Indistinguishable Predictions and Multi-Group Fair Learning*

Guy N. Rothblum[1][0000−0001−5273−6472]

Apple, Cupertino, USA

Abstract. Prediction algorithms assign numbers to individuals that are popularly understood as individual "probabilities"—what is the probability that an applicant will repay a loan? Automated predictions increasingly form the basis for life-altering decisions, and this raises a host of concerns. Concerns about the *fairness* of the resulting predictions are particularly alarming: for example, the predictor might perform poorly on a protected minority group. We survey recent developments in formalizing and addressing such concerns.

Inspired by the theory of computational indistinguishability, the recently proposed notion of *Outcome Indistinguishability (OI)* [Dwork *et al.*, STOC 2021] requires that the *predicted* distribution of outcomes cannot be distinguished from the real-world distribution. Outcome Indistinguishability is a strong requirement for obtaining meaningful predictions. Happily, it can be obtained: techniques from the algorithmic fairness literature [Hebert-Johnson *et al.*, ICML 2018] yield algorithms for learning OI predictors from real-world outcome data.

Returning to the motivation of addressing fairness concerns, Outcome Indistinguishability can be used to provide robust and general guarantees for protected demographic groups [Rothblum and Yona, ICML 2021]. This gives algorithms that can learn a single predictor that "performs well" for every group in a given rich collection G of overlapping subgroups. Performance is measured using a loss function, which can be quite general and can itself incorporate fairness concerns.

1 Introduction

Machine learning tools are used to make and inform increasingly consequential decisions about individuals. Examples range from medical risk prediction to hiring decisions and criminal justice. Automated risk prediction comes with

* This extended abstract overviews the recent developments and contributions in [15] and [42], and borrows liberally from those works. The research described in this extended abstract was done while the author was at the Weizmann Institute of Science and while visiting Microsoft Research. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702), from the Israel Science Foundation (grant number 5219/17), and from the Simons Foundation Collaboration on the Theory of Algorithmic Fairness.

benefits, but it also raises substantial societal concerns. First and foremost, how meaningful are the predictions? Another prominent concern is that these algorithms might discriminate against protected and/or disadvantaged groups. In particular, a learned predictor might perform differently on a protected subgroup compared to the general population.

In a sequence of recent works we tackle these concerns with novel tools and perspectives. Our approach is inspired by the cryptographic and complexity-theoretic literature on indistinguishability, as well as the burgeoning literature on algorithmic fairness. This manuscript aims to highlight these developments, focusing on the following contributions:

Outcome Indistinguishability: a new framework for meaningful predictions. Prediction algorithms "score" individuals, mapping them to numbers in $[0, 1]$ that are popularly understood as "probabilities" or "likelihoods" of observable events: the probability of 5-year survival, the chance that the loan will be repaid on schedule, the likelihood that the student will graduate within four years. What do these numbers actually mean? How can we judge a predicted probability when the event (e.g. 5-year survival) is non-repeatable? The question of "individual probabilities" has been studied for decades across many disciplines without clear resolution (see Dawid [9]).

In recent work with Dwork *et al.* [15] we propose *Outcome Indistinguishability (OI)*: a novel framework for guaranteeing meaningful predictions. In a nutshell, the predictions should be *indistinguishable*, given real-world outcomes, from the true probabilities governing reality. We show that Outcome Indistinguishability is feasible: building on a connection to the notion of multi-calibration [31], we construct algorithms for learning OI predictors from outcome data. These contributions are described in Section 2.

Multi-group fair learning. The literature on (supervised) learning and loss minimization takes a different approach to predicting outcomes. Given an i.i.d. training set of labeled data, the goal is learning a predictor p that performs well on the underlying distribution. Performance is measured using a *loss function*, such as the squared loss or various other measures. In agnostic learning [36], the loss incurred by the predictor p should be competitive with the best predictor in a benchmark class $\mathcal{H}$. These approaches have enjoyed tremendous success, but they does not resolve basic questions about the meaningfulness of predictions. Given a predictor that achieves a certain loss, how should we judge its performance? Both at an aggregate level, over the entire population (what level of loss is "good"?), at the level of protected subgroups, and at the level of individual predictions. Indeed, it has been demonstrated that standard machine learning tools, when applied to standard data sets, produce predictors whose performance on protected demographic groups is quite poor [4].

Motivated by these concerns, in work with Yona [42] we study *multi-group* agnostic learning. For a rich collection $\mathcal{G}$ of (potentially) overlapping groups, our goal is to learn a single predictor p, such that the loss experienced by every group $g \in \mathcal{G}$ (when classified by p) is not much larger than the loss of the best

predictor *for that group* in the class $\mathcal{H}$. This should hold for all groups in $\mathcal{G}$ simultaneously. To capture a wide variety of settings, we aim to be quite general in our treatment of different loss functions. In particular, the loss function itself can also incorporate fairness concerns. We show that this ambitious objective is obtainable! Multi-group fair predictors can be learned for a rich class of loss functions. The learning procedure itself is constructed via a reduction to Outcome Indistinguishability, demonstrating the power and the flexibility of the OI framework. We detail these contributions in Section 3.

Further related work and recent developments. We discussed further related work in Section 2.3 and before Section 3.1. We conclude in Section 4 with a brief discussion of more recent developments that build on the contributions described in this extended abstract.

2 Outcome Indistinguishability

The recently-proposed notion of *Outcome Indistinguishability* (OI) [15] proposes and studies novel criteria for significant predictions. The outputs of a prediction algorithm are viewed as defining a generative model for observational outcomes. Ideally, the outcomes from this generative model should "look like" the outcomes produced by Nature (the real world). A predictor satisfying outcome indistinguishability provides a generative model that cannot be efficiently refuted on the basis of the real-life observations produced by Nature. In this sense, the probabilities defined by any OI predictor provide a meaningful model of the "probabilities" assigned by Nature: even granted full access to the predictive model and historical outcomes from Nature, no analyst can invalidate the model's predictions. This provides a computational / cryptographic perspective on the deeper discussion of what we should demand of prediction algorithms–a subject of intense study in the statistics community for over 30 years (see, *e.g.*, the forecasting work in [8,20,21,44,43])—and how they should be used. For example, the study of Outcome Indistinguishability has led to lower bound results that provide scientific teeth to the political argument that, if risk prediction instruments are to be used by the courts (as they often are in the United States), then at the very least auditors should be given oracle access to the algorithms.

Basic notation. We focus on the fundamental setting of predicting a binary outcome, but note that the OI framework has been extended to deal with more general outcomes [16]. Individuals are represented by a collection of covariates from a discrete domain $\mathcal{X}$, for example, the set of d-bit strings (there might be collisions, or it may be the case that each individual has a unique representation). We model Nature as a joint distribution, denoted $\mathcal{D}^*$, over individuals and outcomes, where $y_x^* \in \{0,1\}$ represents Nature's choice of outcome for individual $x \in X$. We use $x \sim \mathcal{D}_{\mathcal{X}}$ to denote a sample from Nature's marginal distribution over individuals and denote by $p_x^* \in [0,1]$ the conditional probability that Nature assigns to the outcome y_x^*, conditioned on x. We emphasize, however,

that Nature may choose $p^*_x \in \{0,1\}$ to be deterministic; our definitions and constructions are agnostic as to this point.

