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

Amanda Agan Diag Davenport Jens Ludwig
Sendhil Mullainathan

January 29, 2023

**Abstract**

Consumer choices are increasingly mediated by algorithms, which use data on those past choices to infer consumer preferences and then curate future choice sets. Behavioral economics suggests one reason these algorithms so often fail: choices can systematically deviate from preferences. For example, research shows that prejudice can arise not just from preferences and beliefs, but also from the context in which people choose. When people behave automatically, biases creep in; snap decisions are typically more prejudiced than slow, deliberate ones, and can lead to behaviors that users themselves do not consciously want or intend. As a result, algorithms trained on automatic behaviors can misunderstand the prejudice of users: the more automatic the behavior, the greater the error. We empirically test these ideas in a lab experiment, and find that more automatic behavior does indeed seem to lead to more biased algorithms. We then explore the large-scale consequences of this idea by carrying out algorithmic audits of Facebook in its two biggest markets, the US and India, focusing on two algorithms that differ in how users engage with them: News Feed (people interact with friends' posts fairly automatically) and People You May Know (people choose friends fairly deliberately). We 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), but no detectable bias in the PYMK algorithm. Together, these results suggest a need to rethink how large-scale algorithms use data on human behavior, especially in online contexts where so much of the measured behavior might be quite automatic.

*For helpful comments we thank Hunt Alcott, Amanda Coston, Josh Dean, Jonathan Guryan, Reid Hastie, Sam Hirshman, Alex Imas, Erika Kirgios, Jon Kleinberg, Alex Koch, Betsy Levy Paluck, Emma Pierson, Devin Pope, Emma Rackstraw, Manish Raghavan, Ashesh Rambachan, Evan Rose, Jon Roth, Avner Strulov Shlain, Cass Sunstein, Richard Thaler, Alex Todorov, Stefan Uddenberg, Oleg Urminsky, Bernd Wittenbrink, George Wu, attendees at the annual meetings of the American Economic Association and the Society for Judgment and Decision Making, and seminar participants at the University of Chicago and the AI for Behavior Change workshop at the 2022 meetings of the Association for the Advancement of Artificial Intelligence. We are particularly grateful to Center for Applied Artificial Intelligence at the University of Chicago Booth School of Business for funding. For phenomenal programming assistance we thank Khoa Nguyen and Mani Yatam. For research support we thank Becky White, Bryan Baird, Amy Boonstra, and the team of RAs at CDR, Tirtha Patel, Pavan Mamidi, the team of RAs at CSBC at Ashoka University, and the Harvard Decision Science Lab. Agan: Rutgers & NBER, Davenport: Princeton, Ludwig & Mullainathan: U. Chicago & NBER.

# 1 Introduction

Consumers' choices are no longer entirely of their own making. Algorithms now curate choice sets, rank those choices and make recommendations (e.g. movies, news stories, social media posts, music, books, websites, etc.). Though the technical details of these algorithms vary dramatically, the general structure is to infer user preferences by relying on data about people's past choices and behaviors.[1] So these algorithms largely share an (often implicit) assumption: what people choose is what they want. These algorithms thereby ignore a key lesson from behavioral economics: what people choose is not necessarily what they want.

We explore the implications here of an important type of deviation between preferences and choices: a foundational insight from behavioral science is that people can sometimes behave automatically. When thinking fast (automatically), cognitive biases creep in that do not appear when thinking slow (deliberately) and thus may not reflect actual objectives. As such our quick ("system 1") decisions can deviate from the preferences we would display upon more careful consideration ("system 2"). In the discussion that follows we will for convenience refer to system 2 decisions as "preferences," but we recognize that fundamentally there is no way to definitively know someone's "true" preferences.[2] As a result, algorithms trained on our behavior run the risk of automating our automatic tendencies rather than catering to our more deliberate, considered preferences.

We study this general problem in the context of one kind of automatic bias of particular social concern: discrimination. In broad strokes, automaticity itself can induce prejudice above and beyond any explicit preference.[3] Many of the forces that create discrimination operate quickly—stereotypes, gut responses—and can therefore exert stronger influence over automatic choices.[4] As a result, biased behavior is actually a combination of two distinct forces: (i) prejudice that would arise even if people made slow, deliberate choices and (ii) the *additional* prejudice that can arise from how automatically people chose.[5] The additional

[1] See Section 6.1 for a discussion of TikTok's algorithm as an example.

[2] Kleinberg et al. (2022) draw out the theoretical consequences of such misunderstandings in the case of self-control problems, where automatic choices can create unwanted impulsive choices. The present paper considers the implications of automaticity for algorithmic bias, and in addition to presenting a conceptual framework, in this paper we also take these ideas to data.

[3] Notice we allow for the fact that deliberate choices can also contain prejudice. Our focus is on the *additional* gaps that arise from behaving automatically. In practice, it is possible that stated preferences are not our actual preferences, but merely the ones we state to placate a surveyor. We return to this possibility below and in Section 4.

[4] For example, Bordalo et al. 2016 formalize the idea of stereotypes as the result of the representativeness heuristic. Reliance on intuitive heuristics tends to be characteristic of "fast" or system 1 thinking.

[5] Chaiken and Trope (1999), Gawronski and Creighton (2013) and Kahneman (2011a) provide excellent

bias due to automaticity is distinctive because it is an artifact of the choice process rather than a reflection of actual preference. By inferring our preferences from our behavior, the algorithm is codifying biases users themselves do not consciously want. Importantly, the magnitude of that source of bias is not fixed: its extent is shaped by the context in which the behaviors or choices captured by the algorithm's training data are generated. A key prediction of this line of reasoning is that algorithms will inherit more bias when trained on more automatic behaviors.[6] These considerations are particularly troubling because casual inspection suggests algorithms are often trained in contexts (e.g. social media) in which people behave fairly automatically.

