Outcome Indistinguishability

Cynthia Dwork*
Harvard University
Cambridge, MA, USA
dwork@seas.harvard.edu

Michael P. Kim†
UC Berkeley
Berkeley, CA, USA
mpkim@berkeley.edu

Omer Reingold‡
Stanford Unviersity
Stanford, CA, USA
reingold@stanford.edu

Guy N. Rothblum§
Weizmann Institute of Science
Rehovot, Israel
rothblum@alum.mit.edu

Gal Yona¶
Weizmann Institute of Science
Rehovot, Israel
gal.yona@weizmann.ac.il

ABSTRACT

Prediction algorithms assign numbers to individuals that are popularly understood as individual "probabilities"—what is the probability of 5-year survival after cancer diagnosis?—and which increasingly form the basis for life-altering decisions. Drawing on an understanding of computational indistinguishability developed in complexity theory and cryptography, we introduce *Outcome Indistinguishability*. Predictors that are Outcome Indistinguishable yield a generative model for outcomes that cannot be efficiently refuted on the basis of the real-life observations produced by Nature.

We investigate a hierarchy of Outcome Indistinguishability definitions, whose stringency increases with the degree to which distinguishers may access the predictor in question. Our findings reveal that Outcome Indistinguishability behaves qualitatively differently than previously studied notions of indistinguishability. First, we provide constructions at all levels of the hierarchy. Then, leveraging recently-developed machinery for proving average-case fine-grained hardness, we obtain lower bounds on the complexity of the more stringent forms of Outcome Indistinguishability. This hardness result provides the first scientific grounds for the political argument that, when inspecting algorithmic risk prediction instruments, auditors should be granted oracle access to the algorithm, not simply historical predictions.

CCS CONCEPTS

• **Theory of computation** → **Machine learning theory**; **Pseudorandomness and derandomization**.

KEYWORDS

Computational Indistinguishability, Fairness, Prediction

ACM Reference Format:
Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. 2021. Outcome Indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC '21), June 21–25, 2021, Virtual, Italy*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3406325.3451064

1 INTRODUCTION

Prediction algorithms "score" individuals, mapping them to numbers in $[0, 1]$ that are popularly understood as "probabilities" or "likelihoods:" 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, or that it will rain tomorrow. Algorithmic risk predictions increasingly inform consequential decisions, but what can these numbers really mean? Five-year survival, four-year graduation, and rain tomorrow are not repeatable events. The question of "individual probabilities" has been studied for decades across many disciplines without clear resolution.[1]

One interpretation relies on the coarseness of the representation of individuals to the prediction algorithm—the shape of a tumor's boundaries and the age of the patient; the student's grades, test scores, and a few bits about the family situation; the atmospheric

*This work supported in part by the Radcliffe Institute for Advanced Study at Harvard, Microsoft Research, the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, the Sloan Foundation Grant 2020-13941, and NSF CCF-1763665.
†Supported by the Miller Institute for Basic Research in Science. Part of this work completed at Stanford University, supported by NSF Award IIS-1908774.
‡Research supported by the Simons Foundation Collaboration on the Theory of Algorithmic Fairness, the Sloan Foundation Grant 2020-13941, Microsoft Research, and by NSF Award CCF-1763311.
§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), from the U.S.-Israel Binational Science Foundation (grant 2018102), and from the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. Part of this work was done while the author was 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. This research was also partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center and by a research grant from Madame Olga Klein–Astrachan.

STOC '21, June 21–25, 2021, Virtual, Italy

ACM ISBN 978-1-4503-8053-9/21/06…$15.00
https://doi.org/10.1145/3406325.3451064

[1] See the inspiring paper, and references therein, of Philip Dawid [8], discussing several notions of *individual risk* based on different philosophical understandings of probability "including Classical, Enumerative, Frequency, Formal, Metaphysical, Personal, Propensity, Chance and Logical conceptions of Probability" and proposing a new approach to characterizing individual risk which, the author concludes, remains elusive.

pressure, humidity level, and winds—to partition individuals into a small number of "types." This leads to a natural interpretation of the predictions: *amongst the individuals of this type, what fraction exhibit a positive outcome?* In the context of modern data science, however, it is typical to make predictions based on a large number of expressive measurements for each individual—the patient's genome-wide risk factors; the borrower's online transactions and browsing data; the student's social media connections. In this case, when each individual resolves to a unique set of covariates, the frequency-based interpretation fails.

Another view imagines the existence of a party—Nature—who selects, for each individual, a probability distribution over outcomes; then, the realized outcome is determined by a draw from this distribution. Note that Nature may select the outcomes using complete determinism (i.e., probabilities in $\{0, 1\}$). This view of the world gives rise to a statistical model with well-defined individual probabilities, but reasoning about these probabilities from observational data presents challenges. Given only observations of outcomes, we cannot even determine whether Nature assigns integer or non-integer probabilities. Perhaps Nature is deterministic, but we do not have enough information or computing resources to carry out the predictions ourselves. Thus, even if we posit that outcomes are determined by individual probabilities, we cannot hope to recover the exact probabilities governing Nature, so this abstraction does not appear to provide an effective avenue for understanding the meanings of algorithmic risk scores.

1.1 Predictions That Withstand Observational Falsifiability

Given the philosophical uncertainty regarding the very existence of randomness, we explore the criteria for an ideal predictor. We can view the outputs of a prediction algorithm as defining a generative model for observational outcomes. Ideally, the outcomes from this generative model should "look like" the outcomes produced by Nature. To this end, we introduce and study a strong notion of faithfulness—*Outcome Indistinguishability (OI)*. 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. Our study contributes a computational-theoretic 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 [7, 14, 16, 37, 38])—and how they should be used. For example, one of our results provides 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 oracle access to the algorithms should be granted for auditing purposes.

Outcome Indistinguishability presents a broad framework evaluating algorithmic risk predictions. This paper focuses on the fundamental setting of predicting a binary outcome, given an individual's covariates, a simple prediction setup that already highlights many of the challenges and subtleties that arise while defining and reasoning about OI. Nothing precludes extending OI to reason about algorithms that make predictions about more general outcomes. Due to page limits, this conference version of the manuscript represents a technical overview of the work. A full version of the manuscript can be found at [11], which the authors recommend.

Basic notation. We assume that individuals are selected from some discrete domain $\mathcal{X}$, for example, the set of d-bit strings[2]. We model Nature as a joint distribution, denoted $\mathcal{D}^*$, over individuals and outcomes, where $o_i^* \in \{0, 1\}$ represents Nature's choice of outcome for individual $i \in X$. We use $i \sim \mathcal{D}_{\mathcal{X}}$ to denote a sample from Nature's marginal distribution over individuals and denote by $p_i^* \in [0, 1]$ the conditional probability that Nature assigns to the outcome o_i^*, conditioned on i. We emphasize, however, that Nature may choose $p_i^* \in \{0, 1\}$ to be deterministic; our definitions and constructions are agnostic as to this point.

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

Outcome Indistinguishability. Imagine that Nature selects $p_i^* = 1$ for half of the mass of $i \sim \mathcal{D}_{\mathcal{X}}$ and $p_i^* = 0$ 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 $i \sim \mathcal{D}_{\mathcal{X}}$, no efficient observer can guess if $p_i^* = 1$ or $p_i^* = 0$ with probability significantly better than $1/2$. In this case, producing the estimates $\tilde{p}_i = 1/2$ *for every individual* $i \in \mathcal{X}$ 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.

In the most basic form of the definition, a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$ is Outcome Indistinguishable with respect to a family of distinguishers $\mathcal{A}$ if samples from Nature's distribution $(i, o_i^*) \sim \mathcal{D}^*$ cannot be distinguished by $\mathcal{A}$ from samples from the predictor's distribution $(i, \tilde{o}_i) \sim \mathcal{D}(\tilde{p})$, meaning that for each algorithm $A \in \mathcal{A}$, the probability that A outputs 1 is (nearly) the same on samples from each of the two distributions, $\mathcal{D}^*$ and $\mathcal{D}(\tilde{p})$.

Definition (Outcome Indistinguishability). *Fix Nature's distribution $\mathcal{D}^*$. For a class of distinguishers[3] $\mathcal{A}$ and $\varepsilon > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0, 1]$ satisfies $(\mathcal{A}, \varepsilon)$-outcome indistinguishability (OI) if*

[2]Individuals can be arbitrarily complex; they are represented to the algorithm as elements of $\mathcal{X}$. Strictly speaking, distributions over $\mathcal{X}$ are induced distributions over the representations, and our results apply whether or not there are collisions. We do not assume that Nature's view is restricted to the representation.

[3]Like the predictors, distinguishers have access only to the finite *representations* of individuals as elements of $\mathcal{X}$.

for every $A \in \mathcal{A}$,

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

The definition of Outcome Indistinguishability can be extended in many ways, for example to settings where distinguishers receive multiple samples from each distribution, or when they have access to the program implementing $\tilde{p}$, and to the case of non-Boolean outcomes.

In the extreme, when we think of $\mathcal{A}$ as the set of all efficient computations, 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 seems to give an elegant computational perspective on the scientific method, when consider $\tilde{p}$ as expressing a hypothesis that cannot be falsified through observational investigation.

1.2 Our Contributions

The most significant contributions of this work can be summarized as follows:

(1) We define a practically-motivated four-level hierarchy of increasingly demanding notions of *Outcome Indistinguishability*. The levels of the hierarchy arise by varying the degree to which the distinguishers may access the predictive model in question.
(2) We provide tight connections between the two lower levels of the hierarchy to *multi-accuracy* and *multi-calibration*, two notions defined and studied in [25]. Establishing this connection immediately gives algorithmic constructions for these two levels.
(3) We describe a novel algorithm that constructs OI predictors directly. This construction establishes an upper bound on the complexity of OI for the upper levels of the hierarchy (and, consequently, also allows us to recover the results of [25] through the OI framework).
(4) We show a *lower bound* for the upper levels of the hierarchy, demonstrating the tightness of our constructions. We prove that, under plausible complexity-theoretic assumptions, at the top two levels of the hierarchy, the complexity of implementing OI predictors cannot scale polynomially in the complexity of the distinguishers in $\mathcal{A}$ and in the distinguishing advantage $1/\varepsilon$.

Additionally, we revisit the apparent interchangeability of the terms "test" and "distinguisher" in the literature on pseudorandomness, drawing a distinction that is relevant to the forecasting problem. Our reults clarify the mathematical relationship between notions in the two literatures.

Next, we present a colloquial illustration of the different notions of the hierarchy. While very natural, the notions within the hierarchy have not been fully considered in the literature on either forecasting or pseudorandomness.

The Outcome Indistinguishability Hierarchy. 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 $(i, o_i^*) \sim \mathcal{D}^*$ from those sampled according to the predicted distribution $(i, \tilde{o}_i) \sim \mathcal{D}(\tilde{p})$. The board gets no direct access to predictions $\tilde{p}_i$ of the program; we call this variant *no-access-OI*.
(2) Naturally, the board may ask to see the predictions $\tilde{p}_i$ for each sampled individual. In this extension—*sample-access-OI*—the board must distinguish samples of the form $(i, o_i^*, \tilde{p}_i)$ and $(i, \tilde{o}_i, \tilde{p}_i)$, again for $(i, o_i^*) \sim \mathcal{D}^*$ and $(i, \tilde{o}_i) \sim \mathcal{D}(\tilde{p})$.
(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 $\tilde{p}$.

