How Anti-Discriminatory Measures Can Worsen AI Bias

By Anton S. Ovchinnikov , INSEAD and Queen's University, Canada

Removing gender-related data from machine-learning models may paradoxically harm women instead of protecting them.

Algorithms and artificial intelligence (AI) are transforming the ways organisations make decisions in fundamental ways. AI is the engine that revolutionised financial services, enabling home insurance companies to produce a customised quote in **under 60 seconds**. In the lending space, firms may have access to hundreds (and often thousands) of attributes on loan applicants, ranging from personal profile (gender, age, number of children, etc.) to borrowing history. AI has made manually poring through massive volumes of data a thing of the past.

The adoption of AI by organisations, however, has been accompanied by practical and ethical concerns, of which discrimination poses the greatest ethical risk. For example, Apple Card was **accused of discrimination** against women in 2019. The company declined a woman's application for a credit line increase, even though her credit records were better than her husband's. Meanwhile they granted her husband a credit line that was 20 times higher than hers.

The New York State Department of Financial Services found no violations of fair lending by the Apple Card venture since the company had not used gender data in the development of its algorithms. If the company had fully adhered to anti-discrimination laws, what led to the paradoxical outcome?

The effectiveness of regulations

Overcoming gender discrimination by regulatory means is not without challenges. The existing anti-discrimination laws come with their own legal definition of discrimination and these laws are constantly outpaced by rapid advances in AI. As a result, some regulations may paradoxically hurt rather than help the groups they are meant to protect.

While anti-discrimination laws regarding gender and credit differ across countries, their guidance on data management and model-building fall into three main categories. In Singapore, the collection and use of gender data in AI models is allowed, while the legal jurisdiction in the US prohibits the collection and use of gender data. Canada and countries in the European Union are in between these extremes. These jurisdictions permit the collection of gender data but prohibit the use of gender as a feature in the training and screening models used for individual lending decisions.

In a recent **research paper**, we study how the exact same applicant could be issued credit by a financial service firm operating under one jurisdiction but rejected under another because of how the respective anti-discrimination laws limit the machine learning (ML) models these firms could develop to support lending decisions.

Getting to the root through simulation

My co-authors* and I used a realistically large, publicly available dataset on non-mortgage lending from Home Credit, a global fintech company, to train statistical and ML models. Through simulation, we mimicked how a ML algorithm in a fintech firm would decide whether to approve or reject non-mortgage consumer loans depending on which jurisdiction the firm operates in. The study represents an empirical validation of previous studies, shedding light on the real effects of different anti-discrimination laws on gender-based discrimination. This is due to the implications of these laws on data management and model development.

More importantly, we sought to identify the drivers of statistical and ML discrimination. ML algorithms learn or “train” by finding patterns in data that is fed into them so that they can generate accurate predictions for new data. These predictions are used by businesses to make decisions. Along the ML operations pipeline, bias can occur in any part of the process, from selecting data for training, to “feature engineering” (i.e. creating new data to help models learn better) and to choosing the best model and its parameters. The question is how various regulations impact these steps and what that means for discrimination.

A naive approach is to remove the protected attributes from the training data, hoping that this would also eliminate discrimination. Goldman Sachs, a partner in the Apple Card venture, stated: “We have not and never will make decisions based on factors like gender...**we do not know your gender or marital status**.”

Why discrimination persists when protected attribute data is not considered

Goldman’s statement makes intuitive sense: if one does not consider gender (or other protected attributes), they cannot discriminate based on it. Unfortunately, this logic is flawed. To understand why, consider the example illustrated in Figure 1.

Data shows that, all things being equal, women are better borrowers than men, and individuals with more work experience are better borrowers than those with less. Thus, a woman with three years of work experience could be as creditworthy as a man with five years of experience. For the purpose of this example, suppose that both would be deemed good enough (i.e. above the lender’s risk threshold) to be granted loans. In contrast, suppose that in this illustration, men with three years of work experience would not be creditworthy enough (i.e. fall below threshold) and should be denied loans.