We formalize and illustrate these ideas for a particular kind of prejudice: our tendency to favor people like us ("own-group" members) and disfavor people not like us ("out-group" members). We then empirically test the implications of this model in two different ways that have complementary strengths and weaknesses and which together, we believe, tell a compelling story. Note that the key prediction we test is *not* that there will be algorithmic bias, but rather that the magnitude of such bias will be relatively larger for algorithms trained using behavioral data that are relatively more automatic. This is indeed exactly what we see in our first set of empirical tests, which involve laboratory experiments that have a high degree of internal validity. Our second set of tests have relatively higher external validity, by carrying out audits of one of the world's largest social media platforms: Facebook. Specifically, we study two Facebook algorithms that differ in the level of automaticity in the training data: News Feed, which relies on relatively more automatic behavior (scrolling friend posts), and People You May Know, which relies on less automatic behavior (deciding who to friend). While the comparison of bias between the News Feed and PYMK algorithms does not isolate the automaticity mechanism quite so cleanly, given other differences between the algorithms, and while we recognize that we have no way to know the exact inner workings of the Facebook algorithm, our findings are at least suggestive of the potentially immense practical relevance of our theory. Facebook alone, for instance, has 2.9 billion users per month all around the world, and is a source of social capital of the sort that affects a wide

reviews of the history and state of so-called "dual process theories" in social psychology. In appendix A, we provide a more detailed discussion on the relationship between group identity, prejudice, and thinking fast (vs. slow).

[6]The literature on algorithmic bias at this point is immense: see Chouldechova and Roth (2018); Mehrabi et al. (2021); Mitchell et al. (2021); Kearns and Roth (2019); Barocas et al. (2019) for overviews; Chen et al. (2020) discusses bias in recommender systems specifically. A few examples of empirical audit studies that resemble ours to a degree include Datta et al. (2015); Sweeney (2013); Kay et al. (2015); Klare et al. (2012); Buolamwini and Gebru (2018); Caliskan et al. (2017); closest to our study, Lambrecht and Tucker (2019) and Imana et al. (2021) study bias in ad targeting on Facebook (and LinkedIn).

range of life outcomes (Chetty et al. 2022).[7]

For our laboratory experiments we develop a task where subjects select movies recommended to them by strangers, who are randomly assigned an indicator of own- versus out-group status—name, as in Bertrand and Mullainathan (2004) and Kline et al. (2022). Subjects on average have a preference for movies recommended by own-group members. Consistent with our theory, this own-group bias is especially pronounced when choosing in the randomly assigned "rushed" condition.

We then use the subject responses in this lab experiment as the training dataset to build a recommender algorithm. We show 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.[8]

To understand the potential real-world implications of this finding, we carried out audits of two Facebook algorithms. The *News Feed* algorithm ranks the posts of a user's friends, which is the sort of high-frequency online behavior that tends to be quite automatic. For each post we ask subjects to report how much they wanted to see that post. We show that these self-reported preferences are indeed positively correlated with News Feed rankings. Yet we also see a statistically significant difference in rankings for own- versus out-group posts as defined by race, even conditional on user preferences (e.g., a white Facebook user has posts from their Black friends down-ranked). The difference is sizable: own-group posts in the bottom quartile of user preferences are on average ranked nearly as high as out-group posts in the second quartile of user preferences.[9]

[7] Putnam (2000) distinguishes between "bonding" social capital, or "strong ties" with friends and families that provide emotional and other supports, and "bridging" social capital, or "weak ties" (Granovetter 1973), that provide people with valuable new information and perspectives. The advent of social media has added a third category to this list, "maintained" social capital, or the ability to perpetuate ties to people with whom one has lost face-to-face contact (Ellison et al. 2007). Existing research suggests that use of Facebook at all relative to no use, relatively more intensive use of Facebook, and investments in time on "Facebook Relationship Maintenance Behavior" are all associated with increased social capital, particularly the weak ties associated with bridging social capital (Antheunis et al. 2015). Previous research has found that weak ties are positively related to important outcomes like creativity (Baer 2010), employment status and income (Tassier 2006), risk of crime involvement (Patacchini and Zenou 2008), health (Kawachi et al. 2000), and subjective well-being (Sandstrom and Dunn 2014). Previous studies also suggest that intergroup contact over social media, including on Facebook specifically, may reduce prejudice (Alvídrez et al. 2015; Schwab et al. 2019).

[8] While it might be possible to imagine some reason why people might be taking signal about how much they would like a movie from the recommender's name, if that were the case we would expect to see name matter not just under the rushed, automatic decision-making context, but also under the more deliberate decision-making context. But it does not.

[9] Of course one possible interpretation of these results is that user-reported preferences are subject to some sort of social desirability bias. As we discuss further below, one argument against this interpretation is that subjects did not know the study was about discrimination. But perhaps even more convincingly is the

We then turn to *People You May Know* (PYMK), which ranks potential new friends. We again collected data on user preferences, and on the algorithm's ranking of candidate friend recommendations. We show that there is no detectable out-group bias in these rankings. The difference between these findings and those from News Feed may result from the possibility that people are more deliberate in how they decide who to friend on the Facebook platform. And indeed we show using a variety of metrics that users report more automatic behavior when scrolling through posts (the data used to train News Feed) than when scrolling through potential friends (PYMK's data). To summarize, consistently with our theory, while the News Feed rankings (which are built off of relatively more automatic behavioral data) show signs of out-group bias, we find no detectable disparity in the recommendations of the PYMK algorithm (built with less automatic behavioral data).

