Outcome Indistinguishability*

Cynthia Dwork
Harvard University
dwork@seas.harvard.edu

Michael P. Kim
UC Berkeley
mpkim@berkeley.edu

Omer Reingold
Stanford Unviersity
reingold@stanford.edu

Guy N. Rothblum
Weizmann Institute of Science
rothblum@alum.mit.edu

Gal Yona
Weizmann Institute of Science
gal.yona@weizmann.ac.il

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.

Our Contributions.
(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 [1]. 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 [1] 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 $1/\varepsilon$.
Our findings reveal that Outcome Indistinguishability behaves qualitatively differently than previously studied notions of indistinguishability. Specifically, our 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.

References

[1] Ú. Hébert-Johnson, M. P. Kim, O. Reingold, and G. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *ICML*, pages 1939–1948, 2018.

*The full version of the paper, including author acknowledgments, is available here: https://arxiv.org/abs/2011.13426