However, data also shows that women tend to have less work experience than men on average. In addition, the dataset used to train AI algorithms, which comprises information of past borrowers, consists of about 80 percent men and 20 percent women on average globally. Thus, when the algorithm does not have access to gender data, it treats individuals with the same number of years of experience equally. Since women represent a minority of past borrowers, it is unsurprising that the algorithm would predict the average person to behave like a man rather than a woman.

That is, applicants with five years of experience would be granted credit, while those with three year or less would be denied, regardless of gender. This not only increases discrimination but also hurts profitability, as women with three years of work experience are creditworthy enough and should have been issued loans had the algorithm used gender data to differentiate between women and men.

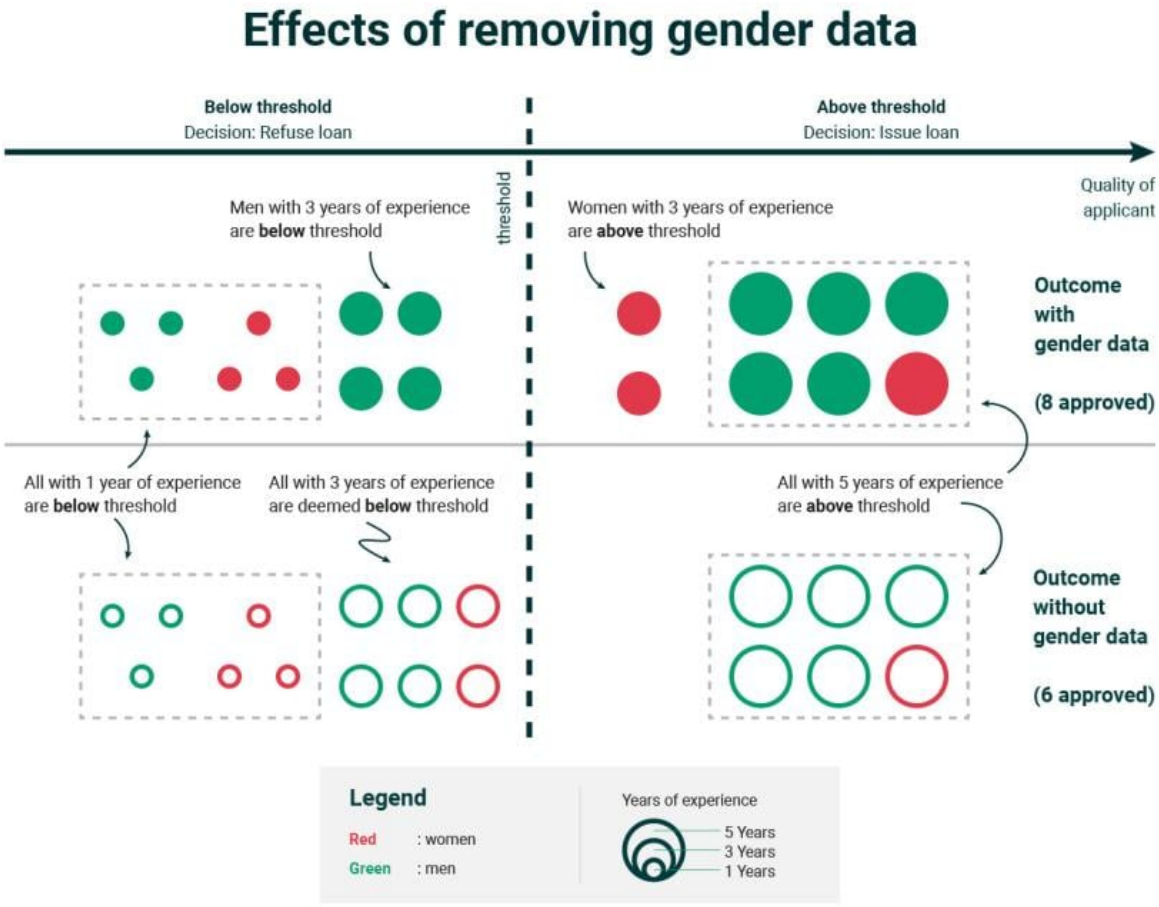

Figure 1. Illustration of the effect of removing gender data on loan application outcomes.