We show that similar results hold in the single largest Facebook market in the world: India. The context here is quite different from that of the US, in that race is not the focus of so much discussion about out-group bias. The challenge stems rather from religious cleavages. So we now define out-groups based on religion: Hindus versus Muslims. Despite the change in geography and a focus on religion, we again find the same pattern of results. News Feed rankings are biased against posts by Hindu friends of Muslim users, and biased against posts by Muslim friends of Hindu users. And, again, we find no detectable evidence of bias with the PYMK algorithm.

Our results argue for much greater attention to the question of what behaviors go into the training data used to construct algorithms. In many decision contexts, several different kinds of behaviors are observed, such as whether to hover for an extra millisecond before scrolling by, or to "like" a post, or even to write a response. These behaviors can happen at different levels of mental deliberation, and so—our results suggest—can vary quite a bit as to their alignment with what users actually want. The implication is that the design of human-facing algorithms must be as attentive to the psychology and behavioral economics of the human users as to the architecture of the machine learning algorithms.

The findings reported here suggest a number of ideas that may stimulate further work in this area. First, algorithms rely on coarser representations of reality than consumers, since not all of the attributes that distinguish product choices out in the world are necessarily captured in the algorithm's training data. That can lead to something like stereotyping by the algorithm that winds up exacerbating the magnitude of bias. Second, this same feature

fact that we see a very similar pattern of findings for Facebook's People You May Know algorithm; there is no obvious reason that social desirability bias should be so substantially different for self-reported user preferences for News Feed posts versus PYMK friend recommendations.

of the algorithm to group data by observed covariates means that algorithms can show detectable bias in finite samples even in situations where no bias can be detected in the raw data on consumer choices directly. And third, bias in the algorithms that are used to curate (rank-order) people's choice sets can in turn lead to more biased choices compared to when people choose from randomly ranked recommendations. That is, because people scan lists from top to bottom, algorithms can exacerbate the problem of human bias in choices among seen content by increasing the chances that users select content recommended by own-group members. We provide some suggestive initial evidence of this sort of "double penalty" in both our laboratory experiment and News Feed audit.

## 2 Conceptual Framework

We build a simple formal framework to state more precisely our core assumptions and the implications we draw out. We focus on the problem of curation. Users face a large number of potential options (a set $S$): posts, tweets, products, movies, job applications, etc. Users must sift through this large set of options to find the ones they like. The fundamental challenge of curation is to help the user sift. In our model, we will assume the user goes through the items one by one; that is, the set is ranked and the user starts with the highest ranked item and proceeds downward. Each piece of content has features $x_s$ (for example, length, topic, etc.), and a binary feature $g_s$ for whether it was produced by an out-group member (1) or own-group member (0).

Engaging with $s$ produces a real valued utility $u(s)$, which is what we mean by user *preferences*. If utility were the only determinant of user choices and out-group posts generated less utility for users ($E[u(s)|g_s = 1] < E[u(s)|g_s = 0]$), then an algorithm trained on user choices would lead to ranking of content $r^u(s, S)$ (with highest-utility content ranked 1 and so on) that would simply reflect user preferences.

The fundamental challenge arises from the fact that the data typically used to train algorithms includes information not about our preferences, but instead about our choices—that is, not $u(s)$ but rather whether we engage with (e.g. "click") a piece of content, $c(s)$. A large body of research from psychology demonstrates that our choices can often deviate from our preferences ("intention-action gaps"), especially in decision settings that are more automatic or "fast" ($f = 1$) where our cognition is relatively more automatic than in settings where our choices are relatively more deliberate or slow ($f = 0$).[10]

[10]See, for example, Kahneman (2011b) and Kahneman et al. (2021). Speed is not the only decision setting that may induce system 1 type thinking, more generally any situation that limits available cognitive

Let the decision to engage with content obey:

$$Pr(c(s)=1) = logit^{-1}(u(s)(1-b_f g_s)) = \frac{e^{u(s)(1-b_f g_s)}}{1+e^{e^{u(s)(1-b_f g_s)}}}. \quad (1)$$

The psychology research implies $b_1 > b_0 \geq 0$ so that fast contexts ($f=1$) have greater bias than slow contexts ($f=0$).[11]

Since utility is unobserved, the algorithm ranks by choices instead, which we call $r^{c,f}(s,S)$. This is sorted by $c(s,f)$ with the most frequently clicked item ranked $r(s)=1$ and so on. We define the *disparity* in a ranking rule $r$ as the expected difference in ranking for own-group versus out-group posts, $\Delta(r^u) = E[r(s)|g_s=0] - E[r(s)|g_s=1]$. It is easy to see that ranking by user behavior or engagement rather than utility increases favoritism for own-group content (because of intention-action gaps), and that these disparities are even larger when engagement is measured in fast contexts.

$$\Delta(r^u) < \Delta(r^{c,f=0}) < \Delta(r^{c,f=1}) \quad (2)$$

Algorithmic ranking increases disparity above and beyond the problem of using engagement as a proxy for preference. The algorithm ranks on *predicted* engagement, which is formed from a dataset of many pairs of the type $(c,x,g)$ in order to predict engagement for new posts for which we have just $(x,g)$ available. Let us (for now) generously assume the algorithm makes the best possible prediction given an infinite number of data points, so for any post the algorithm perfectly predicts $E[c(s,f)|x_s,g_s,f]$. So the algorithmic ranking $r^{a,f}(s,S)$ results from sorting the set of posts by $E[c(s,f)|x_s,g_s,f]$. Algorithms have more bias when trained on data from fast than slow contexts:

$$\Delta(r^{a,f=0}) < \Delta(r^{a,f=1}). \quad (3)$$

Of course the algorithm does not know the actual click rate for every post $c(s)$. Instead it must use the *expected* click rate; that is, the average click rate of similar posts, those with the same $(x,g)$. By ranking on predicted engagement the algorithm replicates the problems of stereotyping.