On a different axis, we also consider multiple-sample variants of OI and show how these relate to the single-sample variants described above. Multiple-sample OI is closer to the problem of online *forecasting* (*e.g.*, daily weather forecasting); we explore connections between this variant of OI and the forecasting literature in Section 1.4.

The Lower Levels of the OI Hierarchy. We begin by examining the relationship between the different levels of the hierarchy. We show that no-access-OI and sample-access-OI are closely related to the notions of multi-accuracy and multi-calibration [25], respectively, studied in the algorithmic fairness literature. Very loosely, for a collection C of subpopulations of individuals, (C, α)-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 $S \in C$. Here, calibration over S means that if we restrict our attention to individuals $i \in S$ for which $\tilde{p}_i = v$, then the fraction individuals with positive outcomes (i.e., $i \in S$ such that $o_i^* = 1$) is roughly v. We prove that sample-access-OI with respect to a set of distinguishers $\mathcal{A}$ is "equivalent" to C-multi-calibration in the sense that each notion can enforce the other, for closely related classes C and $\mathcal{A}$.

Theorem 1 (Informal). *For any class of distinguishers $\mathcal{A}$ and $\varepsilon > 0$, there exists a (closely related) collection of subpopulations $C_{\mathcal{A}}$ and $\alpha_\varepsilon > 0$, such that $(C_{\mathcal{A}}, \alpha_\varepsilon)$-multi-calibration implies $(\mathcal{A}, \varepsilon)$-sample-access-OI. Similarly, for any collection of subpopulations C and $\alpha > 0$, there exists a (closely related) class of distinguishers $\mathcal{A}_C$ and $\varepsilon_\alpha > 0$, such that $(\mathcal{A}_C, \varepsilon_\alpha)$-sample-access-OI implies C-multi-calibration.*

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 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 $S \in C$, rather than calibration.

Feasibility and Constructions. We consider the question of whether efficient OI predictors always exist. In particular, we ask, *Can we bound the complexity of OI predictors, independently of the complexity of Nature's distribution?* The picture we uncover is subtle; we will see that Outcome Indistinguishability differs qualitatively from prior notions of indistinguishability.

First off, leveraging feasibility results for the fairness notions from [25], 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. *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 $O(s/\varepsilon^2)$.*

Turning now to oracle-access-OI and code-access-OI predictors, we obtain a general-purpose algorithm for constructing OI predictors, even when the distinguishers are allowed arbitrary access to the predictor in question. This construction extends the learning algorithm for multi-calibration of [25] 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. In this introduction, we state the bound informally, assuming the distinguishers are implemented by circuits with oracle gates. 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). *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 $s \cdot q^{O(1/\varepsilon^2)}$.*

Intuitively, code-access-OI can implement any of the prior levels through simulation: given the code for $\tilde{p}$, the distinguishers can execute oracle calls (or calls to $\tilde{p}_i$) whenever needed.At the extreme of efficient OI, we consider code-access-OI with respect to the class of polynomial-sized distinguishers. Importantly, we allow the complexity of these distinguishers to grow as a (fixed) polynomial in both the dimension of individuals d and the length of the description of the predictor $\tilde{p}$, which we denote by n. For this most general version of OI, the complexity may scale doubly exponentially in $\mathrm{poly}(1/\varepsilon)$; nevertheless, the bound is independent of the complexity of p^*.

Theorem 4 (Informal). *For some $d \in \mathbb{N}$, let $\mathcal{X} \subseteq \{0,1\}^d$ be represented by d-bit strings. Suppose for some $k \in \mathbb{N}$, $\mathcal{A}$ is a class of distinguishers implemented by circuits of size $(d+n)^k$, on inputs $i \in \mathcal{X}$ and descriptions of predictors in $\{0,1\}^n$. For any $\mathcal{D}^*$ and $\varepsilon > 0$, there exists a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ satisfying $(\mathcal{A}, \varepsilon)$-code-access-OI implemented by a circuit of size $d^{2^{O(1/\varepsilon^2)}}$.*

Hardness via Fine-Grained Complexity. We establish 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 $n^{\Omega(k)}$, we demonstrate that the construction of Theorem 3 is optimal up to polynomial factors. Specifically, we 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. This 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 5 (Informal). *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 $\tilde{O}(d)$ queries, 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. We note that, under the plausible assumption that the fine-grained complexity of known clique counting algorithms is tight, our construction 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^*.

Hardness via BPP $\neq$ PSPACE. We 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.

Theorem 6 (Informal). *Assume that* BPP $\neq$ PSPACE. *Then, there exist: $\mathcal{X} \subseteq \{0,1\}^d$, a distribution $\mathcal{D}^*$ (which can be sampled in* $\exp(\mathrm{poly}(d))$ *time), and a class $\mathcal{A}$ of* $\mathrm{poly}(d)$ *distinguishers that run in time* $\mathrm{poly}(d)$*, s.t. for $\varepsilon = \frac{1}{\mathrm{poly}(d)}$, no predictor $\tilde{p}$ that runs in time* $\mathrm{poly}(d)$ *can satisfy $(\mathcal{A}, \varepsilon)$-oracle-access-OI.*

Discussion. We highlight a few possible interpretations and insights that stem from our technical results. The ability to construct predictors that satisfy outcome indistinguishability can be viewed both positively and negatively. On the 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 detected by efficient inspection.

More generally, the computational perspective of OI underscores an inherent limitation in trying to recover the exact laws governing Nature from observational data alone. We illustrate this perspective through a comparison to pseudorandomness. Traditionally in pseudorandomness, our object of desire is *random* (e.g., a large string of random bits fed to a BPP algorithm), and we show that a simple *deterministic* object suffices to "fool" efficient observers. In outcome indistinguishability, our object of desire is a model of Nature, which may obey highly-complex *deterministic* laws. In this work, we show that a simple *random* model of Nature—namely, $\mathcal{D}(\tilde{p})$ for an OI predictor $\tilde{p}$—suffices to "fool" efficient observers. 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 random 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, where the complexity of the indistinguishable object is usually polynomial in the distinguishers' complexity and distinguishing advantage, regardless of the complexity of the object we are trying to imitate.

The increased distinguishing power of oracle access to the predictor in oracle-access-OI seems to have practical implications. Currently, there are many conversations about the appropriate usage of algorithms when making high-stakes judgments about members of society, for instance 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.

1.3 Technical Overview

Next, we give a technical overview of the main results. Our goal is to convey the intuition for our findings, deferring the technical details to subsequent sections. For a full account of the results, see the full version [11].

Relating OI and multi-calibration. To build intuition for the equivalence, as described informally in Theorem 1, we begin by describing the construction that establishes the lower level of the equivalence, between multi-accuracy and no-access-OI (distinguishers that do not directly observe $\tilde{p}_i$). Informally, for a collection of subpopulations C, multi-accuracy guarantees that the expectations of p^*_i and $\tilde{p}_i$ are approximately the same, even when conditioning on the event that $i \in S$ (simultaneously for every $S \in C$).

Given a subpopulation S, we define the multi-accuracy violation $\nabla_S(\tilde{p})$ to be the absolute value of the above difference in conditional expectations. This can be viewed as a direct analogue of the distinguishing advantage $\Delta_A(\tilde{p})$ (the absolute difference between the acceptance probability of A on a sample from $\mathcal{D}^*$ vs $\mathcal{D}(\tilde{p})$). To translate between the notions, we define two mappings (subpopulations to distinguishers, and distinguishers to subpopulations) that allow us to upper-bound one quantity in terms of the other. Specifically, given a collection of subpopulations C, we define a family of distinguishers $\mathcal{A} = \{A_S\}$, where for all $S \in C$, $A_S(i, b) = \mathbf{1}[\, i \in S \wedge b = 1 \,]$.

Note that the probability that each A_S accepts on a sample $(i, \tilde{o}_i) \sim \mathcal{D}(\tilde{p})$ is the joint probability that $i \in S$ and $\tilde{o}_i = 1$, which can be directly related to the expectation of $\tilde{p}_i$ conditioned on $i \in S$. This allows us to express the multi-accuracy violation in terms of the distinguishing advantage, as $\nabla_S(\tilde{p}) = \frac{\Delta_{A_S}(\tilde{p})}{\mathbf{Pr}_{i\sim\mathcal{D}_X}[i\in S]}$. By taking the accuracy parameter in no-access-OI to be sufficiently small, we can guarantee that no-access-OI implies multi-accuracy. Similarly, to implement outcome indistinguishability from multi-accuracy, we define two sets $S_{(A,0)}$ and $S_{(A,1)}$ for every distinguisher $A \in \mathcal{A}$ and $b \in \{0, 1\}$, $S_{(A,b)} = \{i \in \mathcal{X} : A(i, b) = 1\}$.

Using similar arguments, we show that $\Delta_A(\tilde{p})$ can be upper-bounded by $\nabla_{S_{(A,0)}}(\tilde{p}) + \nabla_{S_{(A,1)}}(\tilde{p})$, for which multi-accuracy (w.r.t the constructed family of subpopulations) provides a bound. Note that the constructions are very closely related; in fact, repeating the translation twice reaches a "stable point". That is, given C (or given $\mathcal{A}$), we can construct a canonical pair $(C', \mathcal{A}')$ such that C'-multi-accuracy implies $\mathcal{A}'$-no-access-OI, and vice versa. Importantly, C' is (essentially) of the same complexity as C (resp., $\mathcal{A}'$ compared to $\mathcal{A}$), and the degradation in the accuracy parameters only results from the fact that multi-accuracy is defined for a collection of arbitrarily small sets.

Showing a similar equivalence for multi-calibration follows the same general construction, but requires more care. We begin with the observation that for a fixed predictor $\tilde{p}$, C-multi-calibration can be viewed as $\tilde{C}$-multi-accuracy, where each subpopulation in $\tilde{C}$ is obtained as the intersection of some subpopoulation $S \in C$ and "level-set" of $\tilde{p}$: $\{i \in S \wedge \tilde{p}_i = v\}$. Thus, at an intuitive level, we can model the constructions similarly in terms of the sets in $\tilde{C}$. A number of technical subtleties arise due to the precise notion of approximate calibration from [25], which is necessary to provide sufficiently strong fairness guarantees.

Constructing OI predictors. We establish the complexity of OI predictors (as in Theorems 3 and 4) by describing a learning algorithm that, given a class of distinguishers $\mathcal{A}$, an approximation parameter ε, and samples from Nature's distribution $(i, o^*_i) \sim \mathcal{D}^*$, constructs a predictor satisfying outcome indistinguishability, for any level of the OI hierarchy. To start, inspired by the approach of [25], we consider a reduction from the task of constructing an OI predictor to auditing for OI. Specifically, the auditing problem receives a candidate predictor p, and must determine whether for all $A \in \mathcal{A}$, the distinguishing advantage $\Delta_A(p) < \varepsilon$ is small; if there is an $A \in \mathcal{A}$ that has nontrivial advantage in distinguishing $\mathcal{D}^*$ from $\mathcal{D}(p)$, then the auditor must return such a distinguisher. Naively, the auditor can be implemented by exhaustive search: for

each $A \in \mathcal{A}$, the auditor—using the samples from $\mathcal{D}^*$ as well as generated samples from $\mathcal{D}(p)$—evaluates the advantage of $\Delta_A(p)$, returning A if $\Delta_A(p) > \varepsilon$.