How big is the resultant discrimination and profit loss? Our results show that depriving models of using gender data increases discrimination by nearly 2.8 times, based on positive predictive values. These values represent the percentages of applicants, by gender, who are correctly granted loans. Restricting the use of gender in ML algorithms also hurts profitability by about 0.25 percent as fewer loans are issued. Therefore, jurisdictions like the

US that prohibit the use of gender data substantially increase discrimination and slightly decrease firm profitability. See Figure 1 for the illustration of this logic.

How can companies reduce discrimination?

Our findings show that ML discrimination is driven by the model training process. Because ML algorithms learn from past data, when historical data is skewed, biased predictions become inevitable. What can firms do?

In jurisdictions like Singapore that allow the collection and use of gender, as well as other prohibited attribute data, firms can use such data in their models. This could reduce discrimination and increase profitability, as our analyses and the stylised example in Figure 1 show. In contrast, in jurisdictions where laws prohibit the collection of such data, firms cannot do anything other than perhaps lobbying to change legislation.

The most interesting scenario arises in the “in between” cases, where laws allow the collection of gender data, but not its use in the final models, as is the case in the EU. In such instances, pre-processing data before ML algorithm training could help. For example, firms could rebalance the training data based on gender by reducing the number of men in the training data (downsampling) or increasing the number of women (upsampling). That is, change the data on which the models are trained, but not use gender in actual decision-making.

Generally, these approaches (illustrated in Figures 2a and 2b) reduce discrimination, but at a cost. Although the model still does not know the gender of the applicants, an average applicant with three years of experience now looks more like a woman, i.e. a “good” borrower. As a result, credit would be issued to all applicants with three years of work experience. More women getting loans reduces discrimination, but so do men with three years of experience (who would otherwise fall below the risk threshold), which hurts profitably.