A *predictor* is a function $\tilde{p} : \mathcal{X} \to [0,1]$ that maps an individual $x \in \mathcal{X}$ to an estimate $\tilde{p}_x$ of the conditional probability of $y^*_x = 1$. For a predictor $\tilde{p} : \mathcal{X} \to [0,1]$, we denote by $(x, \tilde{y}_x) \sim \mathcal{D}(\tilde{p})$ the random process of drawing an individual-outcome pair, where $x \sim \mathcal{D}_{\mathcal{X}}$ is sampled from Nature's distribution over individuals, and then the outcome $\tilde{y}_x \sim \mathrm{Ber}(\tilde{p}_x)$ is sampled from the Bernoulli distribution with parameter $\tilde{p}_x$.

Outcome Indistinguishability. Imagine that Nature selects $p^*_x = 1$ for half of the mass of $x \sim \mathcal{D}_{\mathcal{X}}$ and $p^*_x = 0$ for the remainder. If the two sets of individuals are easy to identify then we can potentially recover a close approximation to p^*. Suppose, however, that the sets are *computationally indistinguishable*, in the sense that given $x \sim \mathcal{D}_{\mathcal{X}}$, no efficient observer can guess if $p^*_x = 1$ or $p^*_x = 0$ with probability significantly better than $1/2$. In this case, producing the estimates $\tilde{p}_x = 1/2$ *for every individual* $x \in \mathcal{X}$ captures the best computationally feasible understanding of Nature: given limited computational power, the outcomes produced by Nature may faithfully be modeled as a random. In particular, if Nature were to change the outcome generation probabilities from p^* to $\tilde{p}$ we, as computationally bounded observers, will not notice. In other words, predictors satisfying OI give rise to models of Nature that cannot be falsified based only on observational data.

Definition 1 (Outcome Indistinguishability). *Fix Nature's distribution $\mathcal{D}^*$. For a class of distinguishers $\mathcal{A}$ and $\varepsilon > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ satisfies $(\mathcal{A}, \varepsilon)$-outcome indistinguishability (OI) if for every $A \in \mathcal{A}$,*

$$\left| \Pr_{(x,y^*_x)\sim\mathcal{D}^*} [\ A(x, y^*_x; \tilde{p}) = 1\ ] - \Pr_{(x,\tilde{y}_x)\sim\mathcal{D}(\tilde{p})} [\ A(x, \tilde{y}_x; \tilde{p}) = 1\ ] \right| \leq \varepsilon.$$

The above definition is purposefully vague about the distinguisher's access to the predictor $\tilde{p}$: we anchor a hierarchy of OI variants around different levels of access to $\tilde{p}$. The definition of Outcome Indistinguishability can be extended in many other ways, for example to distinguishers receive multiple samples from each distribution (this will be used in Lemma 11 below), and to the case of non-Boolean outcomes [16].

In the extreme, when we think of $\mathcal{A}$ as the set of all polynomial-time distinguishers, outcome indistinguishability sets a demanding standard for predictors that model Nature. Given an OI predictor $\tilde{p}$, even the most skeptical scientist—who, for example, does not believe that Nature can be captured by a simple computational model—cannot refute the model's predictions through observation alone. This framing gives a cryptographic or computational perspective on the scientific method, by considering $\tilde{p}$ as expressing a hypothesis that cannot be falsified through observational investigation.

The OI hierarchy. In the most basic variant of the definition, the distinguisher does not get direct access to the predicted probabilities, only to the outcomes

(drawn by p^* or by $\tilde{p}$). A predictor $\tilde{p}$ satisfies this most basic notion of OI if for all $A \in \mathcal{A}$, the probability that A accepts the sample (x, y_x) is (nearly) the same for Nature's distribution and the predictor's distribution. The requirement can be strengthened by also giving the distinguisher direct access the predictor $\tilde{p}$ itself: either access to the predicted probability $\tilde{p}_x$ of the sample at hand, oracle access, or even access to the code. We emphasize, however, that the distinguisher never gets access to p^*: Nature's true probabilities are unknowable.

These differing levels of access to the predictor produce a hierarchy of definitions, which we illustrate through an example. Imagine a medical board that wishes to audit the output of a program $\tilde{p}$ used to estimate the chances of five-year survival of patients under a given course of treatment. We can view the medical board as a distinguisher $A \in \mathcal{A}$. To perform the audit, the board receives historical files of patients and their five-year predicted (*i.e.*, drawn from $\mathcal{D}(\tilde{p})$) or actual (drawn from $\mathcal{D}^*$) outcomes. The requirement is that these two cases be indistinguishable to the board.

1. To start, the board is only given samples, and must distinguish Nature's samples $(x, y_x^*) \sim \mathcal{D}^*$ from those sampled according to the predicted distribution $(x, \tilde{y}_x) \sim \mathcal{D}(\tilde{p})$. The board gets no direct access to predictions $\tilde{p}_x$ of the program; we call this variant *no-access-OI*.
2. Naturally, the board may ask to see the predictions $\tilde{p}_x$ for each sampled individual. In this extension—*sample-access-OI*—the board must distinguish samples of the form $(x, y_x^*, \tilde{p}_x)$ and $(x, \tilde{y}_x, \tilde{p}_x)$, again for $(x, y_x^*) \sim \mathcal{D}^*$ and $(x, \tilde{y}_x) \sim \mathcal{D}(\tilde{p})$.
3. *Oracle-access-OI* allows the board to make queries to the program $\tilde{p}$ on arbitrary individuals, perhaps to examine how the algorithm behaves on related (but unsampled) patients.
4. Finally, in *code-access-OI*, the board is allowed to examine not only the predictions from $\tilde{p}$ but also the actual code, *i.e.*, the full implementation details of the program computing $\tilde{p}$.

2.1 Feasibility and Learnability of OI Predictors

Do efficient OI predictors always exist? In particular, *can we bound the complexity of OI predictors, independently of the complexity of Nature's distribution?* The picture here is subtle, and Outcome Indistinguishability differs qualitatively from prior notions of indistinguishability.

Beyond he question of *existence*, it is also important to understand whether it is possible to *learn* OI predictors from outcome data (we focus on the natural setting where outcomes are all we can hope to observe). A learning algorithm receives outcome data drawn from $\mathcal{D}^*$, with the goal of learning a predictor $\tilde{p}$ that satisfies OI w.r.t a given class $\mathcal{A}$ of distinguishers. Happily, OI predictors can be learned from outcome data at all levels of the hierarchy, with *logarithmic* sample complexity in the size of the family of distinguishers.