Suppose we're given some candidate predictor p; by iteratively auditing and updating, we aim to construct a circuit computing a predictor $\tilde{p}$ that satisfies OI. To start, given the predictor p, if the auditor certifies that p passes $(\mathcal{A}, \varepsilon)$-OI, then trivially we have succeeded in our construction. If, however, there is a distinguisher $A \in \mathcal{A}$ with a nontrivial advantage, then p fails the audit; in this case, the successful distinguisher $A \in \mathcal{A}$ witnesses some "direction" along which p fails to satisfy OI. If we can update along this direction, a standard potential argument (akin to that of boosting or gradient descent) demonstrates that the updated predictor has made "progress" towards satisfying OI. Then, we can recurse, calling the auditor on the updated predictor. We argue, the process must terminate with an OI predictor after not too many rounds of auditing and updating.

Thus, the crux of the construction is to solve the following problem: given a circuit computing a predictor p and a distinguisher $A \in \mathcal{A}$ that witnesses $\Delta_A(p) > \varepsilon$, derive a new circuit computing an updated predictor p' that has made progress towards OI. A subtle issue arises when making this intuition rigorous for oracle-access-OI and code-access-OI. At these levels of OI, the distinguishers may access the predictor in question, so there seems to be some circularity in the construction: to obtain the OI predictor $\tilde{p}$, we need to call the distinguishers $A \in \mathcal{A}$; but to evaluate the distinguishers $A \in \mathcal{A}$, we may need to access $\tilde{p}$. We argue that, in fact, there is no issue; to avoid the circularity, in each iteration, we can use the current predictor $p^{(t)}$ as the "oracle" for the distinguishers in $\mathcal{A}$. If $p^{(t)}$ passes auditing by oracle distinguishers $A^{p^{(t)}}$, then this predictor satisfies oracle-access-OI. If $p^{(t)}$ fails auditing, then we can still use the distinguisher $A^{p^{(t)}}$ to derive an update that we argue makes progress towards Nature's predictor p^*. Of course, because $\mathcal{D}^* = \mathcal{D}(p^*)$, p^* satisfies OI. Thus, the potential argument still works, and we guarantee termination after a bounded number of updates.[4]

To finish the construction, we leverage the concrete assumptions about the model of distinguishers to build up the circuit computing $\tilde{p}$. We focus on obtaining oracle-access-OI for size s oracle circuits that make at most q queries to $\tilde{p}$. The argument above ensures that in the t-th iteration, we can implement each oracle distinguisher using (non-oracle) circuits, where each of the q oracles calls is replaced with a copy of the current predictor $p^{(t)}$ hard-coded in place of the oracle gates. Then, we can derive an updated circuit $p^{(t+1)}$ by combining the circuits computing $p^{(t)}$ and computing $A^{p^{(t)}}$ (taking an addition of the outputs, with appropriate scaling). This recursive construction—where we build the circuit computing $p^{(t+1)}$ by incorporating multiple copies of $p^{(t)}$—suggests a recurrence relation characterizing an upper bound on the eventual circuit size. Intuitively, with a base circuit size of s, and q oracle calls (determining the branching factor), the size of $p^{(t)}$ grows roughly as $s \cdot q^t$. Leveraging an upper bound on the number of iterations $T = O(1/\varepsilon^2)$, the claimed bound follows.

[4] A similar argument holds for code-access-OI, using the description of $p^{(t)}$ as input.

Establishing the upper bound on the complexity of code-access-OI follows by a similar high-level argument, but there are some additional complications. Briefly, because the distinguishers take, as input, the description of a circuit computing the predictor in question, we need to work with a class of distinguishers that *grows with the complexity of the predictor itself*. We deal with the technicalities needed to encode and decode predictors so that we can simulate the lower levels within code-access-OI.

Lower bounds for oracle-access-OI.. To relate the complexity of evaluating oracle-access-OI predictors to complexity-theoretic assumptions such as the hardness of clique counting or PSPACE-complete problems (as done in the informal results stated in Theorems 5 and 6), we consider the task of constructing an OI predictor that needs to withstand the scrutiny of a distinguisher that can make oracle queries. Suppose we guarantee that any such predictor must compute a moderately hard function f on at least part of its domain. Then, a distinguisher could use oracle access to the predictor (on that part of the domain) to perform expensive computations of f at unit cost, while scrutinizing other parts of the domain. As we'll see, pushing this intuition, we show how efficient oracle distinguishers can perform surprisingly powerful tests to distinguish $\tilde{p}$ from p^*.

With this intuition in mind, we divide the domain $\mathcal{X}$ into m disjoint subsets $\mathcal{X}_1, \ldots, \mathcal{X}_m$. As a first step, we want to make sure that it is moderately hard to achieve OI when the distribution $\mathcal{D}^*$ is restricted to $\mathcal{X}_1$: for $i \in \mathcal{X}_1$, we set $o_i^* = f_1(i)$, where f_1 is a moderately hard Boolean function. We will guarantee that any oracle-access-OI predictor $\tilde{p}$ needs to compute f_1 on inputs in $\mathcal{X}_1$ by adding a distinguisher A_1 that verifies, for inputs $i \in \mathcal{X}_1$, that $o_i = f_1(i)$. At this point, it isn't clear that anything interesting is happening: the complexity of achieving OI (i.e., computing f_1) is not yet higher than the complexity of the distinguisher (which also needs to compute f_1). However, we can now add a distinguisher A_2 that verifies, for inputs in $\mathcal{X}_2$, that $\tilde{p}$ also correctly computes a harder function f_2. The key point is for the distinguisher to verify that $o_i = f_2(i)$ *without computing* f_2 *itself*! To achieve this, the distinguisher can use its oracle access to $\tilde{p}$. In particular, assuming that $\tilde{p}$ correctly computes f_1 on inputs $\mathcal{X}_1$, we can use a *downwards self-reduction* from computing f_2 on inputs in $\mathcal{X}_2$ to computing f_1 on inputs in $\mathcal{X}_1$.

The construction proceeds along these lines, using a sequence of functions $\{f_j\}_{j \in [m]}$, where for every $j \in [m]$ and $i \in \mathcal{X}_j$, we set $o_i^* = f_j(i)$. Now, for every $j \in [2, \ldots, m]$, we want the function f_j to be harder to compute than f_{j-1}, and we want a downwards self-reduction from computing f_j to computing f_{j-1}. The distinguisher A_j uses the given predictor $\tilde{p}$ as an oracle to f_{j-1}, and verifies that for $i \in \mathcal{X}_j$, $o_i = f_j(i)$. We emphasize that the complexity of the oracle distinguisher A_j is proportional to the cost of the downwards self-reduction, which can (and will) be significantly smaller than the complexity of computing f_j.

While intuitively appealing, the discussion above ignores an important point: OI only provides an approximate guarantee on the real-valued predictions, not exact recovery of the sequence of fucntions $\{f_j\}$. Starting at the first level of functions, an $(\{A_1\}, \varepsilon)$-oracle-access-OI predictor $\tilde{p}$ only has to correctly compute f_1 in a limited sense. First, $\tilde{p}$ only needs to be correct *on average* for

random inputs; it can err completely on some inputs. Second, while we will choose f_1 (and all the f_j's) to be a Boolean function, the predictor $\tilde{p}$ itself need not be Boolean. Nonetheless, for any input i, the distinguisher A_1 accepts the input $(i, \tilde{o}_i)$ only when $\tilde{o}_i = f_1(i)$, so taking ε small enough guarantees that for any $(\{A_1\}, \varepsilon)$-oracle-access-OI predictor $\tilde{p}$, with all but small probability over a draw of i from $\mathcal{D}_{\mathcal{X}}$ restricted to $\mathcal{X}_1$, rounding $\tilde{p}(i)$ gives the correct value of f_1. The probability of an error raises a new problem: the distinguisher A_2, which uses oracle calls to $\tilde{p}$ to compute f_1, might receive incorrect answers! Indeed, we expect the downwards self-reduction from f_2 to f_1 to make multiple queries (since f_2 is harder to compute), and so the error probability will be amplified. To handle this difficulty, we also want a *worst-case to average-case reduction* for f_1: from computing f_1 on worst-case inputs in $\mathcal{X}_1$, to computing f_1 w.h.p. over random inputs drawn from $\mathcal{D}_{\mathcal{X}}$ restricted to $\mathcal{X}_1$. Indeed, we'll want a similar reduction for each of the functions in the hierarchy. For each $j \in [2, \ldots, k]$, the distinguisher A_j will use the downwards self reduction, to f_{j-1}, using the worst-case to average-case reduction for f_{j-1} to reduce the error probability of $\tilde{p}$ before answering the downward self-reduction's oracle queries.

We can now make an inductive argument: assume that any $\tilde{p}$ that is OI for the distinguishers $\{A_1, \ldots, A_{j-1}\}$ must compute f_{j-1} correctly w.h.p. over inputs drawn from $\mathcal{D}_{\mathcal{X}}$ restricted to $\mathcal{X}_{j-1}$. Then $\tilde{p}$ is a very useful oracle for the j-th distinguisher A_j, which uses the downwards self-reduction and worst-case to average-case reductions, together with its oracle access to $\tilde{p}$, to compute f_j (and verify that $o_i = f_j(i)$). The key point is that A_j can do this (via oracle access to $\tilde{p}$), even though its running time is much smaller than the time needed to compute f_j. In turn, we conclude that any $\tilde{p}$ that is OI for the distinguishers $\{A_1, \ldots, A_j\}$ must compute f_j correctly w.h.p. over inputs drawn from $\mathcal{D}_{\mathcal{X}}$ restricted to $\mathcal{X}_j$. At the top (m-th) level of the induction, we conclude that a predictor that is OI for the entire collection of distinguishers must compute f_m correctly w.h.p. over random inputs. Finally, since we have a worst-case to average-case reduction for f_m, this implies that achieving OI is almost as hard as worst-case (randomized) computation of f_m.

To instantiate this framework, we need a collection of functions $\{f_j : \{0,1\}^n \to \{0,1\}\}_{j=1}^m$ with three properties: (1) "Scalable hardness": The complexity of computing f_j should increase with j. A natural goal is $n^{\Theta(j)}$ time complexity, where the lower bound should apply for randomized (BPP) algorithms; (2) Downwards self-reduction: A reduction from computing f_j to computing f_{j-1}, with fixed polynomial running time and query complexity (ideally $\tilde{O}(n)$, though we will use a collection where the complexity is a larger fixed polynomial); (3) Worst-case to average-case reduction: A reduction from computing f_j in the worst case, to computing f_j w.h.p. over a distribution D_j.