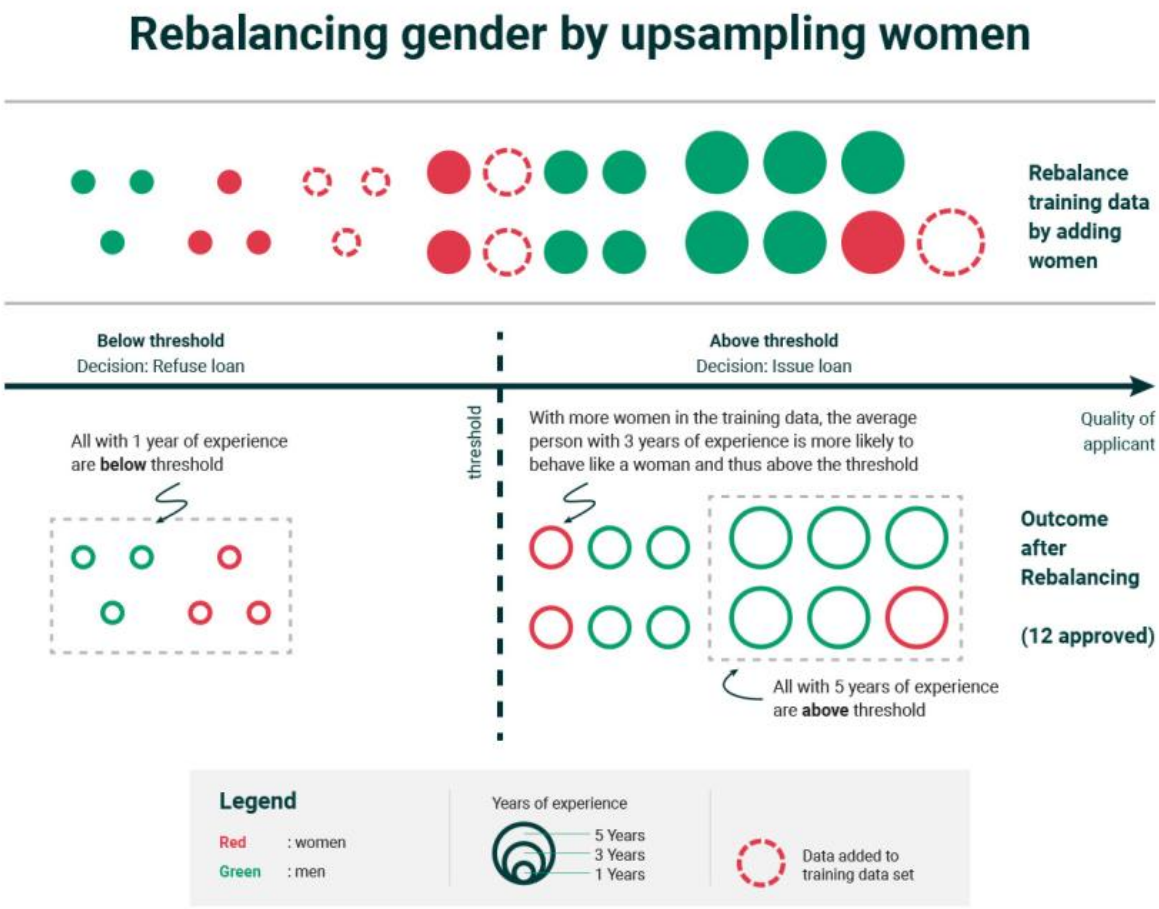

Figure 2a: Illustration of the effect of upsampling women to rebalance training data on loan application outcome.