One problem is that algorithms can diffuse bias. The algorithm is lumping out-group posts in together with other out-group posts. If there is bias against some out-group posts, all

bandwidth can prompt more automatic decisions (e.g. stress) (Mullainathan and Shafir 2013)

[11] See for example Payne et al. (2002), Lueke and Gibson (2015) and the studies reviewed there.

out-group posts can be penalized. And because the algorithm pools data across users, it can propogate implicit biases across people as well. If any users are biased, then:

$$\Delta(r^u) < \Delta(r^{c,f}) < \Delta(r^{a,f}). \tag{4}$$

A second problem is that algorithms do not simply replicate human bias in engagement, but can also *magnify* the bias. The probability that a given own-group post will have a higher click rate than a given out-group post, $P(c(s)|g_s = 1) > P(c(s)|g_s = 0)$ is smaller than the probability that a given own-group post's algorithmically predicted click rate is higher than an out-group post, or $P(E[c(s)|g_s = 1, x_s] > E[c(s)|g_s = 0, x_s])$. To see this assume a simple linear functional form with $c(s) = \alpha + \delta g_s + x_s + z_s$ with $z$ equal to a noise term. The variance of the own- versus out-group difference in clicks equals $\delta^2 + \sigma_x^2 + \sigma_z^2$. But by averaging clicks within cells defined by observable covariates, the own- versus out-group difference in the algorithm's predictions has a smaller variance, equal to $\delta^2 + \sigma_x^2$.

This averaging feature of the algorithm's predictions also makes own- versus out-group differences more easily detectable by improving statistical power. By grouping the data by cells defined by the observable covariates, the algorithm's predictions throw away all of the within-cell variation due to unobserved variables. This makes the between-group comparisons more statistically precise when the outcome of interest is algorithmic predictions rather than actual human choices. We can see this in Table 1 where we present the results of a simple simulation exercise. Assume that $\alpha = 0.5$, $\delta = 0.01$, $x \sim N(0, 0.01)$ and $z \sim N(0, 0.05)$. The rows show the results from simulating 1,000 samples (equally split between own- and out-group) of size 50, 100, 500 and 1,000. We can see that the average p-value for the test of the null hypothesis that the group mean click values are the same is consistently lower when we use simulated "raw" click data (first column) than when we use the algorithm's *predictions* of the click data (second column).[12]

The remainder of our paper tries to empirically test these hypotheses: (1) that automatic behavior generates greater algorithmic bias than does more deliberate behavior; and (2) between-group gaps in algorithmic rankings can be sizable even when between-group differences in click propensities are modest. We also provide some initial suggestive evidence that algorithmic curation of our choices can create a "double penalty" against out-group content.

Testing requires a concrete definition of own-group / out-group boundaries, which have long

[12]The "algorithm" in this case is a simple linear probability model that predicts clicks using group membership and the observable covariate $x$.

been recognized to be context-dependent.[13] In the laboratory experiment task that we describe next, we ask subjects to select movies recommended to them by strangers. Because this task is about product choice (as opposed to social interaction, as with Facebook), and because movie preferences vary substantially by both race and gender,[14] we use both features to define groups in the lab. Since Facebook is a vehicle for social interactions, for our US Facebook audit we define groups based on a very salient characteristic in the US: race.[15] For India, we focus on a feature very salient in that setting: religion (Hindus versus Muslims).

# 3 Lab Experiments

To empirically test our theory, we carried out a series of carefully controlled laboratory experiments, which mimicked the key features of online settings where choices are algorithmically curated for users. These experiments have the advantage of precisely isolating and testing the implications of our model, although of course at the cost of substituting behavior in the lab setting for behavior in the real world. We carried out two lab experiments that show how more automatic decision-making contexts exacerbate the problem of bias, which in turn introduces more bias to algorithms built using data from quicker, more automatic decisions.

## 3.1 Experimental Design

Our lab experiments mimicked the task of an algorithm that must rank content for a user based on past decisions (of the same user and also other users). The strength of these lab experiments is the ability to randomize subjects to both own-group vs. out-group content, and to rushed versus non-rushed decision-making contexts. Additionally, because we built the choice-curating algorithm ourselves using the responses from our own lab study subjects as the training data, we know exactly how the algorithms are constructed (unlike with

[13]The idea that the group identities that are most salient to people vary from context to context has been recognized dating back at least to Asch (1951) and Sherif (1988). For example, for two Americans meeting in Paris, the fact that they are both foreigners in France from the same country may be the most salient group definition. For those same two people meeting on the streets of Chicago, other personal characteristics—gender, race, political affiliation, or favoring the Cubs versus the White Sox—may be more salient.

[14]Survey data show overlap but also notable differences in movie preferences between Black and white adults in the US; see https://morningconsult.com/2020/07/08/black-audiences-polling-hollywood-diversity/. Similarly, the data reveal substantial differences by gender not just in actual movie preferences, but perceptions of what movies people of the opposite gender will like (Wühr et al. 2017).

[15]So, for example, if a white subject has a post from a white friend, a Black subject has a post from a Black friend, or an Asian subject has a post from an Asian friend, those would be coded as "own-group."

Facebook).

