ECONOMIC FINDING • FEBRUARY 2023

Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias

Based on BFI Working Paper 2023-19, "Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias," by Amanda Agan, Rutgers; Diag Davenport, Princeton; Jens Ludwig, UChicago's Harris School of Public Policy; and Sendhil Mullainathan, Chicago Booth

Algorithmic audits of Facebook in the US and India find significant out-group bias in the News Feed algorithm (e.g., whites are less likely to be shown Black friends' posts, and Muslims less likely to be shown Hindu friends' posts), suggesting a need to rethink how large-scale algorithms use data on human behavior.

Prominent among the counterintuitive insights gleaned from behavioral economics is that we often do not choose what we really want. For example, when thinking fast (automatically), cognitive biases take hold and we may choose the donut from the breakfast buffet, but when thinking slow (deliberately) we may choose the banana. In this case, it can be inferred that the banana is our true preference (more on that in a moment).

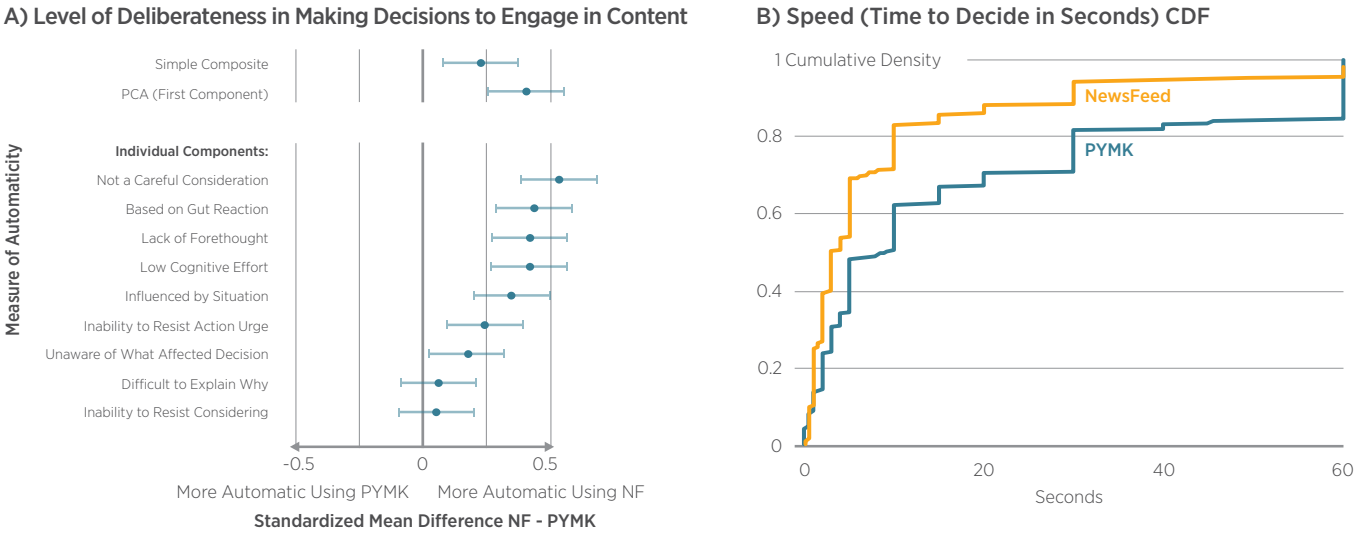
Now imagine that the buffet table is actually Facebook, and the social media platform is offering choices that are curated by an algorithm based on our automatic thinking, meaning that the movies, news stories, posts, books, websites, and so on that Facebook offers us may not reflect our true preferences. In other words, the choices that algorithms offer, which are inferred by reliance on past data and ranked accordingly, are not entirely of our own making. Likewise, when we act automatically on those ranked offerings, we will reinforce the algorithm's choices and the cycle will continue. In such a scenario, our true preferences are often unmet.

This paper explores the implications of this phenomenon in the context of one kind of automatic bias of particular social concern: discrimination. Before describing the authors' methodology and findings, a brief note about terminology: The authors term fast or automatic choices as "system 1" decisions, and more carefully considered choices as "system 2" decisions that likely reflect true preferences. Now, about those "true" preferences: the authors acknowledge that we cannot definitively know what someone truly wants, so by "preferences" the authors are using shorthand for system 2 decisions, given their more considered nature.

