2018).[6] In this paper, we use the latter notation (which is also used by Verma and Rubin (2018) and Mitchell et al. (2021), for example) because it is in line with the framework we propose.

Independence (also called *statistical parity*, *demographic parity*, or *group fairness*) requires the decision to be independent of the sensitive attribute and is formally defined as:

$$P(D = d|A = 1) = P(D = d|A = 0). \tag{1}$$

Thus, the probability of a specific decision d must not depend on the group membership A. For the example of granting a loan, independence requires equal acceptance rates for both groups. *Conditional statistical parity* extends independence in that it allows a set of legitimate features L to affect the decision (Kamiran et al. 2013; Corbett-Davies et al. 2017):

$$P(D = d|L = l, A = 1) = P(D = d|L = l, A = 0). \tag{2}$$

For example, in the loan case, the applicant's requested credit amount could be a possible legitimate feature.

Separation (also called *equalized odds*) takes the individual's Y-value into account:

$$P(D = d|Y = y, A = 1) = P(D = d|Y = y, A = 0). \tag{3}$$

Thus, the requirement of the same probability of a decision d across groups is restricted to individuals with the same value of Y. Separation is equivalent to parity of true positive rates (TPR) and false positive rates (FPR) across groups $a \in A$. Another popular definition of fairness, *equality of opportunity*, is a relaxation of the separation constraint only requiring TPR parity.[7] In the loan granting scenario, this definition of fairness would ensure that "deserving individuals" (the ones who would repay the loan if given one, i.e., $Y = 1$) receive loans proportionately across groups.

In contrast, the fairness notion *sufficiency* conditions not on Y but on the decision D:

$$P(Y = y|D = d, A = 1) = P(Y = y|D = d, A = 0). \tag{4}$$

This means that, among all those individuals who receive the same decision d, the probability of a specific value y must not depend on A. For binary Y and D, sufficiency is equivalent to parity of positive predictive values (PPV) and false omission rates (FOR) across groups – meaning that for subgroups formed by D, an equal share of individuals must belong to the positive class $Y = 1$ across groups A (Baumann et al. 2022). The fairness definition PPV parity (also called *predictive parity* by Chouldechova (2017); Kasy and Abebe (2021)) relaxes sufficiency in that it only requires Y and A to be independent for all individuals who received a positive decision $D = 1$, which amounts to parity of PPV for binary classification (Baumann et al. 2022).[8]

Concluding, we see that all group fairness definitions are based on the equality of a specifically defined probability across the considered groups. Interestingly, the ML literature does not systematically relate the equality of these probabilities to philosophical concepts of social justice and fairness. However, it is beyond the scope of this paper to explore this question of the relation of the mathematical definition of fairness metrics and their moral meaning, even though this is still only rarely discussed, for example, in Hedden (2021), Loi et al. (2019), Baumann and Heitz (2022), Long (2021) and Hertweck et al. (2021). For the current paper, it suffices to state that measures of group fairness are typically based on the equality of conditional probabilities, which corresponds to the normative idea of 'equal shares' across the different groups.

2.3 Generating group fairness

The context-dependent nature of the fairness problems makes it impossible to agree on one universally applicable definition of group fairness. In addition, many fairness criteria are mathematically incompatible (Kleinberg et al. 2016; Chouldechova 2017; Garg et al. 2020; Friedler et al. 2021). This requires making a choice based on the concrete setting of the decision problem.

There are different techniques for ensuring the fairness of prediction-based decision systems, most of which fall into one of three categories (Mehrabi et al. 2019): *Pre-processing* involves manipulating the training data in order to generate a prediction model that leads to a fair decision (Calders and Verwer 2010; Kamiran and Calders 2012). *In-processing* involves integrating fairness requirements directly into the prediction model training itself (e.g., by incorporating a

[6] An example of such a discrepancy in notations is the group fairness metric called separation (which we introduce shortly) as defined in the algorithmic fairness literature. For example, Hardt et al. (2016) defines this notion of fairness for a sensitive attribute $A = \{0, 1\}$ as: $P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), y \in \{0, 1\}$. This definition implicitly assumes that a specific value of the prediction $\hat{Y}$ is equivalent to a specific decision. Others, such as Verma and Rubin (2018), define separation for a sensitive attribute $G = \{m, f\}$ and a decision d as follows: $P(d = 1|Y = i, G = m) = P(d = 1|Y = i, G = f), i \in 0, 1$.

[7] Similarly, FPR parity (also called *predictive equality* by Corbett-Davies et al. 2017) is a relaxation of separation that only conditions on $Y = 0$.

[8] Similarly, FOR parity is a relaxation of sufficiency that only considers individuals who received a negative decision $D = 0$ (Baumann et al. 2022).

fairness constraint for the training of a prediction model) (Kamishima et al. 2012; Bilal Zafar et al. 2017). The third category, *post-processing*, takes the prediction model as given and changes the decision function such that the resulting decisions meet some fairness constraints (e.g., by using group-specific thresholds on predicted scores) (Baumann et al. 2022; Hardt et al. 2016; Menon et al. 2018; Corbett-Davies et al. 2017). Despite their inherent advantages and disadvantages, all of these methods have been shown to be effective (Barocas et al. 2019).[9] Pre-processing and in-processing techniques place the burden of generating fairness on the prediction-modeler. Conceptually, this is only possible if the decision-maker is not an independent actor but instead implements a predefined decision rule applied to the output of the prediction model. In Sect. 3, we will point out that this is an unrealistic assumption in many cases. The much more frequent situation is that a prediction model may be used in different ways by an independent decision-maker, who, e.g., considers additional factors for the decision-making. This implies a focus on post-processing methods.

3 The relation between predictions and decisions

In popular narratives of algorithmic decision-making, the distinction between the idea of decision and that of prediction seems to be blurred and applied to the notion of fairness in a flexible way. Neologisms such as "fair prediction" Chouldechova (2017) or "fairness-aware learning" Kamishima et al. (2012) have become familiar within the ML community fueling, often unintentionally, the idea that fairness is a property of a prediction model. Even studies addressing algorithm-human interaction ultimately assimilate human decisions to a prediction task, e.g by comparing human estimates to algorithmic outcomes (Kleinberg et al. 2018; Dressel and Farid 2018; Vaccaro and Waldo 2019; Green and Chen 2019).

Actually, the conflation of the two concepts does not reflect an explicit ideological position and some studies clearly specify that fairness is an attribute that refers to a decision rule (Corbett-Davies and Goel 2018). However, formal characterizations tend to apply fairness criteria to the prediction model (e.g., the classifier) assuming that the decision consists of the prediction outcome (e.g., see Zafar et al. 2015; Menon et al. 2018; Berk et al. 2021).

Given similar formulations, one might naturally assume that the relation between prediction and decision is fixed and given, i.e., that a specific prediction leads to a specific decision. However, this is not true in many realistic examples, where the optimal decision depends on the prediction and other parameters (for example, in lending decisions, on the interest rate). This is in line with the idea of cost-sensitive learning (Elkan 2001). Thus, it is misleading to qualify a prediction as fair without explicitly assuming how a prediction is converted into a decision. In general, the fairness attribution applies more properly to the full system (i.e., the combination of prediction and decision rule) rather than to the prediction as such.