Our subjects looked at a series of movie recommendations, each of which was attached to a poster by name ("Amanda A. recommends..."), and chose amongst a subset of those movies to potentially watch. The movies and posters were shown three at a time on the screen, much as results from search algorithms or social media sites might show up, and the user could click "load more" to see more. The respondent could also click on a button to read a fuller review of a movie from the recommender. This is similar to many actual online environments in which choices are curated; users initially see a person attached to that content (Twitter handle/Twitter post, Facebook post/name, etc.), but they see limited information about the content, and are able to click to get a little more information before making a decision about whether to engage more fully with the content (e.g. choose to watch a movie). We selected 42 movies total from various genres. Reviews were actual reviews taken from the public dataset used in Maas et al. (2011).[16] The respondents were instructed to choose 4 movies, and told that once chosen they would get a link to watch one of their chosen movies. Since the recommendations showed up 3 at a time, the respondent needed to click on "load more" at least once to choose 4 movies.

We used an audit-study design that randomly assigned names to these real movie reviews, selecting names of the recommenders or posters to saliently signal race and gender. We drew the list of names from previous studies that have used a similar design to understand different economically meaningful decisions like hiring (Bertrand and Mullainathan 2004; Milkman et al. 2012; Agan and Starr 2018; Kline et al. 2022). The distribution of race among our fictional "posters" was designed to mimic the US population overall.[17] Randomization of names to movies means that we held constant true underlying preference for the movies themselves when we looked at how the demographics of the movie recommender affects the subject's movie choices.

We also randomized subjects to make decisions under one of two different choice contexts:

- A *rushed* condition in which the subject was told they would have 5 minutes to make their selections, and were told that this was not much time given the task at hand. To reinforce this, the clearly visible countdown clock on the left of the screen counted in

[16] Additional information about the movies, such as pictures to attach to the movies and ratings, were taken from IMDB.com

[17] In expectation, a user would see 64% white-signalling names, 22% Hispanic signalling names, and 14% Black-signalling names.

milliseconds so that the countdown moved quickly while the respondent decided.[18]

- A *non-rushed* condition in which the respondent was told they had 15 minutes to decide and that this was plenty of time, and the timer counted in minutes, moving much more slowly than in the "rushed" condition.

Screenshots with instructions and images of both the rushed and non-rushed conditions can be seen in Appendix Figure B.1.

We carried out two versions of the lab experiment:

- *Experiment 1* involved $N = 981$ study subjects recruited through the Prolific platform, with a sample intended to be representative of the US population with respect to race.
- *Experiment 2* involved $N = 753$ study subjects from Prolific who were white males, a design modification intended to make the job of defining own-group versus out-group posts (one of the key features of our experiment) easier.

Appendix table E.1 shows summary statistics for our two lab study samples. Experiment 1's sample is roughly representative of the country as a whole with respect to race and ethnicity (63% white, 23% Hispanic, 14% Black), slightly over-represents females (61%), is more highly educated than the population overall (32% have a Bachelor's degree and 31% have more than a Bachelor's degree), and has an average age of 29.6. The study sample for Experiment 2 is older (38.8 years of age) with lower levels of schooling. We have a total of $N = 4,859$ movie selections from Experiment 1 and $N = 3,001$ from Experiment 2.

## 3.2 Results

The results for study subject choices are shown in the two left-hand panels of Figure 1. In the deliberate decision-making condition in Experiment 1, subjects were 1.4 percentage points more likely to choose movies recommended by an own-group member, which represents a 7.8% increase over the mean click-rate for out-group recommended movies of 17.8 ($p < 0.10$). But when rushed, subjects were 3.1 percentage points more likely to choose a movie recommended by an own-group poster ($p < 0.01$; an 18% increase over the mean click-rate for out-group recommended movies).

[18] This seemed to be salient to the participants. At the end they were asked to give us any feedback they had on the task and some quotes from people in the rushed condition include: "The timer was kind of scary to be honest. Reminded me of the stress I felt when playing through The Legend of Zelda: Majora's Mask."; "The timer was too fast to be able to read any of the recommendations. It provided some anxiety to choose the movies."; "I'm curious about the large timer counting down the whole time, seemed to add a stress component.".

One might wonder why subjects would pay attention to the name at all in deciding among movies. The work on automaticity suggests an explanation: when choosing quickly, many pieces of information are used, often including ones that we did not consciously choose to observe. Put differently, subjects did not need to *choose* to pay attention to the name; it simply seeped into their processing. That the name only mattered in the rushed decision-making condition reinforces the idea that attending to it was not a deliberate decision. This is consistent with our larger hypothesis about the role of automaticity and bias.

In Experiment 2, the own-group preference was equal to 0.8 pp in the deliberate condition (6% higher than the out-group mean click rate) and 0.5 pp (12%) in the rushed condition. Neither is statistically significant, but our standard errors here do not allow us to rule out own-group favoritism as large as 2.5 pp in the deliberate condition and 2.2 pp in the rushed condition. See Table 2 Columns (1) and (3) for these regressions.

## 3.3 Experimental Algorithms

We used the data from each of our two lab experiments to build two algorithms—one trained strictly on rushed condition data and the other trained strictly on deliberate condition data. We trained the two algorithms on 70% of the users from the rushed condition and 70% of the users from the deliberate condition, respectively. The hold-out/test set used to assess the two algorithms is the same——we combined the remaining 30% from the fast and slow experiments to construct the single hold-out set. This approach allows us to show how different algorithms would have ranked the options for a fixed set of users. The results we show below are based strictly on the hold-out set.