The first two level of the OI hierarchy. Dwork *et al.* [15] show that no-access-OI and sample-access-OI are closely related to the notions of multi-accuracy and multi-calibration [31], respectively, studied in the algorithmic fairness literature. Very loosely, for a collection $\mathcal{C}$ of subpopulations of individuals, $(\mathcal{C}, \alpha)$-multi-calibration asks that a predictor $\tilde{p}$ be calibrated (up to α error) not just overall, but also when we restrict our attention to subpopulations $S \subseteq \mathcal{X}$ for every set $S \in \mathcal{C}$. Here, calibration over S means that if we restrict our attention to individuals $x \in S$ for which $\tilde{p}_x = v$, then the fraction individuals with positive outcomes (i.e., $x \in S$ such that $y^*_x = 1$) is roughly v. Loosely, by equivalent we mean that each notion can enforce the other, for closely related classes $\mathcal{C}$ and $\mathcal{A}$. Importantly, the relation between the class of distinguishers and collection of subpopulations preserves most natural measures of complexity; in other words, if we take $\mathcal{A}$ to be a class of efficient distinguishers, then evaluating set membership for the populations in $\mathcal{C}$ will be efficient (and vice versa). No-access-OI is similarly equivalent to the weaker notion of multi-accuracy, which requires accurate expectations for each $S \in \mathcal{C}$, rather than calibration.

Leveraging feasibility results for the fairness notions from [31], we can obtain efficient predictors satisfying no-access-OI or sample-access-OI, by reduction to multi-accuracy and multi-calibration. Informally, for each of these levels, we can obtain OI predictors whose complexity scales linearly in the complexity of $\mathcal{A}$ and inverse polynomially in the desired distinguishing advantage ε. The result is quite generic; for concreteness, we state the theorem using circuit size as the complexity measure.

Theorem 2 (Informal [15]). *Let $\mathcal{A}$ be a class of distinguishers implemented by size-s circuits. For any $\mathcal{D}^*$ and $\varepsilon > 0$, there exists a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ satisfying $(\mathcal{A}, \varepsilon)$-sample-access-OI (similarly, no-access-OI) implemented by a circuit of size $O(s/\varepsilon^2)$.*

OI predictors can be learned using only a bounded number of observed outcomes $(x, y^*_x) \sim \mathcal{D}^*$. The learning algorithm, which leverages algorithms for learning multicalibrated predictors, has sample complexity that is logarithmic in the size of the distinguisher class $\mathcal{A}$. The runtime for learning is *linear* in the size of $\mathcal{A}$ and polynomial in $(1/\varepsilon)$. Alternatively, the task of learning an OI predictor can be reduced to an agnostic learning task on a hypothesis class that is related to $\mathcal{A}$. See [31,15] for further details.

The top two layers of the OI hierarchy. There is a general-purpose algorithm for constructing OI predictors, even when the distinguishers are allowed arbitrary access to the predictor in question. This shows the existence and learnability of oracle-access-OI and code-access-OI predictors. This construction of [15] extends the learning algorithm for multi-calibration of [31] to the more general setting of OI. When we allow such powerful distinguishers, the learned predictor $\tilde{p}$ is quantitatively less efficient than in the weaker notions of OI. For the overivew in this manuscript we state the bound informally, assuming the distinguishers are implemented by circuits with oracle gates (see [15] for a full and formal treatment). As an example, if we let $\mathcal{A}$ be the set of oracle-circuits of some fixed

polynomial size (in the dimension d of individual's representations), and allow arbitrary oracle queries, then $\tilde{p}$ will be of size $d^{O(1/\varepsilon^2)}$.

Theorem 3 (Informal [15]). *Let $\mathcal{A}$ be a class of oracle-circuit distinguishers implemented by size-s circuits that make at most q oracle calls to the predictor in question. For any $\mathcal{D}^*$ and $\varepsilon > 0$, there exists a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ satisfying $(\mathcal{A}, \varepsilon)$-oracle-access-OI implemented by a (non-oracle) circuit of size $s \cdot q^{O(1/\varepsilon^2)}$.*

We omit a discussion of the complexity of learning oracle-access-OI, as well as the results (and definitional subtleties) of code-access-OI. We refer the interested reader to [15]. We remark that for code-access-OI, the complexity may scale doubly exponentially in $\text{poly}(1/\varepsilon)$.

Hardness via Fine-Grained Complexity. Dwork *et al.* [15] established a connection between the fine-grained complexity of well-studied problems and the complexity of achieving oracle-access-OI. Under the assumption that the (randomized) complexity of counting k-cliques in n-vertex graphs is $n^{\Omega(k)}$, the construction of Theorem 3 is optimal up to polynomial factors. Specifically, they rule out (under this assumption) the possibility that the complexity of a oracle-access-OI predictor can be a fixed polynomial in the complexity of the distinguishers in $\mathcal{A}$ and in the distinguishing advantage ε. Their hardness result holds for constant distinguishing advantage ε and for an efficiently-sampleable distribution $\mathcal{D}^*$. This hardness results are in stark contrast to the state of affairs for sample-access-OI (see Theorem 2). Concretely, in the parameters of the upper bound, the result based on the hardness of clique-counting rules out any predictor $\tilde{p}$ satisfying oracle-access-OI of (uniform) size significantly smaller than $d^{\Omega(1/\epsilon)}$.

Theorem 4 (Informal [15]). *For $k \in \mathbb{N}$, assume there exist $\alpha > 0$ s.t. there is no $o(n^{\alpha \cdot k})$-time randomized algorithm for counting k-cliques. Then, there exist: $\mathcal{X} \subseteq \{0,1\}^{d^2}$, an efficiently-sampleable distribution $\mathcal{D}^*$, and a class $\mathcal{A}$ of distinguishers that run in time $\tilde{O}(d^3)$ and make $\tilde{O}(d)$ oracle queries to $\tilde{p}$, s.t. for $\varepsilon = \frac{1}{100k}$, no predictor $\tilde{p}$ that runs in time $(d^{\alpha \cdot k} \cdot \log^{-\omega(1)}(d))$ can satisfy $(\mathcal{A}, \varepsilon)$-oracle-access-OI.*

This lower bound is robust to the computational model: assuming that clique-counting requires $n^{\Omega(k)}$-sized circuits implies a similar lower bound on the circuit size of oracle-access-OI predictors. The complexity of clique counting has been widely studied and related to other problems in the fine-grained and parameterized complexity literatures, see the discussion in [15]. We note that, under the plausible assumption that the fine-grained complexity of known clique counting algorithms is tight, this result shows that obtaining oracle-access-OI is as hard, up to sub-polynomial factors, as computing p^*. We emphasize that this is the case even though the running time of the distinguishers can be arbitrarily small compared to the running time of p^*.

Dwork *et al.* also show that, under the (milder) assumption that BPP $\neq$ PSPACE, there exists a polynomial collection of distinguishers and a distribution

$\mathcal{D}^*$, for which no polynomial-time predictor $\tilde{p}$ can be OI. The distinction from the fine-grained result (beyond the difference in the assumptions) is that here $\mathcal{D}^*$ is not efficiently sampleable, and the distinguishing advantage for which OI is hard is much smaller.

2.2 Broader Context and Discussion

We highlight a few possible interpretations and insights that stem from the technical results described above. The ability to construct predictors that satisfy outcome indistinguishability can be viewed both positively and negatively. On one hand, the feasibility results demonstrate the possibility of learning generative models of observed phenomena that withstand very powerful scrutiny, even given the complete description of the model. On the other hand, OI does not guarantee statistical closeness to Nature (it need not be the case that $p^* \approx \tilde{p}$). Thus, the feasibility results demonstrate the ability to learn an *incorrect* model that cannot be refuted by efficient inspection. In this sense, attempting to recover the "true" model of Nature based on real-world observations is futile: no efficient analyst can falsify the outcomes of the model defined by $\tilde{p}$, agnostic to the "true" laws of Nature.