In a similar vein, Kuppler et al. 2022 distinguish between prediction and decision in data-driven decision procedures to highlight the different meanings and roles of fairness and justice. According to the authors, (algorithmic) fairness is concerned with the statistical properties of the prediction model, whereas justice is concerned with the allocation of goods and, therefore, more appropriately associated with the decision step. It is important to note that our approach differs as it builds on the idea that fairness is a concept related to the outcomes of decisions on people's lives. Therefore, we argue that fairness is a property of the entire system and that theories of distributive justice should be reflected in the chosen fairness definition. Instead of fully disentangling the desired properties of a prediction model from the decisions step, we argue that the prediction model's sole purpose is to inform decision-makers. This allows for building prediction-based decision-making systems around social fairness desiderata including theories of distributive justice that are morally appropriate for the context at hand.

3.1 Why a distinction is needed

Abstract formalization facilitates the overlap between the concepts of prediction and decision. For example, in classification tasks, the goal of prediction is to select an option among possible alternatives so that predicting can be viewed as a special form of deciding. Also in cases where the outcome to be predicted is a numerical value (e.g., a risk score), a prediction can be easily translated into a discrete scale (e.g., low - medium - high risk). From this standpoint, there is not much difference between the task performed by a prediction model and that performed by a decision-maker. However, if we go beyond mathematical abstractions and take an ethical stance, a decision is not just a matter of choosing among alternatives. It is a way to act and impinge upon humans and the environment. In other words, decisions change the *status quo*, thus bearing consequences for the decision-maker, the decision subjects, and possibly the external world. In contrast, a prediction, per se, has no impact, and its capacity to influence decision-making is made possible only by a policy or a decision rule. It is the latter that specifies the consequences of future actions.

[9] We point to Pessach and Shmueli (2020) and Caton and Haas (2020) for a more detailed discussion of the different unfairness-mitigation techniques including their (dis)advantages.

Consider, for example, a bank giving loans to individuals, building their decision of accepting a loan applicant on a predicted repayment likelihood. Granting a loan creates a tangible impact in the form of a benefit, consisting of improved financial flexibility and new buying options. This benefit is denied to loan applicants who receive a negative decision. Apparently, the prediction algorithm influences the decision, but the prediction itself is not what creates (un) fairness, it is the decision that specifies how to use the prediction estimate. Note that even if the decision is fully determined by the prediction—a case which is rarely met—the question of whether the prediction algorithm is fair or not is conditioned on the assumed relation between prediction and decision rule. This is why we conceptually suggest to clearly distinguish between the two elements of prediction and decision, which both are ingredients of any prediction-based decision system, whether it be fully automatic or also influenced by humans. Most importantly, the distinction between the two concepts invites us to contextualize algorithmic decision-making into a process of social construction reflecting value judgments and power asymmetries.

Often, apart from the prediction itself, the final outcome of a decision process is determined by additional pieces of information. Consider, again, a bank that needs to decide whom to grant and whom to deny a loan, where the bank's goal is to maximize their profit from the loan business. It is clear that the expected profit depends on the probability of repaying, which is why a prediction model for determining this probability is needed. However, other parameters, such as the interest rate charged for the loan, are also relevant, and these parameters obviously determine the minimum repaying probability that the bank will accept. A change of this threshold changes the decision rule and, thus, this represents a cost-sensitive learning problem (Elkan 2001). However, what is not covered in the literature on cost-sensitive learning is the fact that this often also changes the decision system's fairness properties. If the decision is fair for one threshold, this does not imply that it is fair for another threshold.

Another reason for marking a distinction between prediction and decision lies in the fact that the two concepts are benchmarked against different criteria. From a conceptual point of view, independent of the decision that may follow, a prediction can only be assessed in terms of its predictive power, e.g., accuracy. If one predicts, for example, the probability that a loan applicant will repay their debt, then a given prediction algorithm can be more or less accurate, typically evaluated with observation data, which are called "ground truth." It is, conversely, nonsensical to ask whether a decision is accurate or not since, broadly speaking, there is no such thing as "ground truth" in a decision process. A decision can be "right" or "wrong," but the same decision can be qualified differently depending on a variety of factors. We can judge the quality of a decision based on the purpose it aims to achieve and the consequences it has on the decision-maker and their surrounding environment (including other people), for example, in terms of fairness and accountability. As we will see, decision theory frames this intuition as an optimization problem so that an optimal decision is the one that maximizes a specific goal.

The problem of whether a prediction can be seen as unfair or not connects to a broader philosophical debate. In particular, this issue recalls the attempt to investigate the moral status of beliefs and thoughts going beyond the sphere of actions and deliberations. For example, advocates of epistemic injustice argue that people can commit injustice when they fail to believe someone's testimony due to prejudice (Fricker 2007), and theorists of doxastic wronging postulate that people can wrong others in virtue of what they believe about them, and not just in virtue of what they do (Basu 2019). Therefore, one may ask whether a (un)fair prediction could be regarded as an unjust or discriminatory belief. In this paper, we do not dig into this problem, which would require a separate discussion, and consider unjust beliefs on par with a decision rule operating, more or less consciously, in the decision-maker's mind.

3.2 Related studies on the interaction with prediction-based decisions

Arguing that the assessment of fairness requires a distinction between prediction and decision recalls a growing body of research focusing on how humans and algorithms interact when making decisions. These studies approached the interaction from different perspectives.

Some works aimed to show how algorithms can improve predictive performance (Kleinberg et al. 2018; Miller 2018) especially when there are carefully designed protocols of interaction (Cabitza et al. 2021). Others investigated people's perceptions shedding light on what has been called "algorithmic aversion" (Dietvorst et al. 2015), i.e., the situation in which a human decision-maker prefers human forecast over algorithmic prediction even when the latter is more accurate than the former. Further research pointed out that human decision-makers tend to deviate from algorithmic predictions (Stevenson and Doleac 2021) and struggle to assess algorithmic performance (Green and Chen 2019; Poursabzi-Sangdeh et al. 2021). Similar works suggested best design practices to allow designers to make adjustments in fully automated decision systems that interact with people (the so-called "street-level-algorithms") and make erroneous or unfair decisions when encountering a novel or marginal case (Alkhatib and Bernstein 2019).

Another area of research focuses on the challenge of autonomy in algorithmic decision-making. A key aspect of this work involves clarifying the meaning of autonomy when referring to algorithms, making distinctions between being "autonomous" and being "automatic" (Chiodo 2022; Pianca et al. 2022). Further concerns regard the constraints that algorithms may pose to decision-makers' agency (e.g., in relation to the algorithm's perceived authority) (Hayes et al. 2020) or the liability of algorithms causing injuries to humans or property damages (Barfield 2018).

Our work complements this broader literature and suggests new research directions exploring the interaction between the actors who deal with the prediction and the decision tasks. In this way, we aim at gaining a better understanding of the context of prediction-based decision systems, suggesting the fundamental social nature of systems' construction process and highlighting the informational gaps that must be addressed to improve the accountability of prediction-based decision-making systems.

4 A prediction-based decision system under fairness constraints

Intuitively, a decision is a termination of a process that involves several tasks. In general, a decision-maker identifies preferences, sets requirements and courses of action, analyses the pros and cons of each alternative, and chooses from the available options – the etymology of the term is quite explanatory: from Latin "de" = "off" + "caedere" = "cut."

The idea that distinct tasks are involved in human decision-making is well-entrenched in classical philosophy. For instance, medieval philosophers recognized three distinct operations (see e.g. Aquinas 2005; Saarinen 2006; Hain 2015, for a modern interpretation). The first, known as "consilium," consists of asking for advice and gathering relevant information for the decision at stake. The second operation involves judgment and constitutes more properly the deliberation step, also known as "resolution." The third operation regards the concrete actions that implement what was decided in the previous step.[10]

Our framework focuses on the first two operations of this deliberation scheme ("seeking advice" and "deciding"). Given the focus on decisions under uncertainty, it particularly emphasizes a specific type of advice generated by a prediction model. Additionally, our framework incorporates a specific ethical constraint on group fairness (as described in Sect. 2.2).