The clique counting problem presents a natural candidate, where f_{j-2} counts the number of j-cliques in an unweighted input graph with n vertices.[5] The complexity of this well-studied problem is believed to be $n^{\Theta(j)}$. Goldreich and Rothblum [21] recently showed a worst-case to average-case reduction for this problem, where the reduction runs in $\tilde{O}(n^2)$ times and makes $\text{poly}(\log n)$ queries. The problem also has a downwards self-reduction from counting cliques of size j to counting cliques of size $(j-1)$, which runs in time $O(n^3)$ and makes n oracle queries (on inputs of size $O(n^2)$). The above construction utilized a sequence of Boolean functions, whereas the output of the clique-counting function is an integer in $[n^j]$. We use the Goldreich-Levin hardcore predicate [20] to derive a Boolean function that is as hard to compute as clique counting.

The above framework can be also be instantiated using the algebraic variants of fine-grained complexity problems studied in the work of Ball, Rosen, Sabin, and Vasudevan [2, 3]. Interestingly, downwards self-reducability also comes up in their work [3], where it is used to argue hardness for batch evaluation of many instances. Their algebraic variants of the k-orthogonal-vectors and k-SUM problems seem directly suited to our construction. We focus on clique counting because of the tightness of the upper and lower bounds that have been suggested and studied in the literature.

For PSPACE hardness of oracle-access-OI, we use a PSPACE-complete problem that is both downwards self-reducible and random self-reducible, due to Trevisan and Vadhan [41]. The permanent [35] (or scaled-down variants thereof, see [21]) seems to be another promising candidate for our construction. In a different direction, it is interesting to ask whether moderately hard cryptographic assumptions, as suggested by Dwork and Naor [12], could also provide candidates.

1.4 Broader Context and Related Notions

A Socio-Technical Path of Progress. A sufficiently rich representation of real-life individuals implies a mapping from individuals to their representation as input to the predictor that has no collisions. In other words, given enough bits of information in the representation, each individual will have a unique representation. Still, this richness does *not* mean that the representation contains the right information to determine the values of the p_i^*, even information-theoretically. For example, modulo identical twins, individuals' genomes suffice to uniquely represent every person, but sequences of DNA are insufficient to determine an individual's ability to repay a loan. Alternatively, the necessary information may be present, but its interpretation may be computationally infeasible.

Generally, we assume that the representation of individuals is fixed and informative. Our analysis demonstrates that OI is feasible by using a potential argument. Specifically, we describe an algorithm that iteratively looks for updates to the current set of predictions that will step closer towards indistinguishability. We guarantee that the algorithm terminates in a bounded number of steps by arguing that after sufficiently many updates, the predictor we hold is essentially p^*.

In practice, however, it may be the case that our features will be insufficiently rich to capture p^*. Given a simple representation, even if we require a predictor $\tilde{p}$ that satisfies OI using a very computationally-powerful set of distinguishers (*e.g.*, polynomial-sized circuits), there will be an inherent, information-theoretic limitation that prevents $\tilde{p}$ from converging to p^*. While, given this representation of individuals, it may be impossible to distinguish Nature's outputs o^* from $\tilde{o}$ drawn according to $\tilde{p}$, it may be possible to distinguish the outputs if we obtain an enriched representation of individuals. Moreover, obtaining an enriched representation may

[5] Clique counting is trivial for cliques of size 1 or 2, and begins being interesting for 3-cliques, or triangle counting.

even be easy, in that it can be accomplished (by a human) in time polynomial in the size of the original representation!

The OI framework can be extended naturally, allowing for the representation of individuals to be augmented throughout time. Given such an enriched representation, we can continue iteratively updating $\tilde{p}$, based on the new representation. Specifically, we can obtain new training data, concatenate the old and enriched representations to form a new representation, initialize a new predictor to equal $\tilde{p}$, and enrich the collection of distinguishers to operate on the new representation. The new class of distinguishers retains and adds to the old distinguishing power, so $\tilde{p}$ likely will no longer satisfy OI; thus, we can apply our algorithm, starting at $\tilde{p}$, updating until the predictor fools the new class of distinguishers. By applying the same potential argument, we can can guarantee that this process of augmenting the representations cannot happen too many times. Any augmentation that significantly improves the distinguishing advantage between p^* and $\tilde{p}$ must result in new updates that allow for significant progress towards p^*.

Prediction Indistinguishability. In this work, we also investigate a notion we call *Prediction Indistinguishability (PI)*. In PI we require the stronger condition that (p_i^*, o_i^*) be indistinghishable from $(\tilde{p}_i, o_i^*)$. While Prediction Indistinguishability is intuitively appealing, there are very simple distinguishers that show it is too much to ask for: a predictor $\tilde{p}$ is PI with respect to these distinguishers if and only if $\tilde{p}$ is statistically close to p^*. But indistinguishability as a concept in computational complexity theory is interesting precisely when coming up with $\tilde{p}$ that is statistically close to p^* is impossible. Moreover, since we never see the values p_i^*—we don't even know if randomness exists!—we cannot hope to have indistinguishability of the $\tilde{p}_i$ from the "true" probabilities and it is strange even to pose such a criterion. Nonetheless, we show that PI and OI are equivalent when indistinguishability is with respect to tests that are passed by all natural histories with high probability (Section 3), which we discuss in more detail next.

Tests vs. Distinguishers. 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., [17–19, 42] and references therein). A comparison to the extensive literature on online forecasting clarifies the semantic distinction between two concepts: *tests* (in the forecasting literature) and *distinguishers* (in the complexity literature).

The forecasting 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?). There are many possibilities for Nature; by definition, in this literature, Nature calls the shots, in the following sense: On Day $t-1$, Nature assigns a probability p_t^* that governs whether it will rain or not on Day t. Note that Nature is free to select $p_t^* \in \{0,1\}$, in which case the outcomes are determinsitic, $o_t^* = p_t^*$

In the early 1980s, [7] proposed that, at the very least, forecasts should be calibrated. More stringent requirements were obtained by considering large (countable) numbers of sets of days, such as odd days, even days, prime-numbered days, days on which it has rained for exactly 40 preceding days and nights, and so on, and requiring calibration for each of these sets simultaneously [38]. This is the "moral equivalent" of multi-calibration in the world of infinite sequences of online forecasting.

A signal result in the forecasting literature, due to Sandroni [37] applies to a more general set of tests than calibration tests. Consider infinite histories, that is, sequences of (prediction, outcome) pairs. We say a history is *natural* if it is a sequence $((p_1, o_1), (p_2, o_2), \dots)$ where $\forall t$ we have $o_t \sim \text{Ber}(p_t)$. Note that certain natural histories may have no connection to any real-life weather phenomena, instead corresponding to a valid but unrealistic choice of p. A *test* takes as input a (not necessarily natural) history and outputs a bit. The test is usually thought of as trying to assess whether an algorithm's predictions are "reasonably accurate" with respect to the actual observations. This implicitly focuses attention on tests that natural histories pass with high probability (over the draws from the Bernoulli distributions), and indeed, calibration tests fall into this category. The goal of the Algorithm, then, is to generate predictions $\tilde{p}$ for which the histories $((\tilde{p}_1, o_1^*), (\tilde{p}_2, o_2^*), \dots)$ pass the test. Here $\tilde{p}_i$ is the Algorithm's forecasted probability of rain for Day i and o_i^* is the Boolean outcome, rain or not, that actually occured on Day i.

Sandroni's powerful result [37], proves, non-constructively[6], the existence of an Algorithm that, given any test T, yields a history which passes T with probability at least as great as the minimum probability with which any natural history $((p_1, o \sim \text{Ber}(p_1)), (p_2, o \sim \text{Ber}(p_2)) \dots)$ passes T (again, the probability is over the draws from the $\text{Ber}(p_i)$, $i \geq 1$). The computational complexity of forecasting was studied by Fortnow and Vohra [13] and by Chung, Lui and Pass [6]. The latter work gave a computational analogue of Sandroni's result: for any test T that is computable in polynomial time and that accepts every natural history with high probability, they construct a *polynomial-time* forecasting algorithm that passes the test with high probability, so long as Nature (which generates the outcomes) runs in fixed polynomial time and assuming also that Nature does not use any hidden state.

There are two major differences between tests in the forecasting literature and distinguishers in the complexity-theoretic literature.[7] The first is that tests have semantics—you want to pass the test, and the higher the probability of passing a test the better. In contrast, distinguishers output 0 or 1 with no semantics, and our goal is to produce an object such that the distinguisher outputs 1 with the same probability as the objects that we are imitating. In this case, getting the distinguisher to output 1 with higher probability may be worse. The second difference is that in the forecasting setting we want natural histories to pass the test with high probability: if natural histories fail the test, how can we interpret the Algorithm's inability to pass the test as an indication that the Algorithm is inaccurate? As a result, the Algorithm does not compete with the actual Nature p^*, but only with the hypothetical Nature that passes the test with the lowest probability.

To highlight these differences, consider a distinguisher that considers two sets of individuals, S and T. For each set, the distinguisher estimates the outcome probabilities α_S and α_T—that is, averaged

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

[7] Unfortunately, and confusingly, the literature on indistinguishability often conflates the notions, referring to distinguishers as tests.

over the individuals $i \in S$ (respectively, T), the probability that $o_i^* = 1$—and outputs 0 with probability $|\alpha_S - \alpha_T|$ and 1 otherwise. Note that for some Natures the distinguisher will output 1 with very small probability. Nevertheless, in cases where Nature treats S and T equally, the distinguisher will output 1 with high probability and, OI guarantees that $\tilde{p}$ must also treat S and T equally. Many properties of samples from a distribution are quite naturally and directly specified through the language of distinguishers, and not obviously through the language of tests. In light of this discussion, the connection between sample-access-OI and multi-calibration is very interesting: it shows how to reduce a collection of distinguishers into a collection of tests, and even more specifically to a collection of calibration tests.

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 [10, 25, 31]—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.

Central to evidence-based fairness is the notion of multi-calibration [25], which was also studied in the context of rankings in [10]. Recently, [26] 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. [39] study multi-calibration from a sample-complexity perspective. In a similar vein, [44] 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 [4, 10, 25–28, 31, 32, 39]. This approach to fairness aims to strengthen the guarantees of notoriously-weak group fairness notions, while maintaining their practical appeal. For instance, [27, 28, 32] give notions of multi-group fairness based on parity notions studied in [9] and [23]. [4] 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 [23, 27, 28, 34].

Computational and Statistical Learning. Prediction tasks have also been studied extensively in the theoretical computer science and machine learning communities, both in the offline PAC model [43], as well as in the online model [15, 36]; see [40] and references therein. Relatedly, [5] also show a multi-calibration-style guarantee in the online "sleeping experts" setting. More broadly, our work is also in conversation with more applied approaches to learning distributions and generative models including GANs [22] or auto-encoders [33]. The perspective of generating (statistically) indistinguishabile samples was also recently considered in a work introducing the problem of "sample amplification" [1].

Organization. The remainder of the manuscript is structured as follows. Section 2 defines OI formally and shows a number of propositions relating the various notions of OI to one another. Section 3 introduces the notion of prediction indistinguishability, and investigates the relationship of OI distinguishers to forecasting-style tests. The formal statements of theorems is deferred to the full version of the manuscript [11]. There, we include proofs and further exposition on the connections between the first two levels of the OI hierarchy to multi-accuracy and multi-calibration, our construction of OI predictors, establishing the feasibility of the final two levels, and our construction establishing conditional lower bounds against the final levels.