The most surprising (and potentially-disturbing) aspect of our results may be the complexity of achieving oracle-access-OI and code-access-OI. In particular, for these levels, we show strong evidence that there exist p^* and $\mathcal{A}$ that do not admit efficient OI predictors $\tilde{p}$, *even when $\mathcal{A}$ is a class of efficient distinguishers!* That is, there are choices of Nature that cannot be modeled simply, even if all we care about is passing simple tests. This stands in stark contrast to the existing literature on indistinguishability in cryptography, where the complexity of the indistinguishable object is usually smaller than the distinguishers' complexity, and in complexity theory, where the object is polynomial in the distinguishers' complexity.

Lessons for auditing predictors. The increased distinguishing power of oracle access to the predictor in oracle-access-OI may have bearing on ongoing societal debates regarding appropriate usage of algorithms when making high-stakes judgments about individuals, e.g. in the context of the criminal justice system. Much of the discussion revolves around the idea of *auditing* the predictions, for accuracy and fairness. The separation between oracle-access-OI and sample-access-OI provides a rigorous foundation for the argument that auditors should at the very least have query access to the prediction algorithms they are auditing: given a fixed computational bound, the auditors with oracle-access may perform significantly stronger tests than those who only receive sample access.

The representation is central. The *representation* of individuals is of central importance to the OI framework. If the representation space $\mathcal{X}$ contains little information that is relevant to the prediction task at hand, then p^* itself will not be very informative, and neither will a predictor $\tilde{p}$ that is OI. It is also important to note that a fixed representation of features may be informative

for the general population, but lacking in pertinent information for a protected demographic group. In any setting where automated prediction is considered for deployment, the representation or feature space must be carefully considered.

The OI framework can be extended, allowing for the representation of individuals to be augmented throughout time. Given such an enriched representation, and an enriched class of distinguishers (which take advantage of the new representation), the predictor $\tilde{p}$ can be updated to obtain an improved predictor that fools the new class of distinguishers. A potential argument can be used to show that each such update moves $\tilde{p}$ meaningfully towards the "true" individual probabilities, and thus this representation-augmentation process cannot happen too many times. See [15] and see also the work of [22].

2.3 Further Related Work

The framing of outcome indistinguishability draws directly from the notion of computational indistinguishability, studied extensively in the literature on cryptography, pseudorandomness, and complexity theory (see, e.g., [23,25,46,24] and references therein).

Outcome Indistinguishability is related to the extensive literature on online *forecast testing*. The latter literature focuses on an online setting where there are two players, Nature and the Algorithm. Nature controls the data generating process (e.g., the weather patterns), while the Algorithm tries to assess, on each Day $t-1$, the probability of an event on Day t (e.g., will it rain tomorrow?). In the early 1980s, [8] proposed that, at the very least, forecasts should be calibrated. Later works considered more stringent requirements. A signal result in the forecasting literature, due to Sandroni [43], applies to a more general notion of *tests*. A test tries to assess whether an algorithm's predictions are "reasonably accurate" with respect to the actual observations. It is required to satisfy a strong completeness property: no matter what Nature's true probabilities are, the test should accept them w.h.p. (indeed, calibration tests have this property). Sandroni's powerful result [43], shows, non-constructively[1], how to generate probability forecasts that fool any such complete test. The computational complexity of forecasting was studied by Fortnow and Vohra [19] and by Chung, Lui and Pass [6]. See [15] for a full comparison between the forecast testing literature and the new notion of Outcome Indistinguishability.

Algorithmic fairness. Tests are also implicit in the literature on algorithmic fairness, where they are sometimes referred to as *auditors*. One line of work, the *evidence-based fairness* framework—initially studied in [31,37,14]—relates directly to outcome indistinguishability and centers around tests that Nature always passes. Broadly, the framework takes the perspective that, first and foremost, predictors should reflect the "evidence" at hand—typically specified through historical outcome data—as well as the statistical and computational resources allow.

[1] The result leverages Fan's minimax theorem.

Central to evidence-based fairness is the notion of multi-calibration [31], which was also studied in the context of rankings in [14]. [33] provide algorithms for achieving an extension of multi-calibration that ensures calibration of higher moments of a scoring function, and show how it can be used to provide credible prediction intervals. [45] study multi-calibration from a sample-complexity perspective. In a similar vein, [47] study a notion of individualized calibration and show it can be obtained by randomized forecasters.

Evidence-based fairness is part of a more general paradigm for defining fairness notions, sometimes referred to as "multi-group" notions, which has received considerable interest in recent years [31,34,40,37,35,14,45,1,33]. This approach to fairness aims to strengthen the guarantees of notoriously-weak group fairness notions, while maintaining their practical appeal. For instance, [40,34,35] give notions of multi-group fairness based on parity notions studied in [11] and [30]. [1] extend this idea to the online setting. Other approaches to fairness adopt a different perspective, and intentionally audit for properties that Nature does not necessarily pass. Notable examples are group-based notions of parity [30,41,34,35].

3 Multi-PAC Learning

As discussed in the introduction, one prominent concern about predictors obtained via machine learning is that they might discriminate against protected groups. With fairness in mind, the loss minimization paradigm raises a fundamental concern: since the predictor's loss is measured over the entire underlying distribution, it might not reflect the predictor's performance on sub-populations such as protected demographic groups. Indeed, it has been demonstrated that standard machine learning tools, when applied to standard data sets, produce predictors whose performance on protected demographic groups is quite poor [4].

Motivated by these concerns, in work with Yona [42] (and building on earlier work by Blum and Lykouris [1]) we study *multi-group* agnostic learning. For a rich collection $\mathcal{G}$ of (potentially) overlapping groups, the goal is to learn a single predictor p, such that the loss experienced by every group $g \in \mathcal{G}$ (when classified by p) is not much larger than the loss of the best predictor *for that group* in the class $\mathcal{H}$. We emphasize that this should hold for all groups in $\mathcal{G}$ simultaneously. The study of this question also differs from much of the agnostic learning literature in considering quite general loss functions. In particular, the loss function itself may incorporate fairness considerations (see [42]). The question we ask is: *for which loss functions is multi-group agnostic learning possible?*

To see how this objective is different from the standard agnostic PAC learning setting, consider the simple example in which $\mathcal{H}$ is the class of hyperplanes and we have two subgroups $S, T \subseteq \mathcal{X}$. Suppose that the data is generated such that every group g has a hyperplane h_g that has very low error on it (but that these are different, so e.g. h_T has large loss on S and vice versa). This means that there is no classifier $h \in \mathcal{H}$ that perfectly labels the data. If S is small compared

to T, then the agnostic learning objective could be satisfied by h_T, the optimal classifier for T. For *multi-group* agnostic PAC, the fact that there is some other classifier in $\mathcal{H}$ that perfectly labels S serves to disqualify h_T (more generally, it could be the case that no $h \in \mathcal{H}$ will be multi-PAC). This also highlights that the multi-group objective becomes challenging when the groups in question are intersecting (if the groups are disjoint, we can combine the optimal classifiers for each group [13]).