First, we introduce two formal roles, the role of the prediction-modeler and the role of the decision-maker. We show how these connect to different goals and tasks of prediction-based decision systems, and where fairness constraints (FC) come into play. Then, we describe the two roles in a more formal way specifying the parameters that characterize the tasks of both actors. Finally, we specify how these two interact and, in particular, define a minimum set of deliverables required to construct optimal decision rules meeting fairness constraints.

4.1 Two roles in prediction-based decisions

In our framework, we distinguish two roles that become particularly relevant in the discussion of responsibilities connected to a prediction-based decision system. The first role is the prediction-modeler, who is responsible for the prediction. The second one is the decision-maker who uses the prediction to optimize their own benefit (utility) while, possibly, also considering fairness issues by ensuring that certain fairness measures are met. Note that following the analogy with the aforementioned three-step model of decision-making, the role of prediction-modeler and that of decision-maker fulfill, respectively, the tasks of advising and deliberating.[11]

Usually, these two roles reflect different backgrounds and often different education. Typically, the role of the prediction-modeler is taken by data scientists, engineers, or computer scientists, while the role of the decision-maker can be played by various professionals, such as doctors, product managers, or business strategists. In the bank setting, the prediction-modeler may coincide with an external and independent organization (say, a software company) or an internal but separate department (e.g., the bank's data science lab), while bank managers play the role of the decision-maker. The source of the distinction between the two roles lies in the different goals they aim to achieve. While the goal of the prediction-modeler is to maximize the performance of a prediction model, such as accuracy, the goal of the decision-maker may vary depending on the context and includes, for instance, the increase of profit or the optimization of product development.

Our framework rests on the idea that even though the two roles are conceptually and practically distinct, but need to work in synergy for addressing fairness issues. Our framework specifies the tasks related to each role and, at the same time, the interaction points that allow the decision-maker to

[10] The conceptualization of prediction-based decision as a two-step process has another parallel with the philosophy of science, where a famous distinction regards the generation of new knowledge (the "context of discovery") and its assessment (the "context of justification"). In this work, we recalled the analogy with classical moral philosophy but we acknowledge that the parallelism with the pair "discovery-justification" would offer other important stimuli that would deserve e dedicated discussion.

[11] Here, we present a simplified characterization focused on two roles but more sophisticated descriptions could rely on multiagent system theory (Singh 1994)

adequately integrate fairness concerns into decision-making (see Sects. 4.2-4.4).

The decision-maker is the role directly involved in choosing which fairness metric to use (i.e., how unfairness is measured) and to what extent unfairness should be removed. These choices require the assessment of the social and the business context of the decision system. Typical questions to be answered are: Which subgroups should be considered with respect to fairness (i.e., what are the sensitive attributes)? Which fairness metric is the most appropriate in the given social context? What is the optimum trade-off between optimizing utility and enforcing fairness?

Answering these questions can be challenging, if not impossible, for the prediction-modeler whose task is predicting an unknown, but decision-relevant, quantity Y.[12] On the one hand, one may argue that, in principle, the prediction task should not involve caring for fairness-relevant issues: A good prediction is something else than fair treatment or a socially just distribution of benefits and harms. So, from a conceptual point of view, one may question whether assigning responsibility to the prediction-modeler makes any sense. On the other hand, from a practical perspective, the prediction-modeler is often simply not able to care for fairness because they do not have access to the needed contextual information and do not have the competence to decide on the normative issues involved. This makes clear why, both from a conceptual and a practical viewpoint, the two roles should be distinguished and why these are separated in most real-world cases.

Note that these roles are often left implicit in most ML fairness literature, where the common narrative of "fair ML" or "fair prediction models" would indirectly suggest that caring for fairness is a task of ML engineers. Our framework aims to be more specific than the standard approach in defining the roles and the minimum requirements associated with these roles in prediction-based decision-making. This will allow us to derive ethical responsibilities and support the implementation of fairness governance mechanisms in real-world scenarios.

In the following subsections, we will analyze the two roles and their interaction more closely, which will lay the ground for answering the question of who is responsible for what.

4.2 The decision-maker

Decision-making is a task that can be described in purely abstract terms. This is what decision theory does to frame a variety of decision problems ranging from what movie to watch in the evening to what career to pursue after college.

We consider a decision-theoretic agent[13] who makes decisions based on certain goals and preferences. In its simplest form, the agent chooses an action in a finite set of possible alternatives, and this action has an impact on the surrounding environment. To evaluate the impact, we consider the system's state after the agent's action and assign each possible state a specific value of the so-called *utility*. This refers to a quality that measures the desirability of this future state: the more desirable the state, the higher the utility. Thus, utility formalizes and quantifies the notion of a goal. It allows comparisons among different future states as a function of the chosen action which, in turn, allows one to choose among the different possible actions. In many cases, the relation between action and outcome (and thus utility) is not deterministic: The same action might lead to different outcomes, depending on factors that are not under the decision-maker's control. This situation is referred to as a "decision under uncertainty." It puts a decision-maker in a situation where they have to make a decision without really knowing what utility will be realized. In other words: The utility achieved following a decision is a random variable. Decision-making under uncertainty is about managing this uncertainty, while still trying to achieve a goal.

In the loan example, there are two possible future states or outcomes at the end of the loan contract: The loan plus associated interests may be paid back, or the debtor has defaulted, resulting in a loss of the loan. Obviously, for the bank, the former state is more desirable than the latter. The utility can be measured, e.g., by the amount of money that the bank has in their accounts by the end of the contract duration.

For applying this general decision-theoretic framework to the case of prediction-based decision systems, we identify the concept of "action" with that of "decision." The uncertainty of the outcome is usually attributed to the lack of knowledge of a random variable Y which might take different values y. Note that in real-world situations, many other factors might create uncertainty, but in the following, we analyze the simplest case in which Y is the only source of uncertainty. In the loan example, Y corresponds to the

[12] Note that the notion of a prediction in the context of an ML model is more encompassing than it is in colloquial language. While, in everyday language, we use the term "prediction" to refer to future situations (e.g., whether it will rain tomorrow), in the field of ML and statistics, a prediction simply relates to a fact that is not known when taking some action (e.g., "whether patient x has disease y" or "whether applicant z is trustworthy or not"). This lack of knowledge may be caused by different reasons, for example, due to missing information, but also when referring to an event in the future.

[13] We are aware that the decision-maker (also called the rational agent) assumed in economic theory is an idealization that might be far away from reality (i.e., humans can make irrational decisions for many reasons), but a decision-theoretic approach can also be a good starting point for modeling and analyzing decisions and their consequences.

repayment of the debt by the debtor, which decides which state is reached at the end of the loan contract. We also assume that the decision-maker takes not only one single decision but a sequence of many decisions of the same kind, which is a standard assumption for prediction-based decision systems. In the loan example, we envision a sequence of loan decisions of the bank, following the same decision rules for acceptance.

In such a situation, the goal achievement is measured as an expectation value, i.e., the decision-maker is interested in a decision rule which creates maximum utility in the long run, which means that they try to maximize the *expected utility* $E(U)$ as a function of their decision D:

$$E(U(D)) = \sum_{s} P(s|D) \cdot U(s|D) \tag{5}$$

where each state s represents a possible outcome as a function of the decision D, and $U(s|D)$ are the utilities associated with each outcome s. If Y is the only source of uncertainty, and thus determines the outcome, the different outcome states s correspond to the different values y of the random variable Y:

$$E(U(D)) = \sum_{y} P(Y = y|D) \cdot U(Y = y|D) \tag{6}$$

where now $U(Y = y|D)$ denotes the utility for the state reached in case of $Y = y$, and $P(Y = y|D)$ is the probability that this state is reached. Note that both elements may depend on the decision D.

