Name of Applicant: Tanvir Ahmed Khan

PhD Candidate, Queen's Economics Department, Queen's University

Dunning Hall, 94 University Ave, Kingston, ON K7L 3N6

Email: 20takt@queensu.ca

Year of Enrollment: Fall 2020

Year of Passing Comprehensive Exams: Fall 2021

Year in Ph.D.: 3rd-year

Research Interests: Economics of AI, Digital Economics, Economics of Innovation, Industrial Organization

Supervisor: Dr. Robert Clark

Title of Current Research Project: **'Can Unbiased Predictive AI Amplify Bias?'**

Advisors: Dr. Robert Clark, Dr. Nahim Zahur, Dr. Shota Ichihashi

Dear Concerned,

I am eager to attend the Digital Economics and AI Tutorial on September 21, 2023. I am a 3rd-year Economics Ph.D. candidate at Queen's University. Dr. Robert Clark is my supervisor. My research interest is the economics of AI, digital economics, the economics of innovation, and industrial organization. In my 2nd year, I took courses in Industrial Organization and Competition Policy (ECON 954), Advanced Topics in Microeconomics focusing on labor economics (ECON 912), Experimental Economics (ECON 913), Advanced Topics in Macroeconomics focusing on structural models (ECON 915) and Empirical Microeconomics focusing on structural models and object-oriented programming approach to solving them (ECON 860). Currently, I am working on a project, intended to become a chapter of my PhD thesis, titled ***'Can Unbiased Predictive AI Amplify Bias?'***. As this paper aligns with the topic of the tutorial, I use the rest of my application to highlight an extended abstract of my paper.

In the paper, I show that even a bias-neutral predictive AI can potentially *amplify* exogenous (human) bias in settings where the predictive AI represents a cost-adjusted precision gain to unbiased predictions, and the final judgments are made by biased human evaluators. In the absence of perfect and instantaneous belief updating, '*expected victims of bias' or 'marginal minorities'* become less likely to be *saved by randomness* under more precise predictions if human bias intensity stays fixed in the short run. As a result, discrimination against '*marginal minorities'* is non-decreasing in precision gain. An increase in aggregate

discrimination is possible if this effect dominates. In evaluating AI's impact on discrimination, not accounting for this mechanism may result in AI being unduly blamed for *creating bias* when it is only *amplifying exogenous human bias*. I argue that the distinction is important as a misattribution may slow down the adoption of a beneficial technology.

Predictive algorithms are increasingly used in important decisions of economic and policy consequences. For example, lending institutions use machine learning models to assess creditworthiness. Criminal justice systems use 'predictive recidivism' algorithms to predict recidivism risks and guide bail decisions on defendants. Similar settings exist in healthcare, hiring, law enforcement, university admissions, insurance, etc. How can predictive algorithms exacerbate discrimination? Existing literature recognizes two primary channels - biased model design or biased training data. In this paper, I show the existence of a third channel. An increase in prediction precision, while a desirable feature, might paradoxically increase discrimination for *marginal minorities* due to the decrease in randomness, especially when human bias remains constant over the short term.

The counterintuitive result in this paper arises by capturing in a simple model the interaction between machine predictions and human judgment. This setting with algorithms as prediction machines and humans in the loop making judgments is quite ubiquitous. However, the role of this human-machine interaction has been mostly ignored in the algorithmic bias literature. The adoption of predictive AI is conceptualized as a bias-neutral precision gain for the prediction technology. While the model I study applies to a wide range of settings where predictive AI is used, I consider the setting to be in lending for demonstration. There are two groups of agents in the model, whites, and minorities, with minorities being the discriminated group without loss of generality. Discrimination is defined as the difference in the probability of getting loan approvals between whites and minorities who are otherwise identical in terms of underlying default probability. Aggregate discrimination is defined as the integral of discrimination over the support of default probabilities.

I show that, in the absence of perfect and instantaneous belief updating, which is the likely case in short time horizons over which empirical evaluation of predictive AI's impact on discrimination takes place, discrimination is non-decreasing in precision gain for *marginal minorities* near the cutoff. *Marginal minorities* are the *expected victims of bias*, minority applicants who are expected to be denied loans, but would not be expected to be denied loans if they were white. This result arises because *marginal minorities* become less likely to be *'saved by randomness'* under more precise predictions.

The relaxation of the assumption of perfect and instantaneous belief updating, at least in short time horizons, is further motivated by several features commonly found in the relevant setting being studied. These features are the black-box, trade-secret, and race-blind nature of predictive algorithms, and the limited information and feedback available to human evaluators who make high-volume decisions in fast-paced environments with the possibility of exhaustion.

If the rate of change function of discrimination with precision gain is symmetric or near-symmetric over the support of the default risk distribution, the increase in discrimination for the *marginal minorities* mostly cancels out in aggregate, as there are other minority applicants for whom discrimination decreases. Otherwise, aggregate discrimination can either increase or decrease in response to precision gain, and the magnitude of the effect is increasing in human bias intensity, noisiness of the prediction technology before the adoption of predictive AI, and the extent of precision gain. In general, an increase in aggregate discrimination is possible if the relative mass of *marginal minorities* is large (compared to other minorities), if the distribution of default probabilities is concentrated near the cutoff, and if the first derivative of discrimination with precision is asymmetric over the support of the default risk distribution.

The contribution of this paper is twofold. First, it shows a bias-neutral precision gain of an unbiased prediction algorithm can have a disparate impact on discrimination in the short run, decreasing discrimination for some while increasing discrimination for the *marginal minorities* who are worse off or at least not better off. Second, it shows the counter-intuitive result that even a bias-neutral precision gain of an unbiased prediction algorithm can potentially amplify exogenous (human) bias, resulting in an aggregate rise in discrimination. It also elucidates the underlying mechanism through which this counter-intuitive result arises. Not accounting for this mechanism may result in AI being unduly blamed for creating bias when the impact of prediction AI is being evaluated. Such evaluation often takes place shortly after the adoption of predictive AI and justifies the modeling choice of a short time horizon. Not understanding this mechanism has pitfalls for society at large, especially when fear and distrust of AI is high, as this may inadvertently thwart or slow down the adoption of a beneficial technology.

Considering my research interests and ongoing project, I believe that I can greatly benefit from attending the tutorial. I look forward to learning from research leaders in this field, and the opportunity to discuss ideas and get feedback. My working paper is available upon request. Additionally, I have also submitted it for consideration for the Conference on Economics of Artificial Intelligence on September 22, 2023. I look forward to the possibility of participating in this enriching tutorial.