Multi-group PAC learning via OI. [42] construct a "multi-PAC" agnostic learning algorithm for any loss function that satisfies: (i) a uniform convergence property: it should be possible to estimate the loss of a predictor (or a whole class) from data sampled i.i.d. from the underlying distribution, and (ii) f-proper: meaning that there should be a rule f for transforming Bayes-optimal predictions (the probabilities p^*) into loss-minimizing predictions. Under these two assumptions, there is an algorithm that, for any specified finite collection $\mathcal{G}$ and finite hypothesis class $\mathcal{H}$, learns a multi-group agnostic predictor from labeled data. The sample complexity is logarithmic in the sizes of $\mathcal{G}$ and $\mathcal{H}$. The algorithm is derived by a reduction to *outcome indistinguishability* (OI), drawing a new connection between OI and loss minimization, and demonstrating the power and the flexibility of the OI framework.

Related work. Blum and Lykouris [1] studied this question in an online setting with sequential predictions. Our focus is on the batch setting. They showed that (for every collection of groups and every benchmark hypothesis class) it is possible to achieve competitive loss for all groups, so long as the loss function is *decomposable*: the loss experienced by each group is an average of losses experienced by its members. On the other hand, they showed a loss function (the average of false negative and false positive rates), for which the objective is infeasible even in the batch setting.

See Section 2.3 for a discussion of related work in the algorithmic fairness literature. We briefly discuss the relationship to multi-group fair learning. Many works in the algorithmic fairness literature aim to ensure parity or balance between demographic groups, e.g. similar rates of positive predictions or similar false positive or false negative rates [30,41]. As discussed above, other works consider accuracy guarantees, such as calibration [7] for protected groups. Protections at the level of a single group might be too weak [12], and recent works have studied extending these notions to the setting of multiple overlapping groups [31,34].

3.1 Loss Functions

A loss function L is a mapping from a distribution $\mathcal{D}$ and a predictor p to $[0, 1]$. We use $L_{\mathcal{D}}(p)$ to denote the loss of p w.r.t. a distribution $\mathcal{D}$. For a sample $S = \{(x_i, y_i)\}_{i=1}^{m}$ we use $L_S(p)$ to denote the empirical loss, calculated as $L_{\hat{\mathcal{D}}}(p)$, where $\hat{\mathcal{D}}$ is the empirical distribution defined by the sample S. This setup is extremely general, and assumes nothing about the loss (except that it is bounded

and can't depend on what happens outside $\mathcal{D}$). In machine learning it is common to consider more structured losses, in which $L_{\mathcal{D}}(p)$ is the expected loss of p on a random example drawn according to $\mathcal{D}$. We refer to such structured losses as *decomposable* losses.

Definition 5 (Decomposable losses). *A loss function L is* decomposable *if there exists a function $\ell : X \times Y \times [0,1] \to [0,1]$ such that for every distribution $\mathcal{D}$ and predictor p, $L_{\mathcal{D}}(p) = \mathbf{E}_{(x,y)\sim\mathcal{D}}[\ell(x, y, p(x))]$.*

For example, for binary classifiers a standard decomposable loss is the 0-1 loss, in which $\ell(x, y, p(x)) = \mathbf{1}[p(x) \neq y]$. For predictors, an example of a standard decomposable loss is the squared loss, in which $\ell(x, y, p(x)) = (p(x) - y)^2$.

Beyond decomposable losses. While decomposable losses are standard and common, there are many loss functions of interest that don't have this form – especially in the literature on algorithmic fairness. For this reason, we focus on a general notion of loss functions in our exploration of multi-group agnostic PAC learning. Two prominent examples of such losses are:

- **Calibration.** [41,5,31,45] As discussed above, a predictor is *calibrated* if for every value $v \in [0,1]$, conditioned on $p(x) = v$, the true expectation of the label is close to v. This is a fundamental requirement in forecasting [7,20]. This loss is not decomposable because it is a global function of the predictions, not a property of the prediction for a single $x \in \mathcal{X}$.
- **One-sided error rates** [30,5,1,2,34]: The *false positive rate* (similarly, false negative rate) measures the probability of a random example being labeled as $p(x) = 1$, conditioned on the true label being $y = 0$. This isn't a decomposable loss because the exact contribution of a single misclassification depends on the frequency of the negative labels, which is a global property.

See [42] for further examples and discussion. In this manuscript we focus on loss functions with two additional properties: uniform convergence and f-properness.

Uniform Convergence. We begin by recalling uniform convergence for hypotheses classes:

Definition 6 (Uniform Convergence for hypotheses classes). *We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property (w.r.t. a domain $X \times Y$ and a loss function L) if there exists a function $m^{UC}_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over $X \times Y$, if S is a sample of $m \geq m^{UC}_{\mathcal{H}}(\varepsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, $\forall h \in \mathcal{H}$: $\quad |L_S(h) - L_D(h)| \leq \varepsilon$.*

In our context, we are interested in uniform convergence as a property of the *loss function*. A loss L has uniform convergence (w.r.t finite classes) with sample

complexity $m_L^{UC} : (0,1)^2 \times \mathbb{N} \to \mathbb{N}$ if every finite class $\mathcal{H}$ has the uniform convergence property w.r.t L with sample complexity $m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq m_L^{UC}(\varepsilon, \delta, |\mathcal{H}|)$. Specifically, we will be interested in losses that have the uniform convergence property with sample complexity that depends polynomially on $1/\varepsilon, 1/\delta$ and $\log |\mathcal{H}|$. This gives rise to the following definition:

Definition 7 (Uniform convergence for loss functions). *A loss L has the uniform convergence property (w.r.t finite classes) with sample complexity $m_L^{UC} : (0,1)^2 \times \mathbb{N} \to \mathbb{N}$ if there exists a polynomial $f : \mathbb{R}^3 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and $k \in \mathbb{N}$,*

$$m_L^{UC}(\varepsilon, \delta, k) \triangleq \max_{\mathcal{H}: |\mathcal{H}|=k} m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq f(1/\varepsilon, 1/\delta, \log(k))$$

The uniform convergence property is satisfied by any decomposable loss function. This follows by a combination of Heoffding's bound (for a single h) and a union bound to get a simultaneous guarantee for every $h \in \mathcal{H}$. For calibration, uniform convergence follows as a special case of the bounds in [45]. However, the loss that takes a convex combination of the false positive and the false negative rates does *not* satisfy uniform convergence. See [1] and [42]) for further details, examples and discussion.

f-proper loss functions. Recall that proper losses (or proper scoring functions) are losses that are minimized by the Bayes optimal predictor p^*, i.e. conditional expectation predictor $x \mapsto \mathbf{E}_{\mathcal{D}}[y|x]$ [3]. The f-proper condition is a relaxation: it says that for every distribution, a minimizer can be obtained as some *local* transformation of this predictor (i.e. that does not depend on the rest of the distribution).

Definition 8 (f-proper). *For a function $f : \mathcal{X} \times [0,1] \to [0,1]$, we say that a loss L is f-*proper *if for every distribution $\mathcal{D}$ on $\mathcal{X} \times Y$, the classifier $h_{\mathcal{D}}$ given by $h_{\mathcal{D}}(x) = f(x, p^*(x) = \mathbf{E}_{\mathcal{D}}[y|x])$ minimizes the loss w.r.t $\mathcal{D}$: $h_{\mathcal{D}} \in \arg\min_h L_{\mathcal{D}}(h)$.*

The L_2 loss is a well-known example of a proper loss function (f simply outputs its second argument). The 0-1 loss is another well-known example, where the loss is minimized by $f(x, z) = 1\,[z \geq 0.5]$.