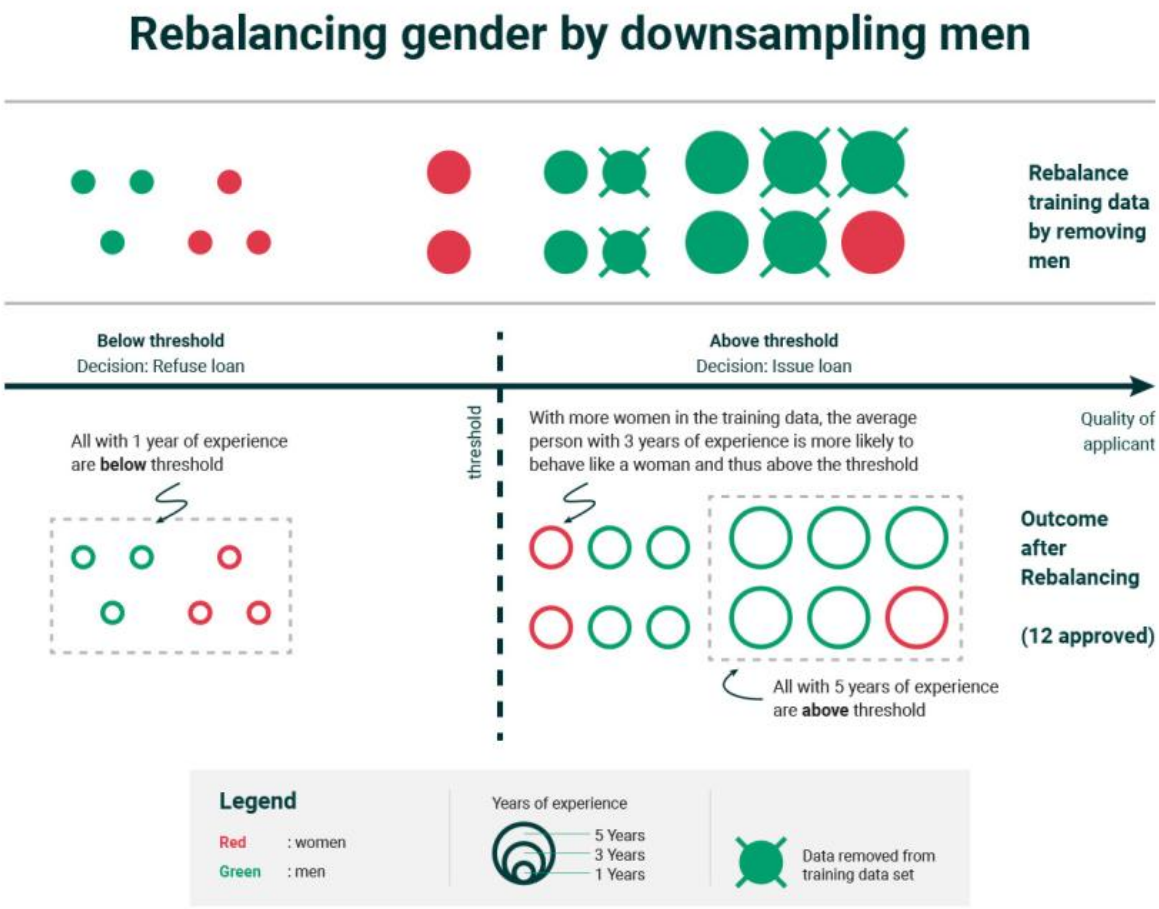

Figure 2b: Illustration of the effect of downsampling men to rebalance training data on loan application outcomes.

A related approach is to tune the model's hyper-parameters with gender, i.e. determine the number of variables in the model with gender data, but determine the coefficients for those variables without gender data, so that an applicants' gender does not impact model predictions. Our results show that this too could reduce discrimination with a minor loss of profitably.

Lastly, firms that operate across multiple jurisdictions have one other "trick" at their disposal. They can use the hundreds of variables on loan applicants in a jurisdiction that allows the collection of gender data to train a secondary model to predict the gender of an applicant in a jurisdiction that prohibits the collection of such data. Then, the predicted gender could be used in lieu of the actual gender in processing loans. According to our study, this process, known as probabilistic gender proxy modelling, can impute gender at up to

91 percent accuracy. Due to such high accuracy, imputing gender reduces discrimination by almost 70 percent and increases profitability by 0.15 percent.

Since all models make mistakes, the efficacy of this approach rests on how well the missing attribute, like gender, can be imputed. Two kinds of mistakes can occur, as illustrated in Figure 3. Incorrectly predicting a male applicant with three years of experience as a woman leads to loss of profitability as the model grants credit when the male applicant should in fact be denied. The opposite mistake of incorrectly predicting a woman to be a man means that a loan that should have been issued was not, thus hurting profitability and amplifying discrimination. Overall, our data shows that if the number of mistakes is small, this approach could work.