There are 10 inputs to the algorithm: genre (whether the movie is in the user's favorite genre or not); movie watching frequency from the user; the movie's rating; the movie recommender's gender and race; and the study subject's age, gender, race, and educational attainment (for more details see Appendix B.2). We use these features to predict user movie selections by estimating a random forest, a widely used classification algorithm (Breiman 2001). Two hyperparameters for the random forest (number of trees to grow and the number of variables to sample as candidates at each split) were tuned with cross validation; R defaults are used for all other hyperparameters. The output of this algorithm is a rank-ordering of movie choices by the predicted likelihood of user selection.

This is basically how actual curation algorithms work; the main difference is that our dataset is much smaller than those used with commercial algorithms (fewer observations and fewer covariates). While industrial-strength machine learning algorithms might not explicitly use

measures of user demographics such as race, it is widely accepted that the rich set of covariates often available in modern-day "big data" make it easy for algorithms to reconstruct close proxies for race.[19]

The top right panel of Figure 1 shows how the algorithm sorts movie recommendations for subjects in Experiment 1, when trained separately on choices from the rushed and non-rushed conditions. The vertical axis depicts where each movie gets ranked in terms of number of slots relative to the average movie ranking. Own-group posts are ranked more towards the top when the algorithm is trained on data from the rushed condition, equal to 2.7 ranking slots above those from out-group posts ($p < .05$), while from the non-rushed condition own-group recommendations only move about 1 ranking slots above out-group posts ($p < 0.05$).

For Experiment 2, even though the own-group favoritism was not statistically significant on average for the rushed or deliberate conditions, when we fed these choices into our algorithm to sort movie recommendations, we saw very similar results to Experiment 1. There was up-ranking of own-group posts 3.3 ranking slots above out-group in the rushed condition ($p < .05$) and 1.9 ranking slots in the deliberate condition ($p < 0.05$). See Table 2 Column (2) and (4) for regression versions of these figures.

These results from Experiment 2 speak to the second hypothesis raised by our conceptual framework: can we detect out-group bias in the algorithm even when we cannot in the algorithm's human-generated training data? The answer, as shown in Figure 1 and Table 2, is yes. As noted above, this follows partly from the fact that the standard error around our estimate of out-group bias is partly a function of the residual (unexplained) variation in the outcome variable being examined. Algorithms, by grouping together all the choices people make by different characteristics of the people or the content, average away some of that residual variation and enable more precise estimates of out-group bias.[20] The algorithm does not only promote the detection of between-group differences, it also increases the magnitude of the bias by working in units of rank-ordering of click rates, rather than click rates.

[19] This view holds not only among data scientists but among the general public as well. In surveys, 4 out of 5 American adults say they think it is easy for social media sites to figure out their race and ethnicity. https://www.pewresearch.org/internet/2019/01/16/facebook-algorithms-and-personal-data/

[20] Another way to see this is to consider situations where the explainable variation in user choices is modest relative to the total variation. In those cases, even if the influence of own-group / out-group status is modest, that influence becomes magnified by the algorithm because the own-group effect is necessarily a larger share of the predictable variation than total variation in user choices.

# 4 US Facebook Audit

Our laboratory experiments are perfectly designed to isolate the effects of automaticity in people's choice context on the level of algorithmic bias, but they do so at the cost of relying on hypothetical behavior in the lab setting. To better understand the potential real-world implications of our theory, we also carried out audit studies of two Facebook algorithms, News Feed and People You May Know (PYMK), in two countries, the US and India. We show that users report spending more time deliberating before making choices on PYMK compared to Facebook, which makes sense given the relatively higher stakes of choosing to friend someone on Facebook versus clicking on a post on News Feed. Our model predicts more pronounced algorithmic bias with News Feed than with PYMK. That is exactly what we find in both countries. Moreover, the magnitude of News Feed's bias relative to PMYK's is substantial.

## 4.1 Automaticity Across Facebook Algorithms

A key prediction of our model is the idea that behavioral biases—and hence algorithmic biases—are most pronounced when behavior is not guided by deliberate thought. We hypothesized that Facebook may provide a sort of "natural experiment" relevant to our model, since the News Feed algorithm curates a large number of low-stakes choices for users (how to rank-order posts from friends) while the PYMK algorithm curates a smaller number of higher-stakes choices (whether to "friend" another user on the platform). The choices that the News Feed algorithm relies on for training data are therefore plausibly more automatic compared to those relied on by the PYMK algorithm.

We first carried out a survey in the US (on the Prolific platform) to measure the amount of cognitive effort, deliberation, and time spent making choices to interact with posts on News Feed versus to add a friend from PYMK (N=300). We draw on existing measures in the literature about, for instance, how well the subject could explain their choices, how much "mental effort" they say they put into the behavior, whether the decisions are based on "gut feelings" or careful consideration, and how much time they usually spend (in seconds) making the decision (Bargh 1994). For more details on our specific measures, see Appendix C.

Figure 3a shows that for each of these nine measures of deliberate interaction, users report higher levels of deliberation when using PYMK than when using News Feed. The difference ranges from 0.05 standard deviations (for inattention) to 0.53 standard deviations (for carefulness). A simple composite index of the standardized measures suggests PMYK has a "deliberation advantage" of 0.22 standard deviations over News Feed. When we do

a principal components analysis across the measures and compare the first principal component across algorithms, PMYK's deliberation advantage equals 0.4 standard deviations. And Figure 3b shows the CDF for responses to the one continuous measure, time it takes the study subject to make a decision (in seconds), which again shows the same pattern. A Kolmogorov-Smirnov test rejects the null that these CDFs are the same with $p < 0.01$.