3.2 Multigroup PAC Learnability via OI

The objective of agnostic PAC learning is outputting a predictor p that satisfies $L_{\mathcal{D}}(p) \lesssim L_{\mathcal{D}}(\mathcal{H})$. Multigroup (agnostic) PAC learning [42] asks for a predictor that satisfies the above, but simultaneously for every group g in a collection $\mathcal{G}$: $L_{\mathcal{D}_g}(p) \lesssim L_{\mathcal{D}_g}(\mathcal{H})$, where $\mathcal{D}_g$ denotes the restriction of $\mathcal{D}$ to samples from g. Moreover, a learning algorithm should be able to find such a solution in sample complexity that is inverse-polynomial in the parameters in question and polylogarithmic in the sizes of $\mathcal{H}$ and $\mathcal{G}$.

Definition 9 (Multi-PAC learnability). *A loss L is* multi-PAC learnable *with sample complexity $m_L^{gPAC} : (0,1)^3 \times \mathbb{N}^2 \to \mathbb{N}$ if there exists a learning algorithm with the following property: For every $\varepsilon, \delta, \gamma \in (0,1)$, for every finite hypothesis class $\mathcal{H}$, for every finite collection of subgroups $G \subseteq 2^{\mathcal{X}}$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times Y$, when running the learning algorithm on $m \geq m_L^{gPAC}(\varepsilon, \delta, \gamma, |\mathcal{H}|, |\mathcal{G}|)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns p such that, with probability at least $1 - \delta$ (over the choice of the m training examples and the coins of the learning algorithm) $g \in \mathcal{G}_\gamma$, $L_{\mathcal{D}_g}(p) \leq L_{\mathcal{D}_g}(\mathcal{H}) + \varepsilon$, where $\mathcal{G}_\gamma \subseteq \mathcal{G}$ is the subset of groups whose mass under $\mathcal{D}$ is at least γ: $\mathcal{G}_\gamma = \{g \in \mathcal{G} : \mathbf{Pr}_{\mathcal{D}}[x \in g] \geq \gamma\}$.*

Additionally, the sample complexity m_L^{gPAC} should be polynomial in $(1/\varepsilon)$, in $(1/\delta)$, in $(1/\gamma)$, in $(\log(|\mathcal{H}|))$, and in $\log(|\mathcal{G}|)$.

When $\mathcal{G}$ consists of intersecting groups, it is not immediately clear that this objective is remotely feasible: it might not be satisfied by *any* predictor $p : \mathcal{X} \to [0,1]$! For a simple (but contrived) example, let h^0, h^1 denote the all-zeroes and all-ones predictors, and consider a loss L that specifies that $L_{\mathcal{D}_S}(h^0) = 0$ and $L_{\mathcal{D}_T}(h^1) = 0$ (and for any other classifier p, the loss of every distribution is always 1). Then the multi-group objective w.r.t $\mathcal{G} = \{S, T\}$ requires that we label the intersection $S \cap T$ as both 1 and 0, which is impossible. See [1,42] for further discussion and natural examples of infeasible loss functions.

Rothblum and Yona [42] show that multi-PAC predictors exist and can be learned for every loss function satistfying the uniform uniform convergence and f-proper conditions.

Theorem 10 (Multi-PAC Learning [42]). *If L is f-proper (Definition 8) and has the uniform convergence property (Definition 7), then L is multi-group learnable (Definition 9).*

The Theorem is proved by a reduction to Outcome Indistinguishability. For a loss function L satisfying the theorem conditions, for and group g and hypothesis h, [42] show how to construct a sample-access-OI distinguisher $A_{L,g,h}$ s.t. if a predictor $\tilde{p}$ is OI w.r.t the distringuisher, then applying f to $\tilde{p}$ gives a predictor whose loss is competitive with h (f is the post-processing function for which L is a proper loss function). This is the crux of the proof of the reduction, and a powerful demonstration of the power of the Outcome Indistinguishability framework. With this reduction in place, multi-PAC learning can be performed using any OI learning algorithm (e.g. the algorithm of Theorem 2): i.e., by learning a predictor $\tilde{p}$ that is OI w.r.t. the class of distinguishers $(A_{L,g,h})_{g \in \mathcal{G}, h \in \mathcal{H}}$. The predictor $\tilde{h}(x) = f(x, \tilde{p}(x))$ will be competitve with $\mathcal{H}$ for all groups $g \in \mathcal{G}$ simultaneously. The heart of the argument is in constructing the distinguishers:

Lemma 11 (Loss Minimization via OI[42]). *Let L be an f-proper loss function that has the uniform convergence property. For a predictor $\tilde{p}$, define the hypothesis $\tilde{h}(x) = f(x, \tilde{p}(x))$.*

Let $\mathcal{D}$ be a distribution, $g \subseteq \mathcal{X}$ a subgroup s.t. $\mathcal{D}_{\mathcal{X}}[g] \geq \gamma$, $h : \mathcal{X} \to [0,1]$ a hypothesis, and $\alpha \in [0,1]$ a desired error parameter. There exists a multi-sample

sample-access-OI distinguisher $A_{L,g,h}$ s.t. if $\tilde{p}$ is $(\{A_{L,g,h}\}, \Theta(\alpha))$-sample-access-OI then:

$$L_{\mathcal{D}_g}(\tilde{h}) \leq L_{\mathcal{D}_g}(h) + \alpha.$$

The distinguisher $A_{L,g,h}$ operates on $k = \tilde{O}((m_L^{UC}(\Theta(\alpha), \Theta(\alpha), 1))/\gamma)$ samples (where m_L^{UC} is the sample complexity for uniform convergence). Its complexity is polynomial in k, in the complexity of determining group membership in g, and in the complexity of the classifier h.

Proof. We want to guarantee that the loss of the hypothesis $\tilde{h}(x) = f(x, \tilde{p}(x))$ is competitive with the loss of h, where both losses are measured on the distribution $\mathcal{D}_g$ over members of the group g. We begin by observing that this is true when the labels are drawn by $\tilde{p}(x)$ (as in the distribution $\tilde{\mathcal{D}}$). We will use OI (with an appropriately constructed distinguisher) to ensure that it is also true for the "real" distribution $\mathcal{D}_g$.

In more detail, since L is an f-proper loss function, we have:

$$L_{\tilde{\mathcal{D}}_g}(\tilde{h}) \leq L_{\tilde{\mathcal{D}}_g}(h),$$

because in $\tilde{\mathcal{D}}$ the labels are indeed generated by $\tilde{p}$, i.e. $\tilde{p}(x) = E_{\tilde{\mathcal{D}}}[y|x]$. By uniform convergence, this will remain true—up to an additive $\Theta(\alpha)$ slack—even if we consider the empirical loss over a (sufficiently large) i.i.d. sample from $\tilde{\mathcal{D}}_g$. We now define the distinguisher $A_{L,g,h}$, which takes as input k samples $\{(x_i, y_i, \tilde{p}_i)\}$ and checks whether, for the samples where $x_i \in g$, it is true that the loss obtained by predicting $f(x_i, \tilde{p}_i)$ for each x_i is competitive with the loss obtained by h on those samples (up to an additive factor of $\Theta(\alpha)$). By the above discussion, when the outcomes y_i are drawn by $\mathrm{Ber}(\tilde{p}_i)$, and assuming that there are sufficiently many samples in g to guarantee uniform convergence for the loss L, the distinguisher will accept with high probability.