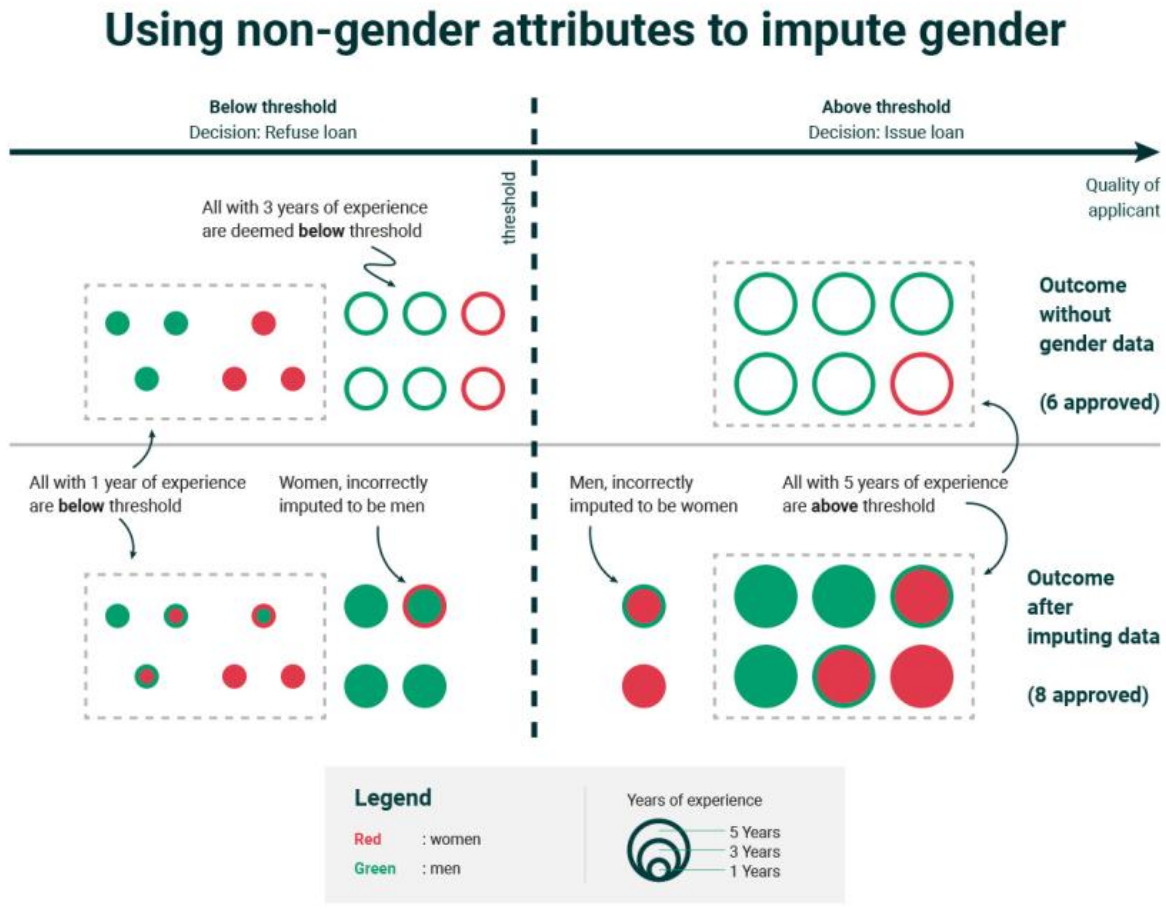

Figure 3: Illustration of the effect of using non-gender attributes to impute data on loan application outcome.

Rethinking approaches to reduce discrimination

We need to rethink the efficacy of anti-discrimination measures and laws, specifically with respect to the collection and use of sensitive data for ML models. Fundamentally, are the measures and laws doing what they set out to do for consumers? Our findings show that access to sensitive data can reduce discrimination substantially and at times increase profitability as well.

The Apple Card example illustrates the risks that businesses face when they are not given access to sensitive data. Importantly, even when businesses are prohibited from using sensitive data, they should be allowed to collect such data in order to assess if their models perpetuate discrimination. If discrimination is identified, they can take steps to reduce it and subsequently evaluate the effectiveness of these steps.

Amazon had trained a ML algorithm to streamline its hiring process by predicting the performance of job applicants based on their resumes. Although the company tried to avoid discrimination by removing gender and names from the data, the algorithm discriminated against women because data of its current employees – predominantly men – was used for training. Thankfully, with access to gender-related data, Amazon was able to **spot the discrimination** before using the algorithm on real applicants.

Companies need to be empowered with data in order to take the necessary steps to reduce discrimination. At the same time, they need to understand that increased data access comes with greater accountability. Government must adopt cognisant regulations to strike a balance between power and responsibility.

* ***Stephanie Kelley***, *Saint Mary University;* ***David R. Hardoon***, *Aboitiz Data Innovation, and* ***Adrienne Heinrich***, *UnionBank of the Philippines*

Find article at
https://knowledge.insead.edu/operations/how-anti-discriminatory-measures-can-worsen-ai-bias

---

About the author(s)

Anton S. Ovchinnikov is a Visiting Professor of Decision Sciences as well as Technology and Operations Management at INSEAD. He is also a Distinguished Professor of Management Analytics and the Scotiabank Scholar of Customer Analytics at the Smith School of Business, Queen's University in Canada.

About the research

"**Antidiscrimination laws, artificial intelligence and gender bias: A case study in nonmortgage fintech lending**" is published in *Manufacturing & Service Operations Management*. This study is named as one of the **Global Top 100 AI solutions** for reaching the UN Sustainable Development Goals by the International Research Centre in Artificial Intelligence (IRCAI) under the auspices of UNESCO. It also won the Best Paper award of the Special Interest Group on the interface between finance and operations management (iFORM SIG) of the INFORMS Manufacturing and Services Operations Management (MSOM) society.

About the series

AI: Disruption and Adaptation

Delve deeper into developments in artificial intelligence, especially the disruptions across value chains. This series examines AI's impact on a range of sectors, including business consulting, education and the media. It also sizes up the regulatory and ethical questions tied to this game-changing technology.