For the sake of simplicity, in the following, we restrict ourselves to a binary variable Y, with two values $y = 0$ and $y = 1$. This gives:

$$\begin{aligned} E(U(D)) = & P(Y = 1|D) \cdot U(Y = 1|D) \\ & + (1 - P(Y = 1|D)) \cdot U(Y = 0|D) \end{aligned} \tag{7}$$

A decision-maker would be called rational if they choose the action that maximizes their expected utility (see the principle of Maximum Expected Utility (Russell and Norvig 2010)):

$$D = \arg\max E(U(D)) \tag{8}$$

For a simple loan example, the decision D is binary, with $D = 1$ corresponding to accepting the loan.[14] If we set the repaying probability $p = P(Y = 1)$, then this reads:

$$\begin{aligned} & E(U(D = 1)) = p \cdot \alpha - (1 - p) \cdot \beta \\ & E(U(D = 0)) = \gamma \end{aligned}$$

where α is the profit of the bank if the customer pays back, β is the loss if the customer defaults, and γ is the profit that can be made by not giving the loan, but instead investing the money into another business line of the bank.

The optimization problem with respect to the decision (see Eq. (8)) can easily be solved, leading to:

$$D = 1 \text{ if } p > \frac{\beta + \gamma}{\alpha + \beta}, D = 0 \text{ else.} \tag{9}$$

This example shows that the decision rule depends not only on the probability p, but also on other parameters (α, β, γ) which are independent on the prediction of Y (given by p), but still decision-relevant.[15] In line with cost-sensitive learning approaches (Elkan 2001), this exemplifies why the prediction alone does not solve the decision problem.

An important insight from this decision-theoretic analysis is that the decision-maker needs the probabilities $P(Y|D)$ to optimize their decisions, which directly leads to the necessity of a prediction model. In fact, the fundamental equation (6) is composed of two elements: the probabilities $P(Y|D)$, and the utilities $U(Y|D)$. The first element is the one that is related to the prediction task, and the second element is related to the decision context, as it implements the desirability of the different possible outcomes. Both elements are independent of each other.

Until now, we assumed that the decision-maker bases their decision strictly on maximizing their utility. The resulting optimum decision rule, given by the solution of Eq. (8), may or may not produce fairness issues. A decision-maker who also wants to consider fairness in their decision-making has to adopt their decision strategy such that the resulting decision fulfills the chosen fairness criterion. While many different ways of how to do this are conceivable, a natural way of extending Eq. (8) to a fairness-sensitive context is to impose a fairness constraint on the utility maximization:

$$\begin{aligned} & D = \arg\max E(U(D)) \\ & \quad \text{subject to } \textit{Fairness Condition FC} \end{aligned} \tag{10}$$

where *FC* is a condition of equality such as the ones mentioned in Sect. 2.2, or a relaxed version of them. From a formal decision-theoretic perspective, this is the optimal combination of the decision-maker's original goal and the additional consideration of fairness.

[14] Notice that decision rules are likely to be more complex in reality. For example, a bank could adjust the interest rate depending on the predicted repayment probability and the applicant's willingness to pay (only denying a loan in cases with a very high default probability). Our framework generalizes to more complex decision rules. However, for simplicity, we consider a binary case.

[15] We assume $\alpha + \beta \neq 0$, as otherwise there would be no need for a prediction in the first place. In the loan example, if $\alpha = -\beta$, the bank's profit would be the same for any value of D.

In the context of a post-processing approach for ensuring fairness, this constraint optimization problem has been solved in Hardt et al. (2016) (for the fairness metrics equalized odds, equality of opportunity, and predictive equality), in Corbett-Davies et al. (2017) (for the fairness metrics statistical parity and conditional statistical parity), and in Baumann et al. (2022) (for the fairness metrics sufficiency, predictive parity, and FOR parity).

4.3 The prediction-modeler

As illustrated in the preceding subsection, the decision-making process requires the probabilities $P(Y|D)$ to solve problems that involve an unknown quantity Y. In the context of machine learning, this corresponds to making a probabilistic prediction of Y that the decision D might depend on. This represents the prediction task that the prediction-modeler is expected to address.

Interestingly, this does not include all versions of prediction models used in ML and discussed in the context of fairness. In particular, a point estimator $\hat{Y}$ with two possible values $\hat{Y} = \{0, 1\}$ is of little use for the decision-maker, as this does not allow to solve the decision problem stated in Eq. (8). Consider, for example, the optimum solution (see Eq. (9)) for $p \gg 0.5$ and for realistic parameters α and β: a typical ML prediction model optimized for maximum accuracy (threshold $p = 0.5$) would lead to far too many instances of $\hat{Y} = 1$ and thus $D = 1$. In general, we can conclude that a prediction-modeler who does not have access to the external parameters α, β, γ is not able to deliver a good point estimator $\hat{Y}$, which allows solving the decision-maker's task.

A typical assumption in the ML (fairness) literature is that the decision is determined by the value of Y (see Murphy 2012; Hardt et al. 2016; Mitchell et al. 2021), e.g., such that $Y = 1$ implies $D = 1$, and vice versa. This means that if only Y can be predicted with high accuracy, then the decision D will be correct. Sometimes it might be possible to achieve a perfect prediction, e.g., in the case of picture recognition. Here, the fact that a picture represents a dog instead of a cat is evidence that could be checked at the time of decision-making (or recognition), even if an ML classifier does not predict the image correctly. However, this is not the case in many decision problems discussed in the algorithmic fairness literature. For instance, for the loan example, there is real uncertainty about the repayment: Y is a random variable whose value cannot be predicted deterministically, and even the best prediction model cannot rule out this uncertainty. Similarly, in the COMPAS case, the fact of re-offending cannot be seen as a deterministic property of a delinquent. In all such cases, point estimators $\hat{Y}$ do not deliver useful information, and the only way of dealing with the uncertainty of the underlying situation is to use probabilities. This is reflected by Eq. (6).

Thus, from a decision-theoretical perspective, the basic task of the prediction-modeler is not to deliver a point estimate, but a probability (even if there are cases where a point estimator may be useful), i.e., the required prediction model is a probabilistic prediction model. The ML task then consists of deriving an estimate $\hat{p}$ of the true probability, based on the analysis of historical data $\{\mathbf{x}_i, i = 1, \ldots n\}$, by specifying a function f with $\hat{p} = f(\mathbf{x})$. Since f is determined from training data, it is prone to errors, and the resulting $\hat{p}$ is not identical to the real p. The goal of the prediction-modeler is thus to create a probability estimate which is as close as possible to the real p, as any deviation will lead to non-optimum decisions if the decision-maker uses the estimator $\hat{p}$ instead of the (unknown) true probability.

If the decision rule is assumed to be given, this requirement can be somewhat relaxed: strictly speaking, the requirement is that $\hat{p}$ leads to the same decisions as the true probabilities p. For example, in the loan context recalled in the last subsection, errors in $\hat{p}$ far away from the threshold specified in Eq. (9) would not make any difference. Thus, in general, our framework is agnostic to the type of prediction model used. However, in all cases where the decision-making is not fully specified from the beginning, or the prediction-modeler does not have full access to all decision-relevant parameters, or the value of the decision-relevant parameters might change over time, the prediction-modeler has to care for generating a prediction model that works over the full range of p.