## 4.2 US Facebook Audit Design

We advertised for study subjects who were Facebook users and were willing to participate in a Zoom-based interview. Details on our sample recruitment protocols can be found in Appendix D. Subjects were first asked to complete a survey asking about their demographic characteristics and basic Facebook usage patterns. Subjects were then asked to log into their Facebook account and share their screen. Data was collected in several waves, and different subjects had different data collected depending on the wave they participated in:

- For all subjects, enumerators captured the News Feed algorithmic ranking, and information about each of the first 60 posts in the user's News Feed (N=662).
- One subset of News Feed subjects then participated in a similar data collection about the first 60 friend recommendations from the PYMK algorithm (N=436)
- A different subset participated in data collection about their 10 most recent *interactions* with News Feed posts on Facebook (N=104)
- No further data was collected from the third subset (N=122)

Which data collection wave a subject participated in was determined by the date and location of data collection—a complete description of this wave structure is in Appendix D. In our main analysis, we use the full samples available for each result—that is, we present results for News Feed for all observations for which we collected News Feed data, results for PYMK in the full sample for which we collected PYMK data, and results for interactions with Facebook posts in the full sample for which we collected interaction data.

To avoid priming study subjects about the topic of race (the way we define own-group / outgroup for this audit), and because of time constraints on our data collection with each subject, we did not ask subjects themselves to report the race/ethnicity of posters on their News Feed. Instead, we asked our enumerators, who could see the subject's Facebook account through screen sharing, to record their perception of the race/ethnicity of the posts from people (not companies) in each of the first 60 total posts and first 60 friend recommendations.

We asked subjects to self-report their own race and ethnicity using the seven-category system from the U.S. Census, where subjects can check as many boxes as they like. We also asked the enumerators to record their perception of the race/ethnicity of the subject before the Facebook data collection began, which matched subject self-reports 85% of the time.[21]

We also collected direct measures of people's explicit utility or preferences:

- For News Feed, we asked subjects to report for each of the posts from people amongst the first 60 total posts they see: "There are more posts than Facebook can possibly show you. How would you rate this post on a scale from 1-7 where 1 means 'can skip' and 7 means 'definitely want to [see]?' "
- For those who participated in the PYMK data collection, for the first 60 friend recommendations on PYMK they are asked: "How familiar are you with this person on a scale from 1-7?" where 1 is not familiar and 7 is very familiar.

The enumerators also recorded ancillary information about each News Feed post, such the post's timestamp, whether the post was made by a person versus company, and whether the post was in a Facebook group. For PYMK recommendations, enumerators recorded additional information such as the number of friends that each friend recommendation shared with the subject.

As noted above, some subjects were also asked about their recent behavior on Facebook (N= 104). Specifically, we recorded the 10 most recent posts on News Feed that the user had some *interaction* with ("liking" or choosing another reaction or commenting), and exactly what action they took. Enumerators then also recorded the perceived race/ethnicity of the poster.[22]

Table 3 shows summary statistics on US subjects for whom we collected News Feed information (N=662), subjects for whom we collected PYMK information (N=436), and subjects for whom we collected recent News Feed interactions (N=104). Our subjects are on average 26.6 years old (sd= 9.186, range=18 to 69). A large fraction of our sample check Facebook at least weekly. Compared to all U.S. adults, our samples tend to have a higher proportion Asian,

[21] This is under a rather strict definition of match: enumerator and subject race choices had to match exactly. However, subjects often chose more than one race but enumerators rarely did so. When we expand the match to be that the enumerator indicated the same race as at least one of the responses of the subject, then the match rate is even higher.

[22] The algorithm presumably has access to a wider range of behaviors than this, such as how long the user lingered on a post, whether the user clicked to expand the post text or comments, whether the user watched a video and how much of the video was watched, etc. Given the constraints on our data collection, interacting with posts was the most feasible measure of actual user behavior on the network.

female, and people with a four-year (bachelor's) college degree. In Appendix Table F.3, we show that re-weighting the data to demographically match the rest of the US with respect to gender, race, age, and education yields results that are qualitatively and quantitatively similar to those reported below.

Note that subjects were *not* made aware that this study was about race or own-group biases, and what data was exactly being recorded by the enumerator was unknown to the subject. For 50 subjects, at the end of the survey we asked, "What do you think the purpose of this study is?" Not one mentioned race, gender, or other indications of own-group biases.[23]

## 4.3 US Facebook Audit Study Results

Our model suggests that the behavioral wedge (relative to preferences) in the direction of own-group bias should be larger for the News Feed algorithm (where decision-making is more rushed and automatic) than for PYMK (where decision-making seems to be more deliberate). This is indeed what we find in our audit study of Facebook users in the US.

Figure 4 shows these results for both News Feed and PYMK. We first normalize people's explicit preference ratings for News Feed and PYMK recommendations to account for the fact that different people use the Likert scale differently, and in particular we see systematic differences in Likert scale distributions for white and Black study subjects.[24] An own-group post in the bottom quartile of the user preference distribution has a *higher* News Feed ranking than an out-group post in the *next-highest* preference quartile. In a regression, own-group posts are ranked by News Feed 1.19 slots higher than out-group posts even conditional on user preference, a difference that is statistically significant at the 5% level (with a standard error of 0.208); see Table 4. In contrast, we see no detectable differences in PYMK rankings conditional on user preferences. Our 95% confidence intervals let us rule out an own-group effect on PYMK rankings larger than 0.187 slots (Table 4).

Our results do not seem to be an artifact of our particular estimation choices. As shown in Appendix Figure F.3, we see similar results if we look at the probability that a post is ranked in the top 5, 10 or 20 slots. Our results are not sensitive to using the actual (non-normalized) Likert scale reports of user explicit preferences over content instead, as shown in Appendix

[23] Similarly, enumerators were not told the purpose of the study, though they were of course aware they were collecting information on race.