Regarding discrimination, algorithmic automaticity can induce prejudice above and beyond any explicit preference because many of the forces that create discrimination operate quickly, via stereotypes or gut responses, and can therefore exert stronger influence over automatic choices. Because behavior is a combination of two distinct forces — prejudice that arises from system 2 and system 1 decisions — the issue at hand

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

Figure 1 • How Automatic are Users When Choosing to Engage with News Feed (NF) Content versus People You May Know (PYMK) Suggestions



Note: This figure illustrates measures of deliberateness in making decisions to engage in content on the News Feed and the People You May Know recommendations, as gathered from an author survey. Panel A shows that for each of nine measures of deliberate interaction, users report higher levels of deliberation when using PYMK than when using News Feed. Panel B shows the cumulative distribution function (CDF) for responses to the one continuous measure, the time it takes the study subject to decide (in seconds), which again shows the same pattern. Please see the working paper for more details.

is the additional bias created by automaticity. Algorithms try to infer preferences from our behavior but the influence of system 1 as reflected in that behavior can lead the algorithm to codify unintended bias. Importantly, the magnitude of that source of bias is not fixed: algorithms will inherit more bias when trained on more automatic behaviors, which the authors describe as particularly troubling because algorithms are often trained in contexts (e.g., social media) in which people behave fairly automatically. Thus, the cycle not only continues, but it also intensifies.

The authors focus on one particular kind of prejudice: the tendency of people to favor those like themselves ("own-group" members) and to disfavor unlike people ("out-group" members). The authors then test the implications of this model in two ways that, together, tell a powerful story. Importantly, the prediction tested here is not so much just about the presence of algorithmic bias, but rather whether the magnitude of such bias will be relatively larger for algorithms trained using behavioral data that are relatively more automatic.

Test #1: Lab Experiment

Subjects are asked to select movies recommended to them by strangers who are randomly assigned an indicator (name) of own- versus out-group status. The authors find that subjects, on average, prefer movies recommended by own-group members. Consistent with the authors' theory, this own-group bias is especially pronounced when choosing in the randomly assigned "rushed" condition.

The authors then take the data from the lab experiments – the data from subject's responses – and use that as a training dataset to build a recommender algorithm. They find that this type of algorithm exhibits more out-group bias in rank-ordering movie reviews when trained using data from the lab experiment's rushed condition than the non-rushed condition. In fact, the algorithm results in even more detectable bias than in the subject responses themselves.

Test #2: Facebook algorithms

To understand the potential real-world implications of these findings, the authors audited two Facebook algorithms: News Feed, which ranks the posts of a user's friends, and People You May Know (PYMK), which ranks potential new friends.

In the first case, when subjects are asked to report on their desire to view a post in the News Feed, their responses are positively correlated with News Feed rankings. However, the authors also find a statistically significant (and sizable) difference in rankings for own- versus out-group posts as defined by race, even conditional on user preferences (e.g., for a white Facebook user, Facebook downranks posts from Black friends).

The authors then collected data on user preferences and the algorithm's ranking of candidate friend recommendations from PYMK friend recommendations to find that there is no detectable out-group bias in these rankings. What explains this difference from the News Feed findings? The authors employ a variety of metrics to show that users report more automatic

behavior when scrolling through posts (the data used to train News Feed) than when scrolling through potential friends (PYMK data).

Similar results hold in the single largest Facebook market in the world, India, where the context for own- and out-group bias is not race but, rather, religion (Hindus vs. Muslims). News Feed rankings are biased against posts by Hindu friends of Muslim users, and biased against posts by Muslim friends of Hindu users, with no detectable evidence of bias with the PYMK algorithm.

Bottom line: Consistent with their theory, the authors find that while the News Feed rankings (derived from relatively more automatic behavioral data) show signs of out-group bias, they find no detectable disparity in the recommendations of the PYMK algorithm (built with less automatic behavioral data). These findings make clear that the design of human-facing algorithms must be as attentive to the psychology and behavioral economics of human users as to the statistical architecture and machine learning engineering of the algorithms. Put another way, more attention must be paid to what behaviors are included in the training data used to construct algorithms, especially in online contexts where so much measured behavior is likely automatic.

READ THE WORKING PAPER

NO. 2023-19 · FEBRUARY 2023

Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias

bfi.uchicago.edu/working-paper/2023-19

ABOUT OUR SCHOLARS

Jens Ludwig
Edwin A. and Betty L. Bergman Distinguished Service Professor, Pritzker Director of UChicago's Crime Lab, Codirector of the Education Lab, Harris School of Public Policy
harris.uchicago.edu/directory/jens-ludwig

THE UNIVERSITY OF CHICAGO
HARRIS SCHOOL OF PUBLIC POLICY

Sendhil Mullainathan
Roman Family University Professor of Computation and Behavioral Science, Chicago Booth
chicagobooth.edu/faculty/directory/m/sendhil-mullainathan

CHICAGO BOOTH
The University of Chicago Booth School of Business