Testimony of Prof. Michael Kearns
House Financial Services Committee
Task Force on Artificial Intelligence
February 12, 2020

My name is Michael Kearns, and I am a professor in the Computer and Information Science Department at the University of Pennsylvania. I hold a PhD in computer science from Harvard University, and for more than three decades my research has focused on machine learning and related topics. I have consulted extensively in the technology and finance sectors, including on legal and regulatory matters. I discuss the topics in these remarks at greater length in the recent book *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* [1].

The use of machine learning for algorithmic decision-making has become ubiquitous in the finance industry and beyond. It is applied in consequential decisions for individual consumers (such as lending or credit scoring), in the optimization of electronic trading algorithms at large brokerages, and in making forecasts of directional movement or volatility in markets and individual assets. With major exchanges now being almost entirely electronic, and with the speed and convenience of the consumer Internet, the benefits of being able to leverage large-scale, fine-grained historical data sets via machine learning have become apparent.

The dangers and harms of machine learning have also recently alarmed both scientists and the general public. These include violations of fairness (such racial or gender discrimination in lending or credit decisions) and privacy (such as leaks of sensitive personal information). It is important to realize that these harms are generally not the result of human malfeasance, such as racist or incompetent software developers. Rather, they are the unintended consequences of the very scientific principles underlying machine learning.

Machine learning proceeds by fitting a statistical model to a training data set. In a consumer lending application, such a data set might contain demographic and financial information derived from past loan applicants, along with the outcomes of granted loans. Machine learning is applied to find a model that can predict loan default probabilities from the properties of applicants, and to make lending decisions accordingly. Because the usual goal or objective is exclusively the accuracy of the model, discriminatory behavior can be inadvertently introduced. For example, if the most accurate model overall has a significantly higher false rejection rate on black applicants than on white applicants, the standard methodology of machine learning will indeed incorporate this bias. Minority groups often bear the brunt of such discrimination since by definition they are less represented in the training data.

Note that such biases routinely occur even if the training data itself is collected in an unbiased fashion, which is rarely the case. Truly unbiased data collection requires a period of what is known as *exploration* in machine learning, which is rarely applied in practice because it involves (for instance) granting loans randomly, without regard for the properties of applicants. When the training data is already biased, and the basic principles of machine learning can amplify such biases or introduce new ones, we should expect discriminatory behavior of various kinds to be the norm and not the exception.

Fortunately, there is help on the horizon. There is now a large community of machine learning researchers who explicitly seek to modify the classical principles of machine learning in a way that avoids or reduces sources of discriminatory behavior. For instance, rather than simply finding the model that maximizes predictive accuracy, we can add the constraint that our model must not have significantly different false rejection rates across different racial groups. This constraint can be seen as forcing a balance between accuracy and a particular notion of algorithmic fairness. The modified methodology generally requires us to specify what groups or attributes we wish to protect (such as racial or gender), and what harms we wish to protect them from (such as high false rejection rates). These choices will always be specific to the context under consideration, and should be made by key stakeholders. The algorithms required to enforce fairness constraints are often more complex than the standard ones of machine learning, but not excessively so.

There are some important caveats to this agenda. First of all, there are “bad” definitions of fairness that should be avoided. One example is forbidding the use of race in lending decisions in the hope that it will prevent racial discrimination. It doesn’t, largely because there are so many other variables strongly correlated with race that machine learning can discover as proxies. Even worse, one can show simple examples where such restrictions will in fact harm the very group we sought to protect [1]. Unfortunately, to the extent that consumer finance law incorporates fairness considerations, they are usually of this flawed form that restricts model inputs. It is usually far better to explicitly constrain the model’s output behavior (as in the example of equalizing false rejection rates in lending).

It is also inevitable that constraining models to be fair will cause them to be less accurate, because we are specifying additional conditions to be met beyond just accuracy. Such trade-offs can and should be made quantitative --- for instance, by varying how much disparity we allow in false rejection rates across racial groups (from 0 percent disparity to 100 percent disparity), we can trace out the numerical curve of accuracies that can be achieved for each disparity. This is as far as science can take us --- again, stakeholders must decide what is the right accuracy-fairness balance. We must also be cognizant of the fact that different notions of fairness may be in competition with each other as well. For example, it is entirely possible that by asking for more fairness by race, we must suffer less fairness by gender. These are painful but unavoidable scientific truths.

I note in closing that while my remarks have focused on the potential for designing algorithms that are better behaved, they also point the way to regulatory reform, since most notions of algorithmic fairness (as well as other social norms such as privacy) can also be algorithmically audited. If we are concerned over false rejection disparities by race, we can systematically test models for such behaviors and measure the violations. I believe that the consideration of such algorithmic regulatory mechanisms is both timely and necessary, and I have elaborated on this in other recent writings [1,2].

Citations and Further Reading:

[1] *The Ethical Algorithm: The Science of Socially Aware Algorithm Design.* Michael Kearns and Aaron Roth. Oxford University Press, 2019.

[2] *Ethical Algorithm Design Should Guide Technology Regulation.* Michael Kearns and Aaron Roth. Brookings Institution Policy Brief, 2020. Available at https://www.brookings.edu/research/ethical-algorithm-design-should-guide-technology-regulation/