Introduction While there is a thriving discourse on various techniques and tradeoffs inherent in making machine learning (and classification in particular) "fair", almost no attention has been paid to the problem of how, once a fair model has been created, one can publicly prove that the model being used is in fact fair. If, for example, a bank wants to demonstrate to the world that its loan policies are fair, it likely would not want to release its entire history of applications, rejections and acceptances. Public verifiability is a necessary component of building fair systems, and future research efforts should consider how to incorporate it into machine learning systems.

Prior Work Most research to this point has focused on the question of what it means for a machine learning algorithm to be fair, either by exploring various fairness metrics, or by measuring discriminatory effects of machine learning algorithms in the real world. While a necessary component for achieving fairness in machine learning, this line of work has not yet started to interrogate issues of trust. In these works, the authors take the position of the model builder, with unlimited access to all data and model decisions.

Prior work in this area has focused most particularly on the problem of binary classification and selection. There are two main strands of thought in this area, herein referred to as the *bandit* and *error rate* approaches. The *bandit* approach, first conceived of in [7], defines unfairness in terms of knowledge. The *bandit* approach gets its name from the term one-armed bandit, which refers to slot machines. In this approach the learner must select bandits to sample from at each time-step. In the naive learning case, when a learner is relatively certain about a majority group of distributions, but uncertain about a minority, it will tend to select those which it is relatively certain about. To know whether or not a process is fair using the *bandit* approach, one needs to know the history of choices the learner was presented with, the choices made by the learner, and the resulting payoffs.

The *error rate* approach refers to a much more disparate set of work, with varying desiderata and strictness of conditions. In this approach, fairness is approached in terms of a combination of the false positive or false negative rates within groups. Some prominent examples of this approach are [5], which takes inspiration from the legal notion of disparate impact, or [6] which requires false positive and false negative rates to be equal across groups. Using this definition, to know whether a classification is fair, one must compute the error statistics within sensitive groups.

Other work has focused on understanding or approximating the influence of variables on black-box models, however this strand also generally assumes access either to model parameters or the ability to submit arbitrary data to the model [1], [3], [2] and furthermore has no formal criterion to which the classifiers must adhere. Overall, nearly all prior work on fairness in machine learning has taken the perspective of the machine learning practitioner, earnestly trying to prevent discriminatory behavior.

While one might hope that all implementers of machine learning models take such a perspective, it would also be nice to have some stronger guarantees. Josh Kroll provides a jumping off point for these guarantees in [8], which describes a protocol for proving to a regulatory observer that an authority correctly followed a secret policy when processing people's data. The applications to fairness here are obvious, and indeed in chapter 6, Kroll shows how Dwork et. al.'s "fairness through awareness" [4] condition can be enforced through this framework. However, accountable algorithms do not allow the people who are actually being classified to verify that their treatment was fair, the subjects must trust the observer investigate and act on unfair actions.

To the best of our knowledge, nobody has yet investigated a trust model where the fairness of the model is publicly verifiable. The accountable algorithms paradigm cannot quite approach that point because of its need to keep the policy details entirely confidential from the data subjects. Since all fairness properties are necessarily informative about the classification policy, the confidentiality condition will necessarily be violated.

Publicly verifiable fairness Relaxing the confidentiality requirement permits the construction of publicly verifiable fair classification protocols. That is, protocols which allow anyone to verify that the classification being carried out is fair, according to some predefined fairness metric. This is a desirable feature for fair systems, because it is a public and transparent commitment to fairness. In many applications of fair systems, a regulatory observer may not exist, may not be interested in engaging in the formal verification process, or may not be a trustworthy regulator. Public verification avoids these pitfalls (which could also be considered excuses!).

At first glance, it may seem that the accountable algorithms paradigm could be trivially modified to take advantage of this change in guarantees. One could try simply allowing anyone request the policy, and associated proof that the policy has been followed. This approach however raises privacy concerns. In this case, the policy is a machine learning model and prior work has shown that it is possible to reconstruct a model based on an observation of inputs and outputs [10], and furthermore that given a model, one can infer training group membership [9]. It is true that under this paradigm, as in under accountable algorithms, data subjects could collaborate, and thereby carry out model stealing and then privacy violating inference attacks. However model inference is an issue with the classifier, rather than the model, and should be addressed through classifier choice rather than threat model. Since some differentially private classifiers do encode the true model parameters, preventing model inference will not repair the naive scheme either.

Future work taking up the problem of proving fairness in machine learning should seriously consider the problem of public verifiability. While the literature has made frequent use of legal or regulatory regimes as motivation, model builders may be interested in fairness out of purely ethical concerns as well. In these cases, a model builder may want to prove that their algorithm is always fair, but there may be no trusted authority to carry out that verification. Furthermore, many of the domains where fairness in machine learning is the most necessary (e.g. banking) are subject to regulatory capture, so generally data subjects may not trust an observer to act in their interests. Public verifiability is a desirable component of fair machine learning, and should be considered further.

References

[1] Julius Adebayo and Lalana Kagal. Iterative Orthogonal Feature Projection for Diagnosing Bias in Black-Box Models. In *FAT/ML*, volume 37, 2016.

[2] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpretability via Model Extraction. *CoRR*, abs/1706.0, 2017.

[3] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems BT - 2016 IEEE Symposium on Security and Privacy, SP 2016, May 23, 2016 - May 25, 2016. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 598–617, 2016.

[4] Cynthia Dwork and Moritz Hardt. Fairness Through Awareness. *arXiv*, 2011.

[5] Michael Feldman, Sorelle Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. *arXiv Prepr. arXiv1412.3756*, pages 1–28, 2014.

[6] Moritz Hardt, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. *Nips*, pages 3315–3323, 2016.

[7] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in Reinforcement Learning. In *International Conference on Machine Learning*, pages 1617—-1626, 2016.

[8] Joshua Kroll. *Accountable Algorithms (A Provocation)*. PhD thesis, Princeton, 2016.

[9] Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership Inference Attacks against Machine Learning Models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 3—-18, 2016.

[10] Florian Tramèr, Fan Zhang, Ari Juels, Michael Reiter, and Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. *Usenix Security*, (Ml), 2016.