4.4 The interaction

In this subsection, we analyze in more detail the interaction between the prediction-modeler and the decision-maker during the creation of a prediction-based decision system. Figure 1 illustrates this from a business process perspective. In addition to the specific activities performed, we visualize the flow of information between the two roles required for developing a prediction-based decision system, focusing on the minimum interaction required between the two actors. The objective is to identify a minimal set of deliverables for this interaction and examine how the introduction of a fairness constraint affects these deliverables.

Table 1 lists the minimum deliverables, i.e., the information that the decision-maker and the prediction-modeler must provide to each other, while Fig. 1 visualizes *when* during the sequence of tasks these deliverables are due. The decision-maker has to specify the prediction task according to the decision problem at stake. The prediction-modeler has to deliver a prediction model with associated additional information, such that the decision-maker can integrate it into the decision procedure. The required deliverables vary based on whether the decision-maker considers fairness requirements, resulting in two different scenarios for both

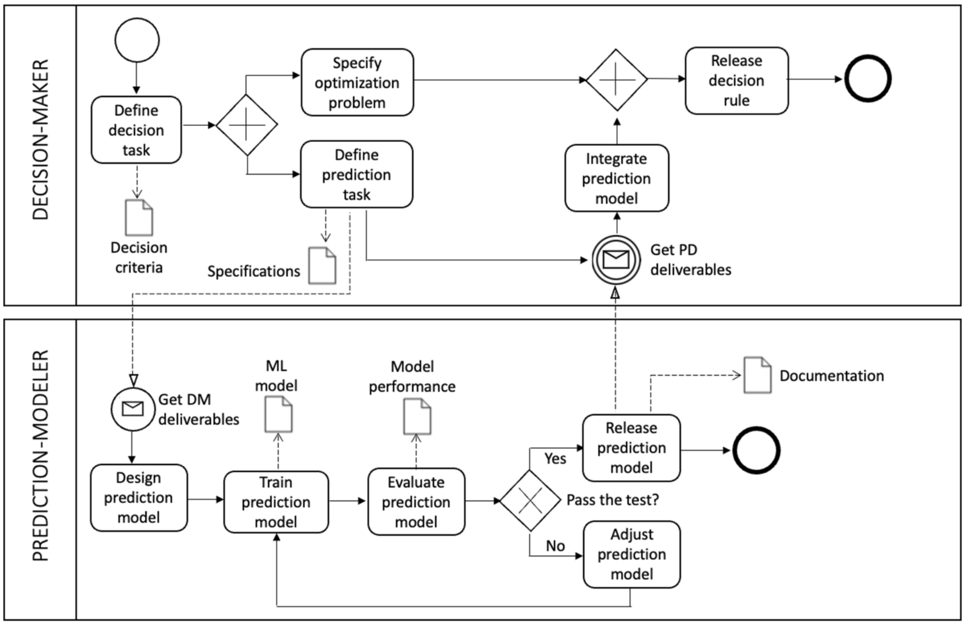

Fig. 1 A BPMN (Business Process Model and Notation) diagram of the tasks involved to generate a prediction-based decision system

Table 1 Sets of minimum deliverables by role

	DM → PM	PM → DM
Unconstrained utility maximization	• Target variable Y	• Prediction model $\hat{p} = f(\mathbf{x})$ • Prediction model performance • Calibration function
utility maximization s.t. FC	• Target variable Y • sensitive attribute A	• Prediction model $\hat{p} = f(\mathbf{x})$ • Prediction model performance • Group-specific calibration functions • Group-specific baseline distributions

DM → PM stands for the minimum set of deliverables the decision-maker (DM) must provide to the prediction-modeler (PM) and **PM → DM** describes the minimum set of deliverables the PM must provide to the DM

roles, as depicted in Table 1: The row described by *unconstrained utility maximization* refers to a decision-maker that bases their decision strictly on maximizing their utility without considering fairness (as formalized in Eq. (8)). In contrast, the row *utility maximization s.t. FC* refers to a decision-maker who also considers fairness, i.e., who adds a fairness constraint to the optimization problem, as is described in Eq. (10). In the following, we will comment on and justify the elements of this table.

4.4.1 Unconstrained utility maximization: DM → PM

At a minimum, defining the prediction task involves specifying the unknown variable Y to be predicted. In practice, additional specification elements such as the considered population or input features are given, which we omit for the sake of simplicity.[16]

4.4.2 Unconstrained utility maximization: PM → DM

Among the information provided by the prediction-modeler to the decision-maker, the prediction model is arguably the most important deliverable. In addition to that, the prediction-modeler also needs to communicate the performance of the model. This is necessary for the decision-maker to assess whether the model fits the domain-specific requirements, i.e., to evaluate if the model should be included in the decision procedure or not.

The decision-maker needs the probabilities p to be able to derive the optimal decision rule (see Eq. (9)). However, many ML models deliver uncalibrated scores instead of an estimate of the probability. To fulfill the needs of the decision-maker, the prediction model needs to be calibrated, delivering an estimate $\hat{p}$ of p. If this is not the case, the prediction model should come with a calibration function that allows the decision-maker to reconstruct the probabilities from the score. Note that calibration defines "a property of

[16] In certain situations, the prediction-modeler might receive training data from the decision-maker (for example, if it makes sense to use the decision-maker's customer data for training). However, in other cases, the prediction-modeler relies on external data sources to develop a model predicting the target variable Y, as specified by the decision-maker. Therefore, our general framework does not foresee the provision of training data by the decision-maker as a minimum deliverable.

the model [more] than of its use since it does not depend on decision thresholds" (Hutchinson and Mitchell 2019, p. 55).

4.4.3 Utility maximization s.t. FC: DM → PM

Consider now the minimum deliverables of the decision-maker in the constrained case, that is, when the decision-maker optimizes their utility subject to some group fairness constraint. Recall that the basic idea of group fairness is to avoid unjustified disadvantages for subgroups of the population, defined by a sensitive attribute A (see Sect. 2.2). The specification of the regarded sensitive attributes is done by the decision-maker. Only with knowledge of the protected subgroups considered, the prediction-modeler, in turn, can transmit the minimum deliverables assigned to them.

4.4.4 Utility maximization s.t. FC: PM → DM

To facilitate the decision-maker's ability to solve the constrained optimization problem outlined in Eq. (10), the prediction-modeler must provide additional information. The type of information might depend on the fairness constraint, and while the problem has been studied for some cases of fairness constraints, the ML literature still has many unexplored areas. In the following, we restrict the discussion to the group fairness metrics that have been studied so far, relating to Hardt et al. (2016), Corbett-Davies et al. (2017), and Baumann et al. (2022).[17]Hardt et al. (2016) and Corbett-Davies et al. (2017) prove that any optimal decision rule d^* that satisfies statistical parity, conditional statistical parity, equality of opportunity, or predictive equality takes the following form of group-specific thresholds, i.e.:

$$d^* = \begin{cases} 1 & p \geq \tau_a \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

where $\tau_a \in [0, 1]$ denote different group-specific constants.[18]Baumann et al. (2022) prove that any optimal decision rule d^* that satisfies predictive parity or false omission rate (FOR) parity takes the following form of group-specific upper- or lower-bound thresholds, i.e.:

$$d^* = \begin{cases} \left.\begin{matrix} 1, \text{ for } p \geq \tau_a \\ 0, \text{ otherwise} \end{matrix}\right\} \text{for } v > P(Y = 1|A = a) \\ \left.\begin{matrix} 1, \text{ for } p \leq \tau_a \\ 0, \text{ otherwise} \end{matrix}\right\} \text{for } v < P(Y = 1|A = a) \end{cases} \tag{12}$$

where v denotes the positive predictive value for the predictive parity fairness constraint – the false omission rate in the case of FOR parity, respectively. $P(Y = 1|A = a)$ denotes the prevalence of group a (also called base rate), which is defined as the share of individuals belonging to the positive class.

The fairness constraint transforms into a condition relating the thresholds τ_0 and τ_1, where the exact form of this relation depends on the chosen fairness constraint. The decision-maker's utility is maximized by selecting the optimum one from all pairs (τ_0, τ_1) defined by this relation, based on the resulting utility. To evaluate the utility, the distributions of p for both groups are needed. This so-called "baseline distribution" describes how each subgroup is distributed over the probability range $p \in [0, 1]$ (for details of the determination of optimum thresholds see Hardt et al. 2016; Corbett-Davies et al. 2017; Baumann et al. 2022). For a given prediction model, the baseline distributions can be determined, at least approximately, from the training data, and this information has to be delivered to the decision-maker as a necessary element for their decision-making. Also, the utility evaluation can only be done if the calibration requirements are met on the level of the subgroups, so either the prediction model must be calibrated separately for each considered subgroup, or group-specific calibration functions need to be provided.

Thus, the fact that the decision-maker is considering fairness constraints leads to additional information requirements from the side of the prediction-modeler. Recall that we have restricted the discussion to the case of a few already studied group fairness criteria, for which we end up with the specification of deliverables in Table 1. For other fairness constraints, the additional requirements may be different. However, as a general rule, we might expect that imposing fairness constraints for the decision system generates additional information requirements that the prediction-modeler must meet. Simply delivering a black-box prediction model without this additional information is, in general, not sufficient for enabling the decision-maker to ensure a fair decision system. In Sect. 5, we will analyze the ethical consequences of this.

Note that the discussed examples in this subsection relate to the so-called post-processing methods for creating fairness (Mehrabi et al. 2019), assuming that a decision-maker accepts the prediction model as given. This is the simplest situation with minimum interaction between the two players.

[17] Hardt et al. (2016) use a concept they call *immediate utility* whereas (Corbett-Davies et al. 2017) use the concept of loss minimization for their proofs. Both of these concepts can easily be translated to what we call the decision-maker's utility. Therefore, their solutions hold for the constrained optimization problem, as we defined it. The problem formalization of Baumann et al. (2022) is in line with the constrained optimization problem, as we defined it.

[18] Note that when choosing conditional statistical parity as the FC, these constants additionally depend on the "legitimate" attributes. Furthermore, for the fairness criteria that combine two parity constraints (equalized odds and sufficiency), some randomization is needed (Hardt et al. 2016; Baumann et al. 2022). For simplicity, we omit this for the rest of the discussion.

However, our framework (as presented in Fig. 1) also holds in cases where pre-processing or in-processing methods are applied. In such cases, the interaction between the two roles is more complicated, as the decision-maker has to inform the prediction-modeler about the fairness constraint and, at least for in-processing methods, specify the decision rule upfront. Thus, the task of generating fairness can be shifted to the prediction-modeler, but at the expense that the decision-maker restricts their freedom to change the decision rule after the prediction model is delivered. Thus, pre-processing and in-processing approaches require a closer collaboration of the two roles, with associated increased requirements for the interaction between the two roles.

5 Discussion

The first important insight is that different actors come with different responsibilities. Here, we focus more specifically on professional responsibility, that is, the set of obligations based on a role played in a certain context.[19] Since our analysis relates to a well-defined problem in algorithmic decision-making (i.e., group fairness), these obligations translate into specific pieces of information that each role is expected to deliver.

The deliverables we suggested are not optional and reflect the strong interdependence between roles. Ultimately, we acknowledge that the responsibility for fair decisions falls on the role of the decision-maker for the reasons we already discussed in Sect. 4. However, their ability to address fairness issues depends heavily on the work of the prediction-modeler. Similarly, the latter cannot take responsibility for group-specific calibration functions and baseline distributions if they do not receive information about the sensitive attributes to be considered.

The interdependence between roles recalls the problem of creating meaningful communication channels among designers, managers, and, more generally, all professionals involved in designing and using artificial intelligence (AI) systems. To this aim, we offer some considerations that might be useful to inform future research and the implementation of prediction-based decision-making systems.

So far most of the literature on algorithmic fairness underestimated practical issues emerging in real-world organizations (see Holstein et al. (2019) for a notable counterexample), but to provide effective and sustainable solutions we need to fill the gap between mathematical abstractions and organizational dynamics and engineering practices (Tubella et al. 2022). Scholars have already addressed the risks of abstracting from the social context of AI applications and highlighted the need to reorient technical work away from solutions to process (Selbst et al. 2019; Scantamburlo 2021). Our framework goes in that direction and tries to figure out which kind of concrete interactions would help the implementation of (group) fairness starting from two key roles and their associated tasks.

Starting from professional roles gives the opportunity to distill important information entering prediction and decision tasks and directs greater attention to organizational aspects, which are often less regarded in the field of AI ethics. In general, analyzing roles and their interactions can reveal the background of values and assumptions that shape the design process (Krijger 2021). This role-based perspective may also serve to highlight a more articulated view of the design and use of algorithmic decision-making systems, where more than one professional might be involved. Usually, the study of human-AI interaction focuses on the exchange occurring between the (end) users and the operating systems. However, our framework suggests that there are other meaningful interactions that are worthy of consideration. An analysis of interactions shaping the design and use of AI systems may reveal conceptual gaps, structural deficiencies, and power imbalances.

Our exercise considers a simple business process scenario, but other elaborations are possible (a finer-grained analysis of tasks in different settings, e.g., medicine). For example, further research might explore connections with existing frameworks that emphasize the context-sensitive nature of computing systems, such as the model of contextual integrity (Nissenbaum 2010). A closer look at the norms and social practices that control, manage, and steer the flow of information within organizations can help gain a richer understanding of prediction-based decision systems. This may result in a description of the flows of information characterizing the context of a prediction-based decision system and the identification of which flows are appropriate to ensure agents (and the organizations) meet established goals and ethical norms.

For supporting interactions among professionals, an essential task is to keep track of relevant information characterizing the elements of the algorithmic decision-making system. In computer science and engineering disciplines, this goal is often fulfilled by devising software documentation that may include a variety of information (e.g., technical requirements, software architecture, codes, etc). Note that documentation is also acknowledged as an important measure to ensure transparency and accountability of AI systems. In this regard, the European Commission's proposal for an AI regulation requires that "technical documentation of a high-risk AI system shall be drawn up before that system is placed on the market or put into service and shall be kept up-to-date" (article 11 (European

[19] Responsibility, of course, extends beyond roles, and for a broader discussion see Van de Poel and Royakkers (2011).

Commission 2021a)). However, it is still unclear which types of information should form a robust documentation. This effort, moreover, should also consider how to make information accessible and useful for the players contributing to the informational exchange. This would require addressing issues of knowledge and language divides which often characterize participatory design processes.

With respect to the documentation task, the results suggested by our framework have a limited scope in that they refer to a particular setting (i.e., group fairness in algorithmic decision-making). However, our results show that, in general, it is necessary to perform an analysis of which specific contents one may need to address the problem at stake. Our effort suggests that, in general, players might have to ask for more specific information due to the context of use and the ethical issues addressed. Also, our attempt shows that it is necessary to address the question of how to modulate the creation and maintenance of documentation among different players. So far, technical documentation is often conceived as a task entirely in charge of engineers and computer scientists, but, in reality, there might be other roles affecting the design and the deployment of AI and ML systems. So we might think of reporting and documenting more as a collaborative practice that involves different roles rather than a duty assigned to a single category of people.

A final consideration regards human oversight, an ethical principle recommending human agency in AI-driven decision processes to ensure human autonomy and prevent adverse effects. While the notion of human-in-the-loop can inspire the structuring of human intervention and monitoring, it is open to discussion what type of duties and actions would be needed in real-world scenarios: What does it mean to intervene in a decision cycle? Who should do it? The intuition of identifying roles and the associated tasks is a way to start answering such questions. This would be particularly beneficial because in real-world decision-making procedures (such as those embedded in administrations or bureaucratic processes) responsibility is often delegated and distributed across multiple actors (Strandburg 2021). In our framework, we envision activities and interactions based on a simplified Business Model Notation, but richer elaboration can provide more details on who supervises what.

The creation of a flow of information between the prediction-modeler and the decision-maker connects to key ethical requirements in the design and deployment of AI and ML systems: Transparency, accountability, and human oversight. The scientific community and policy makers largely acknowledge the centrality of these requirements. However, there is still limited knowledge and experience on translating these requirements into practice. The approach our framework suggests offers meaningful stimuli to articulate these requirements more concretely and raises points that can move the community towards new research and policy directions.

6 Conclusions

In this paper, we argue that a prediction model as such cannot be qualified as fair or unfair. This argument is based on two observations: First, predictions themselves have no direct impact. Second, predictions can be used differently for making decisions. Important examples for the second observation are all post-processing methods to implement fairness constraints, e.g., Baumann et al. (2022), Hardt et al. (2016) and Corbett-Davies et al. (2017). These methods are based on the idea that the fairness properties of a decision system can be shaped by the way in which the prediction model's output is transformed into a decision, e.g., by imposing group-dependent decision thresholds. So, the same prediction model can lead to unfairness (without post-processing) or fairness (with adequate post-processing). Other examples are all human-in-the-loop approaches that combine prediction models with human decision-makers. They assume that humans are at least co-creators of the resulting ethical consequences of prediction-based decision systems, which of course implies that different ways of using the prediction model's output are conceivable and that the activity of the human in the loop consists exactly in influencing the usage of the prediction model's output.

Our framework serves as a tool to identify what is essential for each role in addressing fairness within a prediction-based decision system. It enables us to propose a minimum level of active responsibility (Van de Poel and Royakkers 2011) that one could demand from these roles in similar situations. The identification of deliverables and interactions is not meant to limit the responsibility of ML developers and decision-makers to the delivery of specific pieces of information, but to avoid false or too vague expectations of the obligations for the roles involved. Indeed, a deeper comprehension of the various roles, along with their goals, tasks, and responsibilities, is a crucial first step for implementing ethical requirements in prediction-based decision systems.

Funding Open access funding provided by University of Zurich. TS acknowledges support from the H2020 project AI4EU GA 825619 and from the IRIS Academic Research Coalition (UK government, GA SCH-00001-3391). JB and CH were supported by the National Research Programme "Digital Transformation" (NRP 77) of the Swiss National Science Foundation (SNSF) – Grant Number 187473 – and by Innosuisse – Grant Number 44692.1 IP-SBM.

Data availability Not applicable.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

References

Aler Tubella A, Barsotti F, Koçer RG, Mendez JA (2022) Ethical implications of fairness interventions: what might be hidden behind engineering choices? Ethics Inf Technol 24(1):1–11

Alkhatib A, Bernstein M (2019) Street-level algorithms: A theory at the gaps between policy and decisions. In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1–13

Angwin J, Larson J, Mattu S, Kirchner L (2016) Machine bias. ProPublica 23(2016):139–159

Aquinas T (2005) Summa theologiae. I-II q 14, a. 3

Athey Susan (2017) Beyond prediction: using big data for policy problems. Science 355(6324):483–485

Athey S (2019) The impact of machine learning on economics. In: Agrawal A, Gans J, Goldfarb A (eds) The economics of artificial intelligence: an agenda, chapter the impact. University of Chicago Press, pp 507–552

Barfield Woodrow (2018) Liability for autonomous and artificially intelligent robots. Paladyn J Behav Robot 9(1):193–203

Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning. fairmlbook.org

Barocas S, Selbst AD (2016) Big data's disparate impact. Calif Law Rev 104(3):671–732

Basu Rima (2019) The wrongs of racist beliefs. Philos Stud 176(9):2497–2515

Baumann J, Hannák A, Heitz C (2022) Enforcing group fairness in algorithmic decision making: utility maximization under sufficiency. In: 2022 ACM conference on fairness, accountability, and transparency, FAccT '22, New York, NY, USA. Association for Computing Machinery, pp 2315–2326

Baumann J, Heitz C (2022) Group fairness in prediction-based decision making: from moral assessment to implementation. In: 2022 9th Swiss Conference on Data Science (SDS), pp 19–25

Baumann J, Loi M (2023) Fairness and risk: an ethical argument for a group fairness definition insurers can use. Philos Technol 36:45

Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2021) Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res 50(1):3–44

Bilal Zafar M, Valera I, Gomez Rodriguez M, Gummadi KP (2017) Fairness constraints: mechanisms for fair classification. In: Artificial intelligence and statistics, pp 962–970

Binns R (2018) Fairness in machine learning: lessons from political philosophy. Technical report, 1

Binns R (2020) On the apparent conflict between individual and group fairness. In: FAT* 2020—Proceedings of the 2020 conference on fairness, accountability, and transparency, New York, NY, USA, 1. Association for Computing Machinery, Inc, pp 514–524

Cabitza F, Campagner A, Datteri E (2021) To err is (only) human. Reflections on how to move from accuracy to trust for medical AI. In: Ceci F, Prencipe A, Spagnoletti P (eds) Exploring innovation in a digital world. Springer, Cham, pp 36–49

Cabitza Federico, Campagner Andrea, Sconfienza Luca Maria (2021) Studying human-AI collaboration protocols: the case of the Kasparov's law in radiological double reading. Health Inf Sci Syst 9(1):1–20

Calders T, Verwer S (2010) Three naive bayes approaches for discrimination-free classification. Data Min Knowl Discov 21(2):277–292

Caton S, Haas C (2023) Fairness in machine learning: a survey. ACM Comput. Surv.http://arxiv.org/abs/2010.04053

Chiodo S (2022) Human autonomy, technological automation (and reverse). AI Soc 37:39–48

Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness. J Mach Learn Res, vol. 24.

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, KDD '17, New York, NY, USA. Association for Computing Machinery, pp 797–806

Dieterich W, Mendoza C, Brennan T (2016) COMPAS risk scales: demonstrating accuracy equity and predictive parity. Technical report, Northpoint Inc

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144(1):114

Dressel J, Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. Sci Adv 4(1):eaao5580

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: ITCS 2012—Innovations in Theoretical Computer Science Conference, New York, New York, USA. ACM Press, pp 214–226

Elkan C (2001) The foundations of cost-sensitive learning. In: Proceedings of the 17th international joint conference on artificial intelligence—vol 2, IJCAI'01, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc, pp 973–978

European Commission (2021a) Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on AI and amending certain union legislative acts. Technical report, Brussels

European Commission (2021b) Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Technical report, Brussels

Fricker M (2007) Epistemic injustice: power and the ethics of knowing. Oxford University Press

Friedler SA, Scheidegger C, Venkatasubramanian S (2021) The (im) possibility of fairness. Commun ACM 64(4):136–143

Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A (2017) Predictably unequal? The effects of machine learning on credit markets. SSRN, 11

Garg P, Villasenor J, Foggo V (2020) Fairness metrics: a comparative analysis. In: 2020 IEEE international conference on big data (big data). IEEE, pp 3662–3666

Green B, Chen Y (2019) Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments. In: Proceedings of the conference on fairness, accountability, and transparency, pp 90–99

Green B, Chen Y (2019) The principles and limits of algorithm-in-the-loop decision making. In: Proceedings of the ACM on human–computer interaction, vol 3(CSCW), pp 1–24

Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A (2016) The case for process fairness in learning: Feature selection for fair decision making. In: NIPS symposium on machine learning and the law, vol 1, p 2

Grgić-Hlača N, Zafar MB, Gummadi KP, Weller A (2018) Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 32, no 1

Hain R (2015) Consilium and the foundations of ethics. The Thomist Specul Q Rev 79(1):43–74

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Proceedings of the 30th international conference on neural information processing systems, NIPS'16, Red Hook, NY, USA. Curran Associates Inc, pp 3323–3331

Hayes P, Van De Poel I, Steen M (2020) Algorithms and values in justice and security. AI Soc 35:533–555

Hedden Brian (2021) On statistical criteria of algorithmic fairness. Philos Public Affairs 49(2):209–231

Hertweck C, Heitz C, Loi M (2021) On the moral justification of statistical parity. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, FAccT '21, New York, NY, USA. Association for Computing Machinery, pp 747–757

High-Level Expert Group on Artificial Intelligence (2019) Ethics guidelines for trustworthy AI. Technical report, European Commission, Brussles

Holstein K, Wortman Vaughan J, Daumé III H, Dudik M, Wallach H (2019) Improving fairness in machine learning systems: what do industry practitioners need? In: Proceedings of the 2019 CHI conference on human factors in computing systems, pp 1–16

Hutchinson B, Mitchell M (2019) 50 years of test (un)fairness: lessons for machine learning. In: FAT* 2019—Proceedings of the 2019 conference on fairness, accountability, and transparency, New York, NY, USA, 1. Association for Computing Machinery, Inc, pp 49–58

Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33(1):1–33

Kamiran F, Žliobaitė I, Calders T (2013) Quantifying explainable discrimination and removing illegal discrimination in automated decision making. Knowl Inf Syst 35(3):613–644

Kamishima T, Akaho S, Asoh H, Sakuma J (2012) Fairness-aware classifier with prejudice remover regularizer. In: Flach PA, Bie TD, Cristianini N (eds) Machine learning and knowledge discovery in databases. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 35–50

Kasy M, Abebe R (2021) Fairness, equality, and power in algorithmic decision-making. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, vol 11, New York, NY, USA, 3. ACM, pp 576–586

Kearns Michael, Roth Aaron (2019) The ethical algorithm: the science of socially aware algorithm design. Oxford University Press, Inc., USA

Kleinberg Jon, Lakkaraju Himabindu, Leskovec Jure, Ludwig Jens, Mullainathan Sendhil (2018) Human decisions and machine predictions. Q J Econ 133(1):237–293

Kleinberg J, Ludwig J, Mullainathan S, Rambachan A (2018) Algorithmic fairness. AEA Pap Proc 108:22–27

Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. In proceedings of innovations in theoretical computer science (pp. 43:1–43:23).http://arxiv.org/abs/1609.05807

Krijger J (2021) Enter the metrics: critical theory and organizational operationalization of AI ethics. AI Soc 37:1427–1437

Kuppler M, Kern C, Bach R, Kreuter F (2022) From fair predictions to just decisions? Conceptualizing algorithmic fairness and distributive justice in the context of automated decision-making. Front Sociol 7:883999

Kusner MJ, Loftus J, Russell C, Silva R (2017) Counterfactual fairness. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc.

Li D, Raymond LR, Bergman P (2020) Hiring as exploration. Available at SSRN: https://ssrn.com/abstract=3630630

Liu LT, Dean S, Rolf E, Simchowitz M, Hardt M (2018) Delayed impact of fair machine learning. In: IJCAI international joint conference on artificial intelligence, 2019-August, pp 6196–6200

Loi M, Herlitz A, Heidari H (2019) A philosophical theory of fairness for prediction-based decisions. SSRN Electron J

Long R (2021) Fairness in machine learning: against false positive rate equality as a measure of fairness. J Moral Philos 19:47–78

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35

Menon AK, Williamson RC (2018) The cost of fairness in binary classification. In: Friedler SA, Wilson C (eds) Proceedings of the 1st conference on fairness, accountability and transparency, volume 81 of proceedings of machine learning research, New York, NY, USA. PMLR, pp 107–118

Miller AP (2018) Want less-biased decisions? USE algorithms. Harvard Bus Rev, 26

Miller CC (2015a) Can an algorithm hire better than a human?

Miller CC (2015b) When algorithms discriminate

Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: choices, assumptions, and definitions. Annu Rev Stat Appl 8:141–163

Mulligan DK, Kroll JA, Kohli N, Wong RY (2019) This thing called fairness: disciplinary confusion realizing a value in technology. In: Proceedings of the ACM on human–computer interaction, vol 3(CSCW), pp 1–36

Murphy KP (2012) Machine learning: a probabilistic perspective. MIT Press

Narayanan A (2018) Translation tutorial: 21 fairness definitions and their politics. In: Proc. Conf. Fairness Accountability Transp, New York, USA

Nissenbaum H (2010) Privacy in context: technology, policy, and the integrity of social life. Stanford University Press

Pedreschi D, Ruggieri S, Turini F (2008) Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '08, New York, NY, USA. Association for Computing Machinery, pp 560–568

Pessach D, Shmueli E (2020) Algorithmic fairness. ACM Comput Surv 55(3) Article 51

Petersen Nancy S (1976) An expected utility model for "optimal'' selection. J Educ Stat 1(4):333–358

Pianca F, Santucci VG (2022) Interdependence as the key for an ethical artificial autonomy. AI Soc 38:2045–2059

Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JWW, Wallach H (2021) Manipulating and measuring model interpretability. In: Proceedings of the 2021 CHI conference on human factors in computing systems, pp 1–52

Raghavan M, Barocas S, Kleinberg J, Levy K (2020) Mitigating bias in algorithmic hiring: evaluating claims and practices. In: FAT* 2020—Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 469–481

Russell SJ, Norvig P (2010) Artificial intelligence a modern approach. Pearson Education, Inc

Saarinen R (2006) Weakness of will: philosophical and theological theories of action. In: Intellect et imagination dans la philosophie medievale

Sawyer Richard L, Cole Nancy S, Cole James W L (1976) Utilities and the issue of fairness in a decision theoretic model for selection. J Educ Meas 13(1):59–76

Scantamburlo T (2021) Non-empirical problems in fair machine learning. Ethics Inf Technol 23(4):703–712

Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the conference on fairness, accountability, and transparency, pp 59–68

Singh MP (1994) Multiagent systems. Springer

Stevenson MT, Doleac JL (2021) Algorithmic risk assessment in the hands of humans. Available at SSRN 3489440

Strandburg KJ (2021) Adjudicating with inscrutable decision rules. In: Pelillo M, Scantamburlo T (eds) Machines we trust. Perspectives on dependable AI. The MIT Press, pp 61–85

Vaccaro Michelle, Waldo Jim (2019) The effects of mixing machine learning and human judgment. Commun ACM 62(11):104–110

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

Verma S, Rubin J (2018) Fairness definitions explained. In: Proceedings of the international workshop on software fairness, FairWare '18, New York, NY, USA. Association for Computing Machinery, pp 1–7

Zafar MB, Valera I, Rodriguez MG, Gummadi KP (2015) Learning fair classifiers. In proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research 54:962–970

Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C (2013) Learning fair representations. In: Dasgupta S, McAllester D (eds) Proceedings of the 30th international conference on machine learning, volume 28 of proceedings of machine learning research, Atlanta, Georgia, USA. PMLR, pp 325–333