Now, if $\tilde{p}$ is OI w.r.t. the distinguisher $A^k_{g,h,\alpha}$, then the distinguisher should accept with similar probabilities whether the labeled examples are drawn by $\tilde{\mathcal{D}}$ or by $\mathcal{D}$ (where in both cases the predictions are by $\tilde{p}_i$). I.e., $A_{L,g,h}$ should also accept w.h.p. when the examples are drawn by $\mathcal{D}$. By uniform convergence, this can only happen if the predictor $\tilde{h}$ is competitive with the hypothesis h w.r.t. the distribution $\mathcal{D}_g$: exactly the guarantee we wanted from $\tilde{h}$!

The above reduction, together with the OI learning algorithm of Theorem 2, gives the multi-group agnostic learning algorithm of Theorem 10. The sample complexity of the learning algorithm is governed by the sample complexity of OI learning, which is logarithmic in the number of distinguishers. The reduction includes $|\mathcal{G}| \cdot |\mathcal{H}|$ multi-sample distinguishers. The OI learning algorithm can be modified to handle multi-sample distinguishers, or we can further reduce (a class of) multi-sample distinguishers to (a class of) single-sample distinguishers using a hybrid argument. This all results in sample complexity that is logarithmic in $|\mathcal{G}|$ and in $|\mathcal{H}|$. We note that we need $\mathcal{G}$ and $\mathcal{H}$ to be finite because the known OI learning algorithm works for finite collections of distinguishers.

Even more general losses. Rothblum and Yona [42] separate the questions of multi-group *feasability*: i.e. does a multi-group predictor always exist for a given loss function, from the question of *learnability*. They show a loose characterization of the loss functions for which multi-PAC learning is feasible, and use the connection to OI to construct a learning algorithm for *any* such loss function that also satisfies uniform convergence.

4 Recent Developments

Several recent works have refined, developed and extended the Outcome Indistinguishability and multi-calibration frameworks. The literature has been growing rapidly—we briefly mention some notable examples. Gupta *et al.* [29] consider real-valued predictions and the meaningfulness of the predictor's confidence intervals and moments. As noted above, the study was extended to large outcome spaces in [16], see also [28]. Dwork *et al.* [17] show connections between the literature on multi-calibration and Outcome Indistinguishability, regularity in graph theory and the leakage simulation lemma in cryptography.

An emerging and exciting body of work shows that multi-calibration and Outcome Indistinguishability open the door to machine learning that is quite flexible and robust. An *omni-predictor*, as proposed and studied by Gopalan *et al.* [27], is a single predictor that can be trained once and then adapted to different loss functions. They show that multi-calibration for a collection of sets implies omni-prediction w.r.t. a hypothesis class that is directly related to the collection of sets, and a broad range of loss functions. A similar statement holds for OI, because of the equivalence between OI and multicalibration. We view this as further demonstration of the power and flexibility of the multi-calibration and Outcome Indistinguishability frameworks. Subsequent works (e.g. [26]) sharpen this connection, and use it in the context of optimization under fairness constraints [32]. At a very high level, these results leverage properties of OI (or multicalibration) that are similar in spirit to the "loss minimization to OI" reduction of Lemma 11. There are differences in the types of loss functions that are considered, but the main difference is on the conceptual level: the focus in omni-prediction is on training a predictor that can later be used to handle many loss functions, whereas [42] only use the reduction in the context of a fixed loss function.

Several other works leverage multi-calibration or OI to achieve robustness or adaptability to changes that might be encountered after training. The work of Kim *et al.* [38] on *universal adaptability* shows this in the context of propensity scoring in statistical analysis, where the goal is adapting an analysis to a new target population. Diana *et al.* [10] show a result of this flavor for downstream post-processing of predictions, whereas Kim and Perdomo [39] consider a prediction setting where individuals might exhibit performative behavior.

Finally, Outcome Indistinguishability aims to obtain predictions that cannot be refuted based on real-world outcome data. The real world itself, however, does not treat all demographic groups similarly. In recent work with Dwork and

Reingold [18], we consider corrective transformations τ that aim to map probabilities p^* in the real world to a better world $\tau(p^*)$. We study the goal of learning a predictor that is indistinguishable from the better world, and characterize the transformations for which this goal is achievable.

5 Acknowledgements

This overview is based on the works [15,42] and borrows liberally from those works. We are indebted to our co-authors and collaborators Cynthia Dwork, Michael Kim, Omer Reingold and Gal Yona for many wonderful and illuminating discussions.

References

1. Blum, A., Lykouris, T.: Advancing subgroup fairness via sleeping experts. arXiv preprint arXiv:1909.08375 (2019)
2. Blum, A., Stangl, K.: Recovering from biased data: Can fairness constraints improve accuracy? arXiv preprint arXiv:1912.01094 (2019)
3. Buja, A., Stuetzle, W., Shen, Y.: Loss functions for binary class probability estimation and classification: Structure and applications. Working draft, November **3** (2005)
4. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Friedler, S.A., Wilson, C. (eds.) Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA. Proceedings of Machine Learning Research, vol. 81, pp. 77–91. PMLR (2018), http://proceedings.mlr.press/v81/buolamwini18a.html
5. Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)
6. Chung, K., Lui, E., Pass, R.: Can theories be tested?: a cryptographic treatment of forecast testing. In: Kleinberg, R.D. (ed.) Innovations in Theoretical Computer Science, ITCS '13, Berkeley, CA, USA, January 9-12, 2013. pp. 47–56. ACM (2013). https://doi.org/10.1145/2422436.2422443, https://doi.org/10.1145/2422436.2422443
7. Dawid, A.P.: The well-calibrated bayesian. Journal of the American Statistical Association **77**(379), 605–610 (1982)
8. Dawid, A.: Objective probability forecasts'. Tech. rep., Research Report 14, Department of Statistical Science, University College London (1982)
9. Dawid, P.: On individual risk. Synthese **194**(9), 3445–3474 (Nov 2015). https://doi.org/10.1007/s11229-015-0953-4, http://dx.doi.org/10.1007/s11229-015-0953-4
10. Diana, E., Gill, W., Kearns, M., Kenthapadi, K., Roth, A., Sharifi-Malvajerdi, S.: Multiaccurate proxies for downstream fairness. In: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022. pp. 1207–1239. ACM (2022). https://doi.org/10.1145/3531146.3533180, https://doi.org/10.1145/3531146.3533180
11. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)

12. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd innovations in theoretical computer science conference. pp. 214–226 (2012)
13. Dwork, C., Immorlica, N., Kalai, A.T., Leiserson, M.: Decoupled classifiers for fair and efficient machine learning. arXiv preprint arXiv:1707.06613 (2017)
14. Dwork, C., Kim, M.P., Reingold, O., Rothblum, G.N., Yona, G.: Learning from outcomes: Evidence-based rankings. In: 2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS). pp. 106–125. IEEE (2019)
15. Dwork, C., Kim, M.P., Reingold, O., Rothblum, G.N., Yona, G.: Outcome indistinguishability. In: Khuller, S., Williams, V.V. (eds.) STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021. pp. 1095–1108. ACM (2021). https://doi.org/10.1145/3406325.3451064
16. Dwork, C., Kim, M.P., Reingold, O., Rothblum, G.N., Yona, G.: Beyond bernoulli: Generating random outcomes that cannot be distinguished from nature. In: Dasgupta, S., Haghtalab, N. (eds.) International Conference on Algorithmic Learning Theory, 29-1 April 2022, Paris, France. Proceedings of Machine Learning Research, vol. 167, pp. 342–380. PMLR (2022), https://proceedings.mlr.press/v167/dwork22a.html
17. Dwork, C., Lee, D., Lin, H., Tankala, P.: New insights into multi-calibration. CoRR **abs/2301.08837** (2023). https://doi.org/10.48550/arXiv.2301.08837
18. Dwork, C., Reingold, O., Rothblum, G.N.: From the real towards the ideal: Risk prediction in a better world (2023)
19. Fortnow, L., Vohra, R.V.: The complexity of forecast testing. Econometrica **77**(1), 93–105 (2009). https://doi.org/https://doi.org/10.3982/ECTA7163
20. Foster, D.P., Vohra, R.V.: Asymptotic calibration. Biometrika **85**(2), 379–390 (1998)
21. Fudenberg, D., Levine, D.K.: An easier way to calibrate. Games and economic behavior **29**(1-2), 131–137 (1999)
22. Globus-Harris, I., Kearns, M., Roth, A.: An algorithmic framework for bias bounties. In: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022. pp. 1106–1124. ACM (2022). https://doi.org/10.1145/3531146.3533172
23. Goldreich, O.: Foundations of Cryptography: Volume 1, Basic Tools. Cambridge University Press, USA (2006)
24. Goldreich, O.: Computational Complexity: A Conceptual Perspective. Cambridge University Press, USA, 1 edn. (2008)
25. Goldreich, O.: Foundations of Cryptography: Volume 2, Basic Applications. Cambridge University Press (2009)
26. Gopalan, P., Hu, L., Kim, M.P., Reingold, O., Wieder, U.: Loss minimization through the lens of outcome indistinguishability. In: Kalai, Y.T. (ed.) 14th Innovations in Theoretical Computer Science Conference, ITCS 2023, January 10-13, 2023, MIT, Cambridge, Massachusetts, USA. LIPIcs, vol. 251, pp. 60:1–60:20. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2023). https://doi.org/10.4230/LIPIcs.ITCS.2023.60
27. Gopalan, P., Kalai, A.T., Reingold, O., Sharan, V., Wieder, U.: Omnipredictors. In: Braverman, M. (ed.) 13th Innovations in Theoretical Computer Science Conference, ITCS 2022, January 31 - February 3, 2022, Berkeley, CA, USA. LIPIcs, vol. 215, pp. 79:1–79:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2022). https://doi.org/10.4230/LIPIcs.ITCS.2022.79

28. Gopalan, P., Kim, M.P., Singhal, M., Zhao, S.: Low-degree multicalibration. In: Loh, P., Raginsky, M. (eds.) Conference on Learning Theory, 2-5 July 2022, London, UK. Proceedings of Machine Learning Research, vol. 178, pp. 3193–3234. PMLR (2022), https://proceedings.mlr.press/v178/gopalan22a.html
29. Gupta, V., Jung, C., Noarov, G., Pai, M.M., Roth, A.: Online multivalid learning: Means, moments, and prediction intervals. In: Braverman, M. (ed.) 13th Innovations in Theoretical Computer Science Conference, ITCS 2022, January 31 - February 3, 2022, Berkeley, CA, USA. LIPIcs, vol. 215, pp. 82:1–82:24. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2022). https://doi.org/10.4230/LIPIcs.ITCS.2022.82
30. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in neural information processing systems. pp. 3315–3323 (2016)
31. Hébert-Johnson, Ú., Kim, M.P., Reingold, O., Rothblum, G.: Multicalibration: Calibration for the (computationally-identifiable) masses. In: International Conference on Machine Learning. pp. 1939–1948 (2018)
32. Hu, L., Navon, I.L., Reingold, O., Yang, C.: Omnipredictors for constrained optimization. CoRR **abs/2209.07463** (2022). https://doi.org/10.48550/arXiv.2209.07463
33. Jung, C., Lee, C., Pai, M.M., Roth, A., Vohra, R.: Moment multicalibration for uncertainty estimation. arXiv preprint arXiv:2008.08037 (2020)
34. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In: International Conference on Machine Learning. pp. 2564–2572 (2018)
35. Kearns, M., Neel, S., Roth, A., Wu, Z.S.: An empirical study of rich subgroup fairness for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. pp. 100–109 (2019)
36. Kearns, M.J., Schapire, R.E., Sellie, L.M.: Toward efficient agnostic learning. Machine Learning **17**(2-3), 115–141 (1994)
37. Kim, M.P., Ghorbani, A., Zou, J.: Multiaccuracy: Black-box post-processing for fairness in classification. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 247–254 (2019)
38. Kim, M.P., Kern, C., Goldwasser, S., Kreuter, F., Reingold, O.: Universal adaptability: Target-independent inference that competes with propensity scoring. Proceedings of the National Academy of Sciences **119**(4), e2108097119 (2022). https://doi.org/10.1073/pnas.2108097119
39. Kim, M.P., Perdomo, J.C.: Making decisions under outcome performativity. In: Kalai, Y.T. (ed.) 14th Innovations in Theoretical Computer Science Conference, ITCS 2023, January 10-13, 2023, MIT, Cambridge, Massachusetts, USA. LIPIcs, vol. 251, pp. 79:1–79:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2023). https://doi.org/10.4230/LIPIcs.ITCS.2023.79
40. Kim, M.P., Reingold, O., Rothblum, G.N.: Fairness through computationally-bounded awareness. Advances in Neural Information Processing Systems (2018)
41. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807 (2016)
42. Rothblum, G.N., Yona, G.: Multi-group agnostic PAC learnability. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 9107–9115. PMLR (2021), http://proceedings.mlr.press/v139/rothblum21a.html
43. Sandroni, A.: The reproducible properties of correct forecasts. International Journal of Game Theory **32**(1), 151–159 (2003)

44. Sandroni, A., Smorodinsky, R., Vohra, R.V.: Calibration with many checking rules. Mathematics of operations Research **28**(1), 141–153 (2003)
45. Shabat, E., Cohen, L., Mansour, Y.: Sample complexity of uniform convergence for multicalibration. arXiv preprint arXiv:2005.01757 (2020)
46. Vadhan, S.P.: Pseudorandomness. Now Publishers Inc., Hanover, MA, USA (2012)
47. Zhao, S., Ma, T., Ermon, S.: Individual calibration with randomized forecasting. arXiv preprint arXiv:2006.10288 (2020)