ARTICLE

“A $%^* Sexist Program”: Detecting and Addressing AI Bias

ARTIFICIAL INTELLIGENCE

Published on:
DECEMBER 09TH, 2020
Updated on:
AUGUST 16TH, 2021

7 minutes

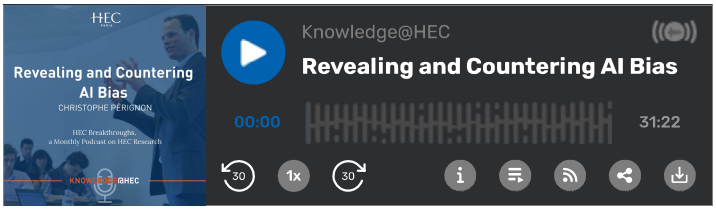
A major issue facing companies that use AI, algorithmic bias can perpetuate social inequalities — as well as pose legal and reputational risks to the companies in question. New research at HEC Paris offers a statistical method of tracking down and eliminating unfairness.

©metamorworks on Adobe Stock

By Christophe Pérignon

DOWNLOAD & PRINT
SHARE
Aa

LISTEN TO THE PODCAST:



Soon after Apple issued its Apple credit card, in August 2019, urgent questions arose. A well-known software developer and author, David Heinemeier Hansson, reported in a Tweet that both he and his wife had applied for the card. “My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time,” Hansson wrote. “Yet Apple’s black box algorithm thinks I deserve 20x the credit limit she does.” He called it a “sexist program,” adding an expletive for good measure.

"If a credit-scoring algorithm is trained on a biased dataset of past decisions by humans, the algorithm would inherit and perpetuate human biases." (Photo ©Siberian Art on Adobe Stock)

Goldman Sachs, the issuing bank for the credit card, defended it, saying that the AI algorithm used to make determinations of creditworthiness didn't even take gender into account. Sounds convincing, except that this ignored the fact that even if specific gender information is removed, algorithms may still use inputs that correlate with gender ("proxies" for gender, such as where a person shops) and thus may still produce unintended cases of bias.

Even Apple's cofounder Steve Wozniak reported that he and his wife had experienced this bias. Wozniak was judged worthy of 10 times more credit than his wife, despite the fact that he and his wife shared assets and bank accounts. The ensuing melee resulted in an investigation of the Apple Card's algorithm by New York regulators.

BIASED DATA LEADS TO BIASED RESULTS

AI/machine learning can process larger quantities of data more efficiently than humans. If applied properly, AI has the potential to eliminate discrimination against certain societal groups. However, in reality, cases of algorithmic bias are not uncommon, as seen in the case of Apple, above.

If a credit-scoring algorithm is trained on a biased dataset of past decisions by humans, the algorithm would inherit and perpetuate human biases.

The reasons for this bias are various. If, for example, a credit-scoring algorithm is trained on a biased dataset of past decisions by humans (racist or sexist credit officers, for example), the algorithm would inherit and perpetuate human biases. Because AI uses thousands of data points and obscure methods of decision making (sometimes described as a black box), the algorithmic biases may be entirely unintended and go undetected.

"When machine learning techniques, which are often difficult to interpret, are poorly applied, they can generate unintended, unseen bias toward entire populations." (Photo ©Nuthawut on Adobe Stock)

In credit markets — the focus of our work — this lack of fairness can place groups that are underprivileged (because of their gender, race, citizenship or religion) at a systemic disadvantage. Certain groups could be unreasonably denied loans, or offered loans at unfavorable interest rates — or given low credit limits. A lack of fairness may also expose the financial institutions using these algorithms to legal and reputational risk.

A "TRAFFIC LIGHT" TEST FOR DETECTING UNFAIR ALGORITHMS

My fellow researchers, Christophe Hurlin and Sebastien Saurin, and I established a statistics-based definition of fairness as well as a way to test for it. To ensure fairness, decisions made by an algorithm should be driven only by those attributes that are related to the target variables, such as employment duration or credit history, but should be independent of gender, for example. Using statistical theory, we derived a formula to compute fairness statistics as well as the theoretical threshold above which a decision would be considered fair.

> We established a statistics-based definition of fairness as well as a way to test for it.

When dealing with an actual algorithm, one can first compute the fairness statistics and compare them to the theoretical value or threshold. It is then possible to conclude whether an algorithm is "green" (when the fairness statistics are greater than our established threshold) or "red" (when the fairness statistics are less than the threshold).

Second, if there is a problem, we offer techniques to detect the variables creating the problem — even if the algorithm's processes are impenetrable. To do so, we developed new AI explainability tools. Third, we suggest ways to mitigate the problem by removing the offending variables.

> We developed new AI explainability tools to detect the variables creating the problem of unfairness.

From a purely practical, business perspective, it is important that banks understand the implications — and potential unintended consequences — of the technology they are using. They may risk running afoul of both the justice system and public opinion — and it goes without saying that reputation and trust are key in the banking industry.

APPLICATION ACROSS DIVERSE FIELDS

While our focus has been on credit scoring, our methodology could potentially be applied in many other contexts in which machine learning algorithms are employed, such as predictive justice (sentencing, probation), hiring decisions (screening of applicants' CVs and videos), fraud detection and pricing of insurance policies.

"Our methodology could potentially be applied in many contexts in which machine learning algorithms are employed, such as predictive justice, hiring decisions, fraud detection and pricing of insurance policies." (Photo ©artinspiring on Adobe Stock)

Applications

In the short run, we aim to help companies and institutions that use AI to better understand the decisions of their algorithms and to detect potential unintended consequences. In the longer term, we hope to contribute to the discussion about guidelines, standards and regulations that public administrations should institute.

[In French on The Conversation] « Un $ % de programme sexiste » : comment détec...

Methodology

...technology ... ...atory questions. When ... which are ... ... applied, they can generate unintended, unseen bias toward entire populations on the basis of ethnic, religious, sexual, racial or social criteria. The opportunities and risks that come with machine learning technologies underscore the call for the implementation of a new form of regulation based on the certification of the algorithms and data used by companies and institutions.

Drawing on work conducted over ... years on risk model validation, we developed new statistical tests that detect a lack of fairness. This "traffic light" test statistically analyzes whether an algorithm's decisions are fair ("green") or unfair ("red") against protected societal groups. If an algorithm's decisions are found to be unfair, we suggest techniques to identify the variables responsible for the bias and to mitigate them.

Based on an interview with Christophe Pérignon and his academic article "The Fairness of Credit Scoring Models," co-written with Christophe Hurlin and Sébastien Saurin, both from the University of Orléans.

RELATED TOPICS: ARTIFICIAL INTELLIGENCE DATA SCIENCE FINANCE DIVERSITY & INCLUSION PODCASTS DOWNLOAD & PRINT SHARE Aa + -

Christophe Pérignon
Professor
Finance

Christophe Pérignon is Associate Dean for Research and Professor of Finance at HEC Paris, France. He is also the co-holder of the ACPR (Banque de...

RELATED CONTENT ON ARTIFICIAL INTELLIGENCE

ARTIFICIAL INTELLIGENCE

Increased Use of AI in Private Equity Will Cause an Industry Shakeout

OCTOBER 11TH, 2021

Thomas Åstebro
Professor