2 OUTCOME INDISTINGUISHABILITY

Throughout this work, we study how to obtain predictors that generate outcomes that cannot be distinguished from natural outcomes. Specifically, we model Nature as a joint distribution over individuals and outcomes, denoted $\mathcal{D}^*$. Individuals come from a discrete domain $\mathcal{X}$; throughout, we will assume that each $i \in \mathcal{X}$ can be resolved to some d-dimensional boolean string $i \in \{0,1\}^d$, which represents the "features" of the individual. In this work, we focus on boolean outcomes $\mathcal{Y} = \{0,1\}$. Thus, $\mathcal{D}^*$ is supported on $\mathcal{X} \times \mathcal{Y} \subseteq \{0,1\}^d \times \{0,1\}$. We use $i \sim \mathcal{D}_\mathcal{X}$ to denote a sample from Nature's marginal distribution over individuals.

We say that a *predictor* is a function $p : \mathcal{X} \to [0,1]$ that maps individuals to an estimate of the conditional probability of the individual's outcome being 1. For ease of notation, we use $p_i = p(i)$ to denote a predictor's estimate for individual i. Note that the marginal distribution over individuals $\mathcal{D}_\mathcal{X}$ paired with a predictor induce a joint distribution over $\mathcal{X} \times \mathcal{Y}$. Given a predictor p, we use $(i, o_i) \sim \mathcal{D}(p)$ to denote an individual-outcome pair, where $i \sim \mathcal{D}_\mathcal{X}$ is sampled from Nature's distribution over individuals, and the outcome $o_i \sim \text{Ber}(p_i)$ is sampled—conditional on i—according to the Bernoulli distribution with parameter p_i.

With this basic setup in place, we are ready to introduce the main notion of this work—outcome indistinguishability (OI). Intuitively, when developing a prediction model, a natural goal would be to learn a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ whose outcomes "look like" Nature's distribution $\mathcal{D}^*$. Outcome indistinguishability formalizes this intuition, and is parameterized by a family of distinguisher algorithms $\mathcal{A}$. In the most basic form of OI, each $A \in \mathcal{A}$ receives as input a labeled sample from one of two distributions, Nature's distribution $\mathcal{D}^*$ or the predictor's distribution $\mathcal{D}(\tilde{p})$.

$$\underbrace{(i, o_i^*) \sim \mathcal{D}^*}_{\text{Nature's distribution}} \qquad \underbrace{(i, \tilde{o}_i) \sim \mathcal{D}(\tilde{p})}_{\text{Predictor's distribution}}$$

In other words, in each distribution the individual i is sampled according to nature's marginal distribution on inputs, denoted $i \sim \mathcal{D}_\mathcal{X}$. The distribution over outcomes, however, varies: conditioned on the individual i, the distinguisher either receives the corresponding natural outcome o_i^*, or receives an outcome sampled as $\tilde{o}_i \sim \text{Ber}(\tilde{p}_i)$. In its most basic form, a predictor $\tilde{p}$ satisfies OI over the family $\mathcal{A}$ if for all $A \in \mathcal{A}$, the probability that A accepts the sample (i, o_i) is (nearly) the same for Nature's distribution and the predictor's distribution. In addition to the sample from $\mathcal{D}^*$ versus $\mathcal{D}(\tilde{p})$, we can also allow the distinguishers to access the predictor $\tilde{p}$ itself. This setup allows us to define a prototype for a notion of OI.

Definition 2.1 (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 $A \in \mathcal{A}$,*

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

The subsequent sections introduce multiple variants of outcome indistinguishability, highlighting four distinct access patterns to $\tilde{p}$. By allowing the distinguishers increasingly liberal access to the predicitve model $\tilde{p}$, the indistinguishability guarantee becomes increasingly strong.

Remark on nature. We primarily model Nature $\mathcal{D}^*$ as a fixed and unknown joint distribution over $\mathcal{X} \times \mathcal{Y}$. The presentation of some result benefits from an equivalent view, based on the agnostic PAC framework [24, 29, 30]. In this view, we imagine that individuals $i \sim \mathcal{D}_\mathcal{X}$ are sampled from the marginal distribution over $\mathcal{X}$, and then Nature selects outcomes conditioned on i. Throughout, we will use $p^* : \mathcal{X} \to [0,1]$ to denote the function that maps individuals to the true conditional probability of outcomes given the individual. That is, for all $i \in \mathcal{X}$:

$$p_i^* = \Pr_{(i,o_i^*)\sim\mathcal{D}^*} \left[o_i^* = 1 \,\middle|\, i \right].$$

In our notation, we can imagine that Nature specifies the distribution over individuals $i \sim \mathcal{D}_\mathcal{X}$, then specifies the "natural predictor" p^* and samples the outcome for an individual i as $o_i^* \sim \mathrm{Ber}(p_i^*)$; in other words, $\mathcal{D}^* = \mathcal{D}(p^*)$. We emphasize that this view—of Nature selecting a predictor—is an abstraction that is sometimes instructive in our analysis of OI. Nevertheless, we make no assumptions about p^* other than it defines a valid conditional probability distribution for each $i \in \mathcal{X}$. In particular, p^* need not come from any realizable or learnable class of functions.

Expectations and norms. We take expectations over Nature's marginal distribution over individuals, possibly conditioned on membership in particular subpopulations $S \subseteq \mathcal{X}$. A simple but important observation is that for any subpopulation $S \subseteq \mathcal{X}$, the expected outcome is equal to the expectation of p^*.

$$\Pr_{(i,o_i^*)\sim\mathcal{D}^*} \left[o_i^* = 1 \,\middle|\, i \in S \right] = \mathop{\mathbf{E}}_{i\sim\mathcal{D}_\mathcal{X}} \left[p_i^* \,\middle|\, i \in S \right]$$

Similarly, we may compare predictors over the distribution of individuals. For any two predictors $p, p' : \mathcal{X} \to [0,1]$, we use the following ℓ_1-distance to measure the statistical distance between their outcome distributions $\mathcal{D}(p)$ and $\mathcal{D}(p')$.

$$\left\| p - p' \right\|_1 = \mathop{\mathbf{E}}_{i\sim\mathcal{D}_\mathcal{X}} \left[\left| p_i - p_i' \right| \right]$$

We only use the $\|\cdot\|_1$ notation when the distribution on individuals $\mathcal{D}_\mathcal{X}$ is clear from context.

Supported predictions. In many definitions, we can reason about predictors as arbitrary functions $p : \mathcal{X} \to [0,1]$, but to be an effective definition, we need to discuss functions that are implemented by a realizable model of computation. Importantly, this means we will think of predictors as mapping individuals $i \in \mathcal{X}$ to a range of values p_i that live on a discrete subset of $[0,1]$. We assume for any predictor $p : \mathcal{X} \to [0,1]$, the predictor's support is a discrete set of values in $[0,1]$ that receive positive probability over $\mathcal{D}_\mathcal{X}$. For any subpopulation $S \subseteq \mathcal{X}$, we denote the support of p on S as

$$\mathrm{supp}_S(p) = \left\{ v \in [0,1] : \Pr_{i\sim\mathcal{D}_\mathcal{X}} \left[p_i = v \mid i \in S \right] > 0 \right\}$$

In this way, for any $v \in \mathrm{supp}(p)$, the conditional distribution over individuals $i \sim \mathcal{D}_\mathcal{X}$ where we condition on the event $p_i = v$ is well-defined.

When possible, we obtain results agnostic to the exact choice of discretization. Sometimes, we need to reason about the discretization explicitly and map the values of $\mathrm{supp}(p)$ onto a known grid with fixed precision; we introduce additional technical details as needed in the subsequent sections.

Distinguishers and subpopulations. The notion of outcome indistinguishability is parameterized by a family of distinguishing algorithms, which we denote as $\mathcal{A}$. To instantiate a concrete notion of OI (at any of the four levels we define), we must specify $\mathcal{A}$ within a fixed realizable model of computation. In practice, it may make sense to use a class of learning-theoretic distinguishers, (e.g., decision-trees, halfspaces, neural networks). In this work, we focus on more abstract models of distinguishers. When our proofs allow, we will treat $\mathcal{A}$ as an arbitrary class of computations, but for certain results, it will be easier to assume something about the model of computation in which $A \in \mathcal{A}$ are implemented (e.g., time-bounded uniform, size-bounded non-uniform).

Recall, we assume the domain of individuals $\mathcal{X} \subseteq \{0,1\}^d$ can be represented as d-dimensional boolean vectors for $d \in \mathbb{N}$, and that the distinguishing algorithms $A \in \mathcal{A}$ take as input an individual $i \in \mathcal{X}$ and an outcome $o_i \in \{0,1\}$. Often, we will think of the dimension d as fixed. In this case, we can think of $\mathcal{A}$ as a fixed class of distinguishers of concrete complexity: for example, if the class $\mathcal{A}$ is implemented by circuits, then we can reason about their size as $s(d) = s$ for some fixed $s \in \mathbb{N}$. When we think of the dimension as growing, then we need to consider ensembles of distinguishing families, where the family is parameterized by $d \in \mathbb{N}$.

The same issues arise when we discuss multi-calibration, which is parameterized by a collection of "efficiently-identifiable" subpopulations $C \subseteq \{0,1\}^\mathcal{X}$. Here, efficiently-identifiable refers to the fact that we assume for each $S \in C$, there exists some efficient computational model that given $i \in \mathcal{X}$, can compute the predicate $\mathbf{1}[\, i \in S \,]$ (i.e., the characteristic function of S). Again, whenever possible, our treatment does not depend on the exact model of computation.

Circuits. As suggested above, some results are most naturally stated with a concrete model of computation in mind. In these cases, we will describe the distinguishers and subpopulations as computed by a collection of circuits. Fixing such a model of circuits will be useful when relating the complexity of a class of distinguishers $\mathcal{A}$ to that of a class of subpopulations C, as well as showing the feasibility of obtaining circuits that implement OI predictors. Analogous results could be proved instead for uniform classes.

Throughout, we say that a family of distinguishers $\mathcal{A}$ (resp., subpopulations C) for $\mathcal{X} \subseteq \{0,1\}^d$ is implemented by a family of circuits of size $s(d)$, if for each $A \in \mathcal{A}$ (resp., $S \in C$), there exists a bounded fan-in circuit over the complete boolean basis c_A that computes the distinguisher A on all inputs, with at most $s(d)$ gates (or equivalently, by bounded fan-in, $\Theta(s(d))$ wires).

Specifying the model of computation for the most stringent levels of OI requires some care. The third level—oracle-access-OI—allows the distinguishers oracle-access to the predictor in question. For each $A \in \mathcal{A}$, we denote the oracle distinguisher as $A^{\tilde{p}}$, which has random access to $\tilde{p}_i$ for any $i \in \mathcal{X}$. The fourth level—code-access-OI—allows the distinguishers direct access to the description of the predictor in question, denoted $\langle \tilde{p} \rangle$. In this case, it makes sense to allow the ensemble of distinguishers to be parameterized by the length of the description $n = |\langle \tilde{p} \rangle|$ in addition to the dimension d. We discuss the specific assumptions about the implementation of these notions in subsequent sections.

2.1 Defining the Levels of Outcome Indistinguishability

With the general framework and preliminaries in place, we are ready to define the various levels of outcome indistinguishability. In this section, we focus on the definitions of each notion—no-access-OI, sample-access-OI, oracle-access-OI, and code-access-OI. Along the way, we show some relations between the notions, but defer most of our investigation of the notions to subsequent sections. We begin introducing each notion in the single-sample setting, and discuss OI for distinguishers that receive multiple samples in Section 2.2.

2.1.1 No-Access-OI. The weakest model of distinguisher receives no direct access to the predictive model $\tilde{p}$, and must make its judgments based only on the observed sample. In this framework, the only access to the predictor is indirect, through the sampled outcomes.

Definition 2.2 (No-Access-OI). *Fix Nature's distribution $\mathcal{D}^*$. Let $\mathcal{Z} = \mathcal{X} \times \{0,1\}$. For a class of distinguishers $\mathcal{A} \subseteq \{\mathcal{Z} \to \{0,1\}\}$ and $\varepsilon > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{A}, \varepsilon)$-no-access-OI if for every $A \in \mathcal{A}$,*

$$\left| \mathop{\mathbf{Pr}}_{(i,o_i^*)\sim\mathcal{D}^*} \left[A(i, o_i^*) = 1 \right] - \mathop{\mathbf{Pr}}_{(i,\tilde{o}_i)\sim\mathcal{D}(\tilde{p})} \left[A(i, \tilde{o}_i) = 1 \right] \right| \leq \varepsilon.$$

We remark that from a statistical perspective, no-access-OI already defines a strong framework for indistinguishability. Even at this baseline level, if we allow a computationally-inefficient class of distinguishers, no-access-OI can be used to require closeness in statistical distance. For instance, consider a family of "subset" distinguishers, where for a subset $S \subseteq \mathcal{X}$, the distinguisher A_S is defined as follows.

$$A_S(i, o_i) = \begin{cases} o_i & \text{if } i \in S \\ 0 & \text{o.w.} \end{cases}$$

If we take $\mathcal{A} = \{A_S : S \subseteq \mathcal{X}\}$ to be the family of all subset distinguishers, then the only predictors $\tilde{p}$ that satisfy no-access-OI will be statistically close to Nature's predictor p^*. Of course, this class of distinguishers includes inefficient tests (necessary to certify $\|p^* - \tilde{p}\|_1$ is small). Our interest will be on the guarantees afforded by OI when we take $\mathcal{A}$ to be a class of efficient distinguishers.

2.1.2 Sample-Access-OI. To strengthen the distinguishing power, we define sample-access-OI, which allows distinguishers to observe the prediction for the individual in question. Specifically, in addition to the sampled individual $i \sim \mathcal{D}$ and outcome o_i (drawn according to nature or the predictor $\tilde{p}$), the distinguisher receives the prediction $\tilde{p}_i$.

Definition 2.3 (Sample-Access-OI). *Fix Nature's distribution $\mathcal{D}^*$. Let $\mathcal{Z} = \mathcal{X} \times \{0,1\} \times [0,1]$. For a class of distinguishers $\mathcal{A} \subseteq \{\mathcal{Z} \to \{0,1\}\}$ and $\varepsilon > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{A}, \varepsilon)$-sample-access-OI if for every $A \in \mathcal{A}$,*

$$\left| \mathop{\mathbf{Pr}}_{(i,o_i^*)\sim\mathcal{D}^*} \left[A(i, o_i^*, \tilde{p}_i) = 1 \right] - \mathop{\mathbf{Pr}}_{(i,\tilde{o}_i)\sim\mathcal{D}(\tilde{p})} \left[A(i, \tilde{o}_i, \tilde{p}_i) = 1 \right] \right| \leq \varepsilon.$$

Sample-access-OI is a strengthening of no-access-OI: for any no-access-OI distinguisher, on input $(i, o_i, \tilde{p}_i)$, a sample-access-OI distinguisher can simply ignore the prediction $\tilde{p}_i$, and simulate the original no-access-OI distinguisher.

2.1.3 Oracle-Access-OI. The next strengthening of OI allows distinguishers to make queries to $\tilde{p}$, not just on the sampled individual $i \sim \mathcal{D}$, but also on any other $j \in \mathcal{X}$. Such a query model needs to be formalized; at a high level, we assume the distinguishers in the class $A \in \mathcal{A}$ are augmented with oracle access to $\tilde{p}$, denoted as $A^{\tilde{p}}$.

Definition 2.4 (Oracle-Access-OI). *Fix Nature's distribution $\mathcal{D}^*$. Let $\mathcal{Z} = \mathcal{X} \times \{0,1\}$. For a class of oracle distinguishers $\mathcal{A} \subseteq \{\mathcal{Z} \to \{0,1\}\}$ and $\varepsilon > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{A}, \varepsilon)$-oracle-access-OI if for every $A^{\tilde{p}} \in \mathcal{A}$,*

$$\left| \mathop{\mathbf{Pr}}_{(i,o_i^*)\sim\mathcal{D}^*} \left[A^{\tilde{p}}(i, o_i^*) = 1 \right] - \mathop{\mathbf{Pr}}_{(i,\tilde{o}_i)\sim\mathcal{D}(\tilde{p})} \left[A^{\tilde{p}}(i, \tilde{o}_i) = 1 \right] \right| \leq \varepsilon.$$

The exact formulation of such oracle distinguishers will vary based on the model of computation in which $\mathcal{A}$ is defined. Independent of the exact model, oracle-access-OI can implement sample-access-OI: on input (i, o_i), the oracle-access-OI distinguisher can access $\tilde{p}_i$ using a single query and then simulate the sample-access-OI distinguisher.

Lunchtime-OI and Sample-Access-OI.. Oracle-access-OI generally defines a stronger notion of indistinguishability than sample-access-OI, but we show that if the oracle-access-OI distinguishers are non-adaptive—asking only pre-processing queries—then they can be simulated by a family of (non-uniform) sample-access-OI distinguishers. This result demonstrates that the power of oracle-access distinguishers over sample-access distinguishers derives from the ability to query $\tilde{p}$ *adaptively*, based on the sample in question. In particular, we show that oracle-access-OI is strictly more powerful than sample-access-OI. The construction follows by exploiting correlations within $\mathcal{D}^*$ across different $i, j \in \mathcal{X}$, which can be tested efficiently by querying $\tilde{p}$ adaptively.

In fact, this collapse from oracle-access-OI to sample-access-OI will hold for an even more powerful class of distinguishers, which are allowed "lunchtime attack" style pre-processing on $\tilde{p}$. Consider the following model of pre-processing analysis. For some $t \in \mathbb{N}$, given a family of distinguishers $\mathcal{A}$, suppose that for each $A \in \mathcal{A}$, there exists a pre-processing algorithm $R_A^{\tilde{p}} : 1^d \to \{0,1\}^t$ with oracle access to $\tilde{p}$. Given access to $\tilde{p}$ for input domain $\mathcal{X} \subseteq \{0,1\}^d$, the pre-processing algorithm $R_A^{\tilde{p}}(1^d)$ produces an advice string $a \in \{0,1\}^t$. Then, oracle access to $\tilde{p}$ is revoked, and the distinguisher A receives a individual-outcome-prediction sample $(i, o_i, \tilde{p}_i)$ from one

of the two distributions, given access to a.[8] That is, the lunchtime variant of $(\mathcal{A}, \varepsilon)$-oracle-access-OI holds if for every $A \in \mathcal{A}$ and $a = R_A^{\tilde{p}}(1^d)$,

$$\left| \Pr_{(i,o_i^*)\sim\mathcal{D}^*} \left[A^a(i, o_i^*, \tilde{p}_i) = 1 \right] - \Pr_{(i,\tilde{o}_i)\sim\mathcal{D}(\tilde{p})} \left[A^a(i, \tilde{o}_i, \tilde{p}_i) = 1 \right] \right| \leq \varepsilon.$$

Note that computing $R_A^{\tilde{p}}(1^d)$ need not be efficient, but importantly, its analysis of $\tilde{p}$ must be summarized into t bits. For this variant of oracle-access-OI, we show the following inclusion.

Proposition 2.5. *Fix Nature's distribution $\mathcal{D}^*$. Let $\mathcal{Z} = \mathcal{X} \times \{0,1\} \times [0,1]$. Suppose $\mathcal{A} \subseteq \{\mathcal{Z} \to \{0,1\}\}$ is a class of lunchtime distinguishers implemented by size-s circuits. Then, there exists a class of sample-access distinguishers $\mathcal{A}'$ implemented by size-s circuits, such that any predictor $\tilde{p} : \mathcal{X} \to [0,1]$ that satisfies $(\mathcal{A}', \varepsilon)$-sample-access-OI must also satisfy $(\mathcal{A}, \varepsilon)$-oracle-access-OI.*

Proof. Given a class of lunchtime distinguishers $\mathcal{A}$, we define a new class $\mathcal{A}'$ of sample-access distinguishers as follows.

$$\mathcal{A}' = \left\{ A'_a : A \in \mathcal{A},\ a \in \{0,1\}^t \right\}$$

where A'_a is defined as

$$A'_a(i, o_i, p_i) = A^a(i, o_i, p_i)$$

for all $i \in \mathcal{X}$ and $o_i \in \{0,1\}$ and $p_i \in [0,1]$. In other words, for each $A \in \mathcal{A}$, we introduce 2^t fixed distinguishers that have the possible output of $R_A^{\tilde{p}}(1^d)$ hard-coded. If A is implemented by a circuit with access to the advice string output by $R_A^{\tilde{p}}(1^d)$, then for any $a \in \{0,1\}^t$, A'_a can be implemented by circuits with the same number of wires. We argue that $(\mathcal{A}', \varepsilon)$-sample-access-OI implies $(\mathcal{A}, \varepsilon)$-oracle-access-OI.

Suppose there is some $A \in \mathcal{A}$ such that A^a distinguishes between the natural and modeled distribution. Then, by construction, there exists some $A'_a \in \mathcal{A}'$ that also distinguishes the distributions with the same advantage. Thus, by contrapositive, if a predictor $\tilde{p}$ satisfies $(\mathcal{A}', \varepsilon)$-sample-access-OI, then it also satisfies $(\mathcal{A}, \varepsilon)$-oracle-access-OI. □

Note that we state and prove Proposition 2.5 for distinguishers implemented by circuits, but the construction is quite generic. This style of hard-coding works very naturally for any non-uniform class model of distinguishers. Even if we work with a uniform model of distinguishers, if the length of the advice string $t \in \mathbb{N}$ is a constant (independent of d the dimension of individuals $\mathcal{X}$), then for each $A \in \mathcal{A}$ we can define a TM that has $a \in \{0,1\}^t$ hard-coded as part of its description. The number of distinguishers in $\mathcal{A}'$ grows by a factor of 2^t.

2.1.4 Code-Access-OI. The strongest notion of distinguishers we consider receive—as part of their input—the description $\langle \tilde{p} \rangle$ of a circuit that computes $\tilde{p}$. In this model, which we call code-access-OI, the distinguishers can accept or reject their sample based on nontrivial analysis of the circuit computing $\tilde{p}$, not just its evaluation on domain elements. We assume that $|\langle \tilde{p} \rangle| = n$ for some $n \in \mathbb{N}$.

[8] Note that, as in sample-access-OI, we additionally give $\tilde{p}_i$ as input to the lunchtime distinguisher. We exclude the prediction $\tilde{p}_i$ as input in Definition 2.4 because, in general, an adaptive oracle-access-OI distinguisher can query $\tilde{p}_i$ as desired. Without feeding the prediction as input, lunchtime-OI actually collapses to no-access-OI.

Definition 2.6 (Code-Access-OI). *Fix Nature's distribution $\mathcal{D}^*$. Let $Z = \mathcal{X} \times \{0,1\} \times \{0,1\}^n$ for $n \in \mathbb{N}$. For a class of distinguishers $\mathcal{A} \subseteq \{\mathcal{Z} \to \{0,1\}\}$ and $\varepsilon > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{A}, \varepsilon)$-code-access-OI if for every $A \in \mathcal{A}$,*

$$\left| \Pr_{(i,o_i^*)\sim\mathcal{D}^*} \left[A(i, o_i^*, \langle \tilde{p} \rangle) = 1 \right] - \Pr_{(i,\tilde{o}_i)\sim\mathcal{D}(\tilde{p})} \left[A(i, \tilde{o}_i, \langle \tilde{p} \rangle) = 1 \right] \right| \leq \varepsilon.$$

There are a number of subtle technicalities in how we define code-access-OI, relating to how we encode $\langle \tilde{p} \rangle$. In particular, if we want to be able to simulate the prior notions of OI within code-access-OI, then we need to allow the complexity of the distinguishers in $\mathcal{A}$ to scale with the complexity of $\tilde{p}$. Even evaluating $\tilde{p}$ on a single domain element requires that $\mathcal{A}$ can compute circuit evaluation. This technicality sets code-access-OI apart from the prior notions, where it sufficed to think of the domain as fixed in dimension, and thus think of the distinguishers' complexity as fixed as well.

2.2 Multiple Sample OI

Throughout this work, we focus on distinguishers that receive a single sample from nature or the modeled distribution, with varying levels of access to $\tilde{p}$. A natural generalization of this model allows distinguishers to access multiple samples. We define the generic variant as follows (where each of no-access-OI, sample-access-OI, oracle-access-OI, and code-access-OI follow by allowing distinguishers the analogous degree of access to $\tilde{p}$).

Definition 2.7. *Fix Nature's distribution $\mathcal{D}^*$. Let $m \in \mathbb{N}$ and $\mathcal{Z} = (\mathcal{X} \times \{0,1\})$. For a class of multi-sample distinguishers $\mathcal{A}_m \subseteq \{\mathcal{Z}^m \to \{0,1\}\}$ and $\varepsilon > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{A}_m, \varepsilon)$-OI if for every $A_m \in \mathcal{A}_m$,*

$$\left| \Pr_{(i_1,o_{i_1}^*),\ldots,(i_m,o_{i_m}^*)\sim(\mathcal{D}^*)^m} \left[A_m\left((i_1, o_{i_1}^*), \ldots, (i_m, o_{i_m}^*)\right) = 1 \right] \right.$$

$$\left. - \Pr_{(i_1,\tilde{o}_{i_1}),\ldots,(i_m,\tilde{o}_{i_m})\sim\mathcal{D}(\tilde{p})^m} \left[A_m\left((i_1, \tilde{o}_{i_1}), \ldots, (i_m, \tilde{o}_{i_m})\right) = 1 \right] \right| \leq \varepsilon.$$

We leave full exploration of multi-sample-OI to future work, but make the following observation. If the class of distinguishers we use admits a hybrid argument, then the multi-sample distinguishers' advantage can be bounded generically in terms of the single-sample advantage. As an example, we show the following proposition for oracle-access-OI.

Proposition 2.8. *Fix Nature's distribution $\mathcal{D}^*$. Let $\mathcal{A}$ be the class of size-s single-sample distinguishers, and for $m \in \mathbb{N}$ let $\mathcal{A}_m$ be the class of size-s m-sample distinguishers. Suppose we allow $\mathcal{A}$ pre-processing samples from $\mathcal{D}^*$ and oracle-access to $\tilde{p}$. For $\varepsilon > 0$, if a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is $(\mathcal{A}, \varepsilon/m)$-oracle-access-OI, then it is $(\mathcal{A}_m, \varepsilon)$-oracle-access-OI.*

Proof. Suppose there exists some m-sample distinguisher $A_m \in \mathcal{A}_m$ that distinguishes between nature and the model $\tilde{p}$ with advantage at least ε. We show that there is a single-sample randomized distinguisher $A \in \mathcal{A}$ that distinguishes between nature and the model with advantage at least ε/m. By contrapositive, if $\tilde{p}$ is $(\mathcal{A}, \varepsilon/m)$-oracle-access-OI, then it must be $(\mathcal{A}_m, \varepsilon)$-oracle-access-OI.

Consider the following sequence of hybrid distributions over m samples, $(i_1, o_{i_1}), \ldots, (i_m, o_{i_m})$, where $\mathcal{D}_k = (\mathcal{D}^*)^{m-k} \times \mathcal{D}(\tilde{p})^k$ is a product distribution of $m-k$ independent samples from nature and k samples from the model. Note that assuming pre-processing access to samples from $\mathcal{D}^*$ and oracle access to $\tilde{p}$, each $\mathcal{D}_k$ is sampleable. Specifically, to obtain a sample from $\mathcal{D}_k$, we will draw m samples from $\mathcal{D}^*$, and then for each $j \in \{m-k+1, \ldots, m\}$, we resample the outcome by evaluating $\tilde{p}_{i_j}$ and then randomly drawing $\tilde{o}_{i_j} \sim \mathrm{Ber}(\tilde{p}_{i_j})$.

Observing that $\mathcal{D}_0 = (\mathcal{D}^*)^m$ and $\mathcal{D}_m = \mathcal{D}(\tilde{p})^m$, we can write the distinguishing probability of A_m as a telescoping sum over distinguishing probabilities over the hybrid distributions.

$$\begin{aligned}
&\Pr_{(i_1,o^*_{i_1}),\ldots,(i_m,o^*_{i_m})\sim(\mathcal{D}^*)^m}\left[A_m^{\tilde{p}}\left((i_1,o^*_{i_1}),\ldots,(i_m,o^*_{i_m})\right)=1\right] \\
&- \Pr_{(i_1,\tilde{o}_{i_1}),\ldots,(i_m,\tilde{o}_{i_m})\sim\mathcal{D}(\tilde{p})^m}\left[A_m^{\tilde{p}}\left((i_1,\tilde{o}_{i_1}),\ldots,(i_m,\tilde{o}_{i_m})\right)=1\right] \\
&= \sum_{j=1}^{m}\left(\Pr_{(I,O)\sim\mathcal{D}_{j-1}}\left[A_m^{\tilde{p}}(I,O)=1\right] - \Pr_{(I,O)\sim\mathcal{D}_j}\left[A_m^{\tilde{p}}(I,O)=1\right]\right) \\
&\geq \varepsilon
\end{aligned}$$

Thus, the following randomized single-sample oracle-access-OI distinguisher succeeds with advantage at least ε/m: as pre-processing, sample a random index $j \sim [m]$ and draw a sample from the hybrid distribution $\mathcal{D}_j$; on input (i, o_i), replace the jth sample with the input (i, o_i), and run A_m on the resulting m-sample input. If the input is drawn from nature, then the resulting sample is drawn from $\mathcal{D}_{j-1}$, whereas if the input is from the model, then the resulting sample is drawn from $\mathcal{D}_j$. Thus, the distinguishing advantage of A is the average distinguishing advantage between $\mathcal{D}_{j-1}$ and $\mathcal{D}_j$, or ε/m. □

We state Proposition 2.8 for oracle-access-OI (and thus, by simulation, code-access-OI), due to the ease of running the hybrid argument with oracle access to $\tilde{p}$. Note that we use circuit-size as the complexity measure for concreteness, but the argument will go through for most complexity measures of $\mathcal{A}_m$. Similar hybrid arguments can also be made for no-access-OI and sample-access-OI, provided the model of computation of the distinguishers admits "hard-coding" the outcome values $\{\tilde{o}_{i_{m-k+1}}, \ldots, \tilde{o}_{i_m}\}$, and $\{\tilde{p}_{i_1}, \ldots, \tilde{p}_{i_{m-k}}\}$ if needed (for sample-access-OI). In particular, for any non-uniform class of multi-sample distinguishers $\mathcal{A}$, there exists a class $\mathcal{A}'$ of single-sample distinguishers that simulates the distinguishers in $\mathcal{A}$ with the choices hard-coded.

3 PREDICTION INDISTINGUISHABILITY

We turn our attention to an idealized notion of indistinguishability, which we refer to as prediction indistinguishability (PI). Distinguishers receive as input an individual-outcome pair $(i, o_i^*) \sim \mathcal{D}^*$ from Nature's distribution, and either Nature's prediction p_i^* or the model's estimate of the parameter $\tilde{p}_i$. We show that achieving PI may require learning Nature's predictor p^* very precisely, even when $\mathcal{A}$ is a very simple class of distinguishers. This result shows that PI is generally infeasible due to the ability to access p_i^* directly: even computationally-weak PI distinguishers are incredibly powerful at distinguishing between p^* and $\tilde{p}$. In a sense, the hardness of PI motivates our focus on OI.

Statistical closeness through PI. Prediction indistinguishability requires that the joint distribution of such individual-outcome-prediction triples cannot be significantly distinguished by a family of algorithms $\mathcal{A}$.

Definition 3.1 (Prediction Indistinguishability). *Fix Nature's distribution $\mathcal{D}^*$. Let $\mathcal{Z} = \mathcal{X} \times \{0,1\} \times [0,1]$. For a class of distinguishers $\mathcal{A} : \mathcal{Z} \to [0,1]$ and $\varepsilon > 0$, a predictor $\tilde{p} : \mathcal{X} \to [0,1]$ satisfies $(\mathcal{A}, \varepsilon)$-Prediction Indistinguishability (PI) if for every $A \in \mathcal{A}$,*

$$\left| \Pr_{(i,o_i^*)\sim\mathcal{D}^*}\left[A(i, o_i^*, p_i^*) = 1\right] - \Pr_{(i,o_i^*)\sim\mathcal{D}^*}\left[A(i, o_i^*, \tilde{p}_i) = 1\right]\right| \leq \varepsilon.$$

We emphasize that prediction indistinguishability departs from outcome indistinguishability in an essential way, by assuming the distinguisher may receive direct access to Nature's prediction p_i^*.[9]

We show that prediction indistinguishability is too strong a notion of indistinguishability to be broadly useful. Specifically, we show that using a very simple distinguisher, we can test for statistical closeness between nature's predictor p^* and the model's predictor $\tilde{p}$. Given the hardness of recovering individual-level predictions in statistical distance (both information-theoretic and computational), this reduction allows us to conclude that, in general, prediction indistinguishability is infeasible.