[24] As shown in Appendix Figure F.1, Black subjects tend to be more likely to use higher Likert preference rankings, indicating that they more prefer a post. Black subjects also see fewer posts by own-group friends on average (48% versus 51% for non-Black subjects). The normalized preferences re-scale scores relative to each subject's own average reported preference, and so take into account differences across subjects in the use of the Likert scale.

Figure F.1.

One may worry this is about family members. If Facebook is aware of which existing friends are family members, and family members are highly likely to be "own-group," then the upranking of these posts could simply reflect family status. In Wave 5 of our data collection, we gathered information on how participants knew the poster, and in Appendix Table F.2, we use these data to interact own-group status and preference with a binary indicator for family. We see that even amongst non-family posts, Facebook upranks own-group posts even conditional on user explicit preferences.

Our results also would not seem to be due to social desirability bias, which might lead study subjects to misreport their preferences about Facebook content. We blinded study subjects to the purpose of the study. Moreover, if subject reports were prone to social desirability bias, and that were leading us to under-state the bias of the PYMK algorithm, it is unclear why subjects would not be prone to similar social desirability bias in their preference reports for News Feed, where we do detect bias.

Relatedly (and interestingly), even though News Feed rankings seem to have an own-group bias, our measures of people's explicit preferences about News Feed content does not. Table 5 shows that for normalized self-reported user preferences, the quartile rankings of own-group and out-group posts are very similar; none of the differences are statistically significant at the 5% level. In Appendix Figure F.2, we also show a CDF of the normalized preferences people report for own-group and out-group content, which correspond heavily. We might worry about social desirability bias in subject responses, but subjects did not know the study was about race. Moreover, intentionally misreporting and hiding any explicitly biased preferences would require subjects to keep a running tally of their reported preferences across all the top 60 News Feed posts separately by own-group and out-group.[25] The most straightforward explanation for these results is that algorithms that learn preferences from quick decisions can become biased even when learning from users with no explicitly biased preferences.

# 5 India Facebook Audit

Part of what makes Facebook an interesting setting in which to explore our hypothesis is its massive scope. Billions of people around the world regularly use Facebook and rely on its algorithms. So far, we have presented evidence for the world's second-largest Facebook

[25] As noted above, we asked 50 participants what they thought the study was about, and not one mentioned "race," "bias," "discrimination," "ingroup" or any similar phrases.

market, the US. But the US, with its particular cleavages by race and ethnicity, accounts for just 10% of all Facebook users world-wide.[26] Do our results hold more generally?

To answer this, we replicated our entire audit study in the world's largest Facebook market: India. Rather than define own-group and out-group by race, signaled by user images in the US Facebook context, we define this now by religion, signaled by name in the Indian Facebook context.[27] Our results are qualitatively similar in India.

## 5.1 India Facebook Audit Study Design

Other than defining own-group/out-group status by religion, the study design of our India Facebook audit was identical to our audit in the US.

We recruited $N = 200$ study subjects via the Ashoka University Centre for Social and Behaviour Change (CSBC). Appendix Table G.1 shows that our sample somewhat over-represents men (60.7% or our study sample) as well as Muslims (30.1% of our sample, compared to 14.2% of the general population). In Appendix Table G.4 we show the results we present below are qualitatively similar when we re-weight our analysis to make our weighted sample more nationally representative by gender and religion.

## 5.2 India Facebook Audit Study Results

In Figure 6 we show that, as in the US data, there is a sizable difference in News Feed post rankings for own-group content (defined by religion) relative to out-group content, even conditional on explicit user preferences. For example, own-group posts in the bottom quartile (least preferred) of user preferences have an average News Feed ranking that is not substantially different from the average News Feed ranking for out-group content in the third (next-to-most-preferred) quartile. In Table G.3, own-group posts are ranked 1.110 slots closer to the top than are out-group posts (standard error 0.398, $p < .05$).

In contrast (and again consistent with the findings from the US data), we see little difference in the algorithm's rankings for own-group versus out-group recommendations with PYMK. Given our smaller sample size for India relative to the US, our PYMK results are somewhat noisier (our 95% confidence interval does not let us rule out an own-group versus out-group

[26] https://worldpopulationreview.com/country-rankings/facebook-users-by-country

[27] Names are quite distinct in India for Hindus (roughly 80% of the country's population) versus Muslims (14% of the population). Like race in the US, religion is a fraught fault line in Indian society. Hundreds of thousands of people died when Muslim and Hindu populations were partitioned in 1946 into the countries of Pakistan and India, respectively. Discrimination and even violence on the basis of religion remains common in India to this day.

difference in rankings less than 0.839 slots), but the point estimate is quite small, about one-fifth of a ranking slot. (Additional results for the India sample are in Appendix G.)

As with the US data, even though we see a difference in News Feed rankings of own-group versus out-group content in India, we do not see much detectable bias in explicit user preferences. Table G.2 shows that the average quartile rankings of subject explicit preferences for News Feed posts are very similar for own-group versus out-group content. Only one of the pairwise differences is statistically significant (second quartile).

The tendency of algorithms to learn our preferences from our worst selves does not appear to be specific to any particular type of bias (race versus religion versus other salient own-group vs. out-group distinctions), or to any particular setting or country context.

# 6 Implications for Algorithms and Platforms

We next consider how broadly these findings might apply, and what sorts of constructive responses might be possible.

## 6.1 Platforms and Behaviors

Survey data suggest our results are likely to apply beyond our controlled lab setting or Facebook. We surveyed a sample of 576 adults about their behavior on different broad categories of online applications they might engage with: social media (like Instagram, TikTok or Twