Fairness

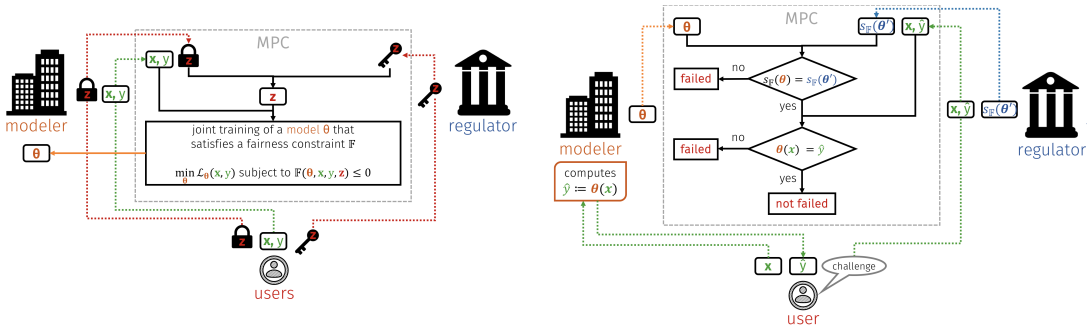

Figure 1.4: Protocols based on secure multi-party computation to learn (left) and verify (right) a fair model using only encrypted sensitive attributes; for details, see [142].

Algorithmic decision making processes are increasingly becoming automated and data-driven in both online (e.g., spam filtering, product personalization) as well as offline (e.g., pretrial risk assessment, mortgage approvals) settings. However, as automated data analysis supplements and even replaces human supervision in decision making, there are growing concerns from civil organizations, governments, and researchers about potential unfairness of these algorithmic decision systems towards people from certain demographic groups (e.g., gender or ethnic groups).

To assay the set of problems and alleviate these concerns, a number of recent studies in the emerging field of ethical machine learning have proposed and analyzed mechanisms to ensure that algorithmic decision systems do not lead to unfair outcomes, or perpetuate historic biases and harmful stereotypes. Such work has the potential to ensure that AI systems are used in a way that is compatible with human values, and it can also help us better understand some of our own biases as we take decisions driven by data and prior knowledge.

Definitions, Metrics, and Mechanisms Different forms of legally-problematic discrimination are commonly divided into several categories, the first one being *disparate treatment* (or *direct discrimination*), which occurs if individuals are treated differently according to their sensitive attributes (with all others equal). To avoid disparate treatment, one should not inquire about individuals' sensitive data ("fairness by unawareness"). While this has some intuitive appeal, a significant concern is that sensitive attributes may often be accurately predicted ("reconstructed") from non-sensitive attributes.

Hence, the second form, *disparate impact* (or *indirect discrimination*), occurs when the outcomes of decisions disproportionately benefit or hurt individuals from subgroups with particular sensitive attribute settings. Much recent work in fair learning has focused on approaches to avoiding various notions of disparate impact. Specificially, *demographic parity* demands that the proportion of people in each sensitive group receiving the favorable outcome must be equal. Building on similar ideas, *equality of opportunity* or *disparate mistreatment* demand that among all people who deserve to receive the favorable outcome (for instance, people who would pay back a loan, or people who do not go on to reoffend if released from prison on parole), the fractions of people actually receiving it are equal across sensitive groups. Most of the existing criteria are observational: They depend only on the joint distribution of predictor, protected attribute, features, and outcome, and are formulated as potentially approximate conditional independencies.

In our work, we have worked on defining, measuring, and efficiently mitigating the (un)fairness of a decision-making process regarding people from different sensitive groups. Specifically, we proposed a novel preference-based notion of (un)fairness, which is inspired by the fair division and envy-freeness literature in economics and game theory [159]. This definition provides a more flexible alternative to previous notions that are mostly based on parity of distributions of outcomes. It focuses on whether any group of users would collectively prefer its treatment regardless of the (dis)parity as compared to the other groups, when they are given the choice

between various sets of decision treatments.

Moreover, we contributed to algorithmic solutions to mitigate unfairness by developing flexible constraint-based frameworks to enable the design of fair margin-based classifiers [13]. The main technical innovation of our framework is a general and intuitive measure of decision boundary unfairness, which serves as a tractable proxy to several of the most popular computational definitions of unfairness from the literature, such as disparate impact and mistreatment, or preferred fairness. We can thus reduce the design of fair margin-based classifiers to adding tractable constraints on their decision boundaries.

Avoiding both disparate impact and disparate mistreatment is a major challenge. First, to avoid disparate mistreatment, the modeler often needs access to sensitive attributes. However, actively taking sensitive attributes into account introduces disparate treatment, an apparent contradiction. Second, individuals are unlikely to want to entrust sensitive attributes to modelers in all application domains. Finally, legal barriers – for example EU's General Data Protection Regulation (GDPR) – may limit collection and processing of sensitive personal data. We introduce cryptographic methods to resolve these tensions in the intersection of privacy, accountability, and fairness. By encrypting sensitive attributes, we show in our recent ICML paper how a fair model may be learned, checked, or have its outputs verified and held to account, without users revealing their sensitive attributes, cf. title figure [142].

Given the local expertise on causality, we early on realized a major limitation of fairness criteria based solely on the distribution of the observational data [173]. The way humans reason about fairness in decision-making processes crucially hinges on the causal relations underlying the observed decisions. It is conceptually insightful as well as practically relevant to incorporate the causal pathways into fair training procedures. Most fairness notions are based solely on the joint distribution of all variables at play and can thus not distinguish between different causal structures (leading to the same observation distribution) that can have vastly different intuitive interpretations of what is fair. Within the causal framework, we propose novel fairness criteria as well as training methods to achieve fair classifiers that come with theoretical guarantees under certain regularity assumptions.

In particular, we distinguish two context-dependent scenarios. In the *skeptic viewpoint*, we assume that any causal influence from the sensitive attribute on the outcome amounts to harmful discrimination unless it is mitigated via *resolving variables*, which we deem fair to use in our decision as measured. In the *benevolent viewpoint*, we specifically identify *proxy variables* (causal descendents) of the sensitive attribute that must not influence the decision (e.g., the name of a job applicant), but allow for causal pathways from the sensitive attribute to the outcome that do not go through proxies.

Fairness in Human Decision Making As described above, there has been a flurry of work on developing computational mechanisms to make sure that the machine learning methods that fuel algorithmic decision making are fair. In contrast, there is a lack of machine learning methods to ensure accuracy and fairness in human decision making, which is still prevalent in a wide range of critical applications such as, e.g., jail-or-release decisions by judges, or accept-or-reject decisions in academia. In this context, each decision is taken by an expert who is typically chosen uniformly at random from a pool of experts. In our recent NeurIPS paper [125], we showed that a random assignment might result in undesirable results in terms of both accuracy and fairness, and propose an algorithm to perform an assignment between decisions and experts which allows optimizing the accuracy and fairness of the overall decision-making process.

More information: https://ei.is.mpg.de/project/fairness