Testimony to the House Committee on Financial Services Task Force on Artificial Intelligence

Hearing: "Equitable Algorithms: Examining Ways to Reduce AI Bias in Financial Services"

February 12, 2020

Submitted by Dr. Philip S. Thomas

Assistant Professor, University of Massachusetts Amherst

Chairman Foster, Ranking Member Loudermilk, and members of this task force, thank you for the opportunity to testify today.

I am Philip Thomas, an assistant professor at the University of Massachusetts Amherst. My goal as a machine learning researcher is to ensure that machine learning algorithms are safe and fair – properties that may be critical for the responsible use of AI in finance.

Towards this goal, in a recent *Science* paper, my co-authors and I proposed a new type of machine learning algorithm, which we call a *Seldonian* algorithm. Seldonian algorithms make it easier for the people using AI to ensure that the systems they create are safe and fair. We have shown how Seldonian algorithms can avoid unfair behavior when applied to a variety of applications including optimizing online tutorials to improve student performance, influencing criminal sentencing, and deciding which loan applications should be approved.

While our work with loan application data may appear most relevant to this task force, that work was in a subfield of machine learning called *contextual bandits*. The added complexity of the contextual bandit setting would not benefit this discussion, and so I will instead focus on an example in a more common and straightforward setting called *regression*. In this example, we used entrance exam scores to predict what the GPAs of new university applicants would be if they were accepted. This GPA prediction problem resembles many problems in finance, for example rating applications for a job or loan. The fairness issues that I will discuss are the same across all these applications.

In the GPA prediction study, we found that three standard machine learning algorithms over-predicted the GPAs of male applicants on average and under-predicted the GPAs of female applicants on average, with a total bias of around 0.3 GPA points in favor of male applicants. A Seldonian algorithm successfully limited this bias to below 0.05 GPA points with only a small reduction in predictive accuracy.

The rapidly growing community of machine learning researchers studying issues related to fairness has produced many similar AI systems that can effectively preclude a variety of types of unfair behavior across a variety of applications. With the development of these fair algorithms, machine learning is reaching the point where it can be applied responsibly to financial applications, including influencing hiring and loan approval decisions.

I will now discuss technical issues related to ensuring the fairness of algorithms, which might inform future regulations aimed at ensuring the responsible use of AI in finance. First, there are many definitions of fairness. Consider our GPA-prediction example:

- One definition of fairness requires the average predictions to be the same for each gender. Under this definition, a system that tends to predict a lower GPA if you are of a particular gender would be deemed unfair.
- Another definition requires the average error of predictions to be the same for each gender. Under this definition, a system that tends to over-predict GPAs for one gender and under predict for another would be deemed unfair.

Although both of these might appear to be desirable requirements for a fair system, for this problem it is not possible to satisfy both simultaneously. Any system, human or machine, that produces the same average prediction for each gender necessarily over-predicts more for one gender, and vice versa. The machine learning community has generated more than twenty possible definitions of fairness, many of which are known to be incompatible in this way.

In any effort to regulate the use of machine learning to ensure fairness, a critical first step is to define precisely what fairness means. This may require recognizing that certain behaviors that appear to be unfair may necessarily be permissible, in order to enable enforcement of a conflicting and more appropriate notion of fairness. Although the task of selecting the appropriate definition of fairness should likely fall to regulators and social scientists, machine learning researchers can inform this decision by providing guidance with regard to which definitions are possible to enforce simultaneously, what unexpected behavior might result from a particular definition of fairness, and how much or little different definitions of fairness might impact profitability.

Regulations could also protect companies. Fintech companies that make every attempt to be fair, using AI systems that satisfy a reasonable definition of fairness, may still be accused of racist or sexist behavior for failing to enforce a conflicting definition of fairness. Regulation could protect these companies by providing an agreed-upon, appropriate, and satisfiable definition of what it means for their systems to be fair.

Once a definition of fairness has been selected, machine learning researchers can work on developing algorithms that will enforce the chosen definition. For example, our latest Seldonian algorithms are already compatible with an extremely broad class of fairness definitions and might be immediately applicable. Still, there is no “silver bullet” algorithm for remedying bias and discrimination in AI. The creation of fair AI systems may require use-specific considerations across the entire AI pipeline, from the initial collection of data through to monitoring the final deployed system.

Another observation that might inform efforts at regulation is that, for many reasonable definitions of fairness, it is not possible to ensure with certainty that any system, human or

machine, is fair. Any data used to evaluate the fairness of a system might not be representative of the actual population that the system will be applied to in the future. So, a system that appears to be fair based on the available data may not actually be fair. However, as we obtain more data, we can become increasingly confident that the data resembles the larger population, and hence that the system will be fair when used. In this way, when fairness cannot be guaranteed with certainty, it can usually be guaranteed with high probability. While this motivated my research into creating systems that are safe and fair with high probability, this observation might also inform how AI systems are regulated. Requiring companies using AI to ensure that their systems are fair with certainty may be asking the impossible. Hence, one might regulate the process rather than the outcome – to require the use of algorithms that are fair with high probability and the use of mechanisms to quickly identify and repair unfair behavior when it inevitably occurs.

Several other questions must be answered for regulations to be effective and fair. For example: Will fairness requirements that appear reasonable in the short-term have the long-term impact of reinforcing existing social inequalities? How should fairness requirements account for the fact that changing demographics can result in a system that was fair last month being unfair today? When unfair behavior occurs, how can regulators determine whether this is due to the aforementioned inevitability of unfair behavior, or the improper use of machine learning?

Thank you again for the opportunity to testify today. I look forward to your questions.