Consider the randomized distinguisher A_{ℓ_1} defined as follows.

$$A_{\ell_1}(i, o_i, p_i) = \begin{cases} 0 & \text{w.p. } |o_i - p_i| \\ 1 & \text{o.w.} \end{cases}$$

We argue that if a candidate $\tilde{p}$ passes this single PI-distinguisher, it must have small statistical distance to p^*.

Proposition 3.2. *Fix Nature's distribution $\mathcal{D}^*$ and constant $\varepsilon, \tau \geq 0$; suppose Nature's predictor $p^* : \mathcal{X} \to [0,1]$ is such that $p^* = f + \delta$ for a boolean function $f : \mathcal{X} \to \{0,1\}$ and $\delta : \mathcal{X} \to [-1,1]$ where $\|\delta\|_1 \leq \tau$. Then any $(\{A_{\ell_1}\}, \varepsilon)$-PI predictor $\tilde{p} : \mathcal{X} \to [0,1]$ is statistically close to p^*, satisfying*

$$\left\|p^* - \tilde{p}\right\|_1 \leq 4\tau + \varepsilon.$$

Proof. Consider the difference in probabilities of acceptance under that natural and modeled distributions.

$$\begin{aligned}
&\Pr_{(i,o_i^*)\sim\mathcal{D}^*}\left[A_{\ell_1}(i, o_i^*, p_i^*) = 1\right] - \Pr_{(i,o_i^*)\sim\mathcal{D}^*}\left[A_{\ell_1}(i, o_i^*, \tilde{p}_i) = 1\right] \\
&\qquad = \mathop{\mathbf{E}}_{i\sim\mathcal{D}_{\mathcal{X}}}\left[p_i^* \cdot (p_i^* - \tilde{p}_i) + (1 - p_i^*) \cdot (\tilde{p}_i - p_i^*)\right] \qquad (1)
\end{aligned}$$

Assuming that $\tilde{p}$ is $(\{A_{\ell_1}\}, \varepsilon)$-PI, we can upper bound this quantity by ε. Under the assumption that $p^* = f + \delta$ for boolean f, we will

[9] The assumption that p^* meaningfully exists such that p_i^* can be given as input to a distinguisher breaks the abstraction of $\mathcal{D}^*$, but is a common assumption in the forecasting literature. Still, this is another sense in which PI is an idealized variant of OI, because we can never actually generate individual-outcome-prediction samples from $\mathcal{D}^*$.

lower bound the quantity in terms of $\|p^* - \tilde{p}\|$ and τ.

$$
\begin{aligned}
&= \mathop{\mathbf{E}}_{i\sim\mathcal{D}_X}\left[\ (f_i+\delta_i)\cdot(f_i+\delta_i-\tilde{p}_i)+(1-f_i-\delta_i)\cdot(\tilde{p}_i-f_i-\delta_i)\ \right]\\
&\geq \mathop{\mathbf{E}}_{i\sim\mathcal{D}_X}\left[\ f_i\cdot(f_i-\tilde{p}_i)+(1-f_i)\cdot(\tilde{p}_i-f_i)\ \right]-3\left\|\delta\right\|_1\\
&= \mathop{\mathbf{E}}_{i\sim\mathcal{D}_X}\left[\ |f_i-\tilde{p}_i|\ \right]-3\left\|\delta\right\|_1\\
&\geq \left\|p^*-\tilde{p}\right\|_1+4\tau
\end{aligned}
$$

Thus, in combination, we can conclude $\|p^* - \tilde{p}\|_1 - 4\tau \leq (1) \leq \varepsilon$ and the proposition follows. □

We can therefore port any hardness results for recovering p^* in statistical distance to obtaining prediction indistinguishability. For example, if we take p^* to be a random boolean function, then ℓ_1-recovery is information-theoretically impossible unless we observe the outcome o_i^* for a $1 - O(\varepsilon)$ fraction of inputs $i \in \mathcal{X}$. If we restrict ourselves to relatively simple functions, ℓ_1-recovery may be information-theoretically feasible, but computationally infeasible: for instance, if p^* is a pseudorandom function, then any computationally-efficient estimate of $\tilde{p}$ will fail $(\{A_{\ell_1}\}, \varepsilon)$-PI.

REFERENCES

[1] Brian Axelrod, Shivam Garg, Vatsal Sharan, and Gregory Valiant. 2019. Sample Amplification: Increasing Dataset Size even when Learning is Impossible. *arXiv preprint arXiv:1904.12053* (2019).

[2] Marshall Ball, Alon Rosen, Manuel Sabin, and Prashant Nalini Vasudevan. 2017. Average-case fine-grained hardness. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 483–496.

[3] Marshall Ball, Alon Rosen, Manuel Sabin, and Prashant Nalini Vasudevan. 2018. Proofs of Work From Worst-Case Assumptions. In *Advances in Cryptology - CRYPTO 2018 - 38th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2018, Proceedings, Part I (Lecture Notes in Computer Science)*, Hovav Shacham and Alexandra Boldyreva (Eds.), Vol. 10991. Springer, 789–819. https://doi.org/10.1007/978-3-319-96884-1_26

[4] Avrim Blum and Thodoris Lykouris. 2019. Advancing subgroup fairness via sleeping experts. *arXiv preprint arXiv:1909.08375* (2019).

[5] Avrim Blum and Yishay Mansour. 2007. From external to internal regret. *Journal of Machine Learning Research* 8, Jun (2007), 1307–1324.

[6] Kai-Min Chung, Edward Lui, and Rafael Pass. 2013. Can theories be tested?: a cryptographic treatment of forecast testing. In *Innovations in Theoretical Computer Science, ITCS '13, Berkeley, CA, USA, January 9-12, 2013*, Robert D. Kleinberg (Ed.). ACM, 47–56. https://doi.org/10.1145/2422436.2422443

[7] AP Dawid. 1982. *Objective probability forecasts'*. Technical Report. Research Report 14, Department of Statistical Science, University College London.

[8] Philip Dawid. 2015. On individual risk. *Synthese* 194, 9 (Nov 2015), 3445–3474. https://doi.org/10.1007/s11229-015-0953-4

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[10] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. 2019. Learning from outcomes: Evidence-based rankings. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 106–125.

[11] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. 2020. Outcome Indistinguishability. arXiv:2011.13426 [cs.LG]

[12] Cynthia Dwork and Moni Naor. 1992. Pricing via Processing or Combatting Junk Mail. In *Advances in Cryptology - CRYPTO '92, 12th Annual International Cryptology Conference, Santa Barbara, California, USA, August 16-20, 1992, Proceedings (Lecture Notes in Computer Science)*, Ernest F. Brickell (Ed.), Vol. 740. Springer, 139–147. https://doi.org/10.1007/3-540-48071-4_10

[13] Lance Fortnow and Rakesh V. Vohra. 2009. The Complexity of Forecast Testing. *Econometrica* 77, 1 (2009), 93–105. https://doi.org/10.3982/ECTA7163 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA7163

[14] Dean P Foster and Rakesh V Vohra. 1998. Asymptotic calibration. *Biometrika* 85, 2 (1998), 379–390.

[15] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[16] Drew Fudenberg and David K Levine. 1999. An easier way to calibrate. *Games and economic behavior* 29, 1-2 (1999), 131–137.

[17] Oded Goldreich. 2006. *Foundations of Cryptography: Volume 1, Basic Tools*. Cambridge University Press, USA.

[18] Oded Goldreich. 2008. *Computational Complexity: A Conceptual Perspective* (1 ed.). Cambridge University Press, USA.

[19] Oded Goldreich. 2009. *Foundations of Cryptography: Volume 2, Basic Applications*. Cambridge University Press.

[20] Oded Goldreich and Leonid A Levin. 1989. A hard-core predicate for all one-way functions. In *Proceedings of the twenty-first annual ACM symposium on Theory of computing*. 25–32.

[21] Oded Goldreich and Guy N. Rothblum. 2018. Counting t-Cliques: Worst-Case to Average-Case Reductions and Direct Interactive Proof Systems. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, Mikkel Thorup (Ed.). IEEE Computer Society, 77–88. https://doi.org/10.1109/FOCS.2018.00017

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[23] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[24] David Haussler. 1992. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and computation* 100, 1 (1992), 78–150.

[25] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy Rothblum. 2018. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. 1939–1948.

[26] Christopher Jung, Changhwa Lee, Mallesh M Pai, Aaron Roth, and Rakesh Vohra. 2020. Moment Multicalibration for Uncertainty Estimation. *arXiv preprint arXiv:2008.08037* (2020).

[27] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. 2564–2572.

[28] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 100–109.

[29] Michael J Kearns and Robert E Schapire. 1994. Efficient distribution-free learning of probabilistic concepts. *J. Comput. System Sci.* 48, 3 (1994), 464–497.

[30] Michael J Kearns, Robert E Schapire, and Linda M Sellie. 1994. Toward efficient agnostic learning. *Machine Learning* 17, 2-3 (1994), 115–141.

[31] Michael P. Kim, Amirata Ghorbani, and James Zou. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 247–254.

[32] Michael P. Kim, Omer Reingold, and Guy N. Rothblum. 2018. Fairness Through Computationally-Bounded Awareness. *Advances in Neural Information Processing Systems* (2018).

[33] Diederik P Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML]

[34] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[35] Richard J. Lipton. 1989. New Directions In Testing. In *Distributed Computing And Cryptography, Proceedings of a DIMACS Workshop, Princeton, New Jersey, USA, October 4-6, 1989 (DIMACS Series in Discrete Mathematics and Theoretical Computer Science)*, Joan Feigenbaum and Michael Merritt (Eds.), Vol. 2. DIMACS/AMS, 191–202. https://doi.org/10.1090/dimacs/002/13

[36] Nick Littlestone and Manfred K Warmuth. 1994. The weighted majority algorithm. *Information and computation* 108, 2 (1994), 212–261.

[37] Alvaro Sandroni. 2003. The reproducible properties of correct forecasts. *International Journal of Game Theory* 32, 1 (2003), 151–159.

[38] Alvaro Sandroni, Rann Smorodinsky, and Rakesh V Vohra. 2003. Calibration with many checking rules. *Mathematics of operations Research* 28, 1 (2003), 141–153.

[39] Eliran Shabat, Lee Cohen, and Yishay Mansour. 2020. Sample Complexity of Uniform Convergence for Multicalibration. *arXiv preprint arXiv:2005.01757* (2020).

[40] Shai Shalev-Shwartz and Shai Ben-David. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge university press.

[41] Luca Trevisan and Salil P. Vadhan. 2007. Pseudorandomness and Average-Case Complexity Via Uniform Reductions. *Comput. Complex.* 16, 4 (2007), 331–364. https://doi.org/10.1007/s00037-007-0233-x

[42] Salil P. Vadhan. 2012. *Pseudorandomness*. Now Publishers Inc., Hanover, MA, USA.

[43] Leslie G Valiant. 1984. A theory of the learnable. *Commun. ACM* 27, 11 (1984), 1134–1142.

[44] Shengjia Zhao, Tengyu Ma, and Stefano Ermon. 2020. Individual Calibration with Randomized Forecasting. *arXiv preprint arXiv:2006.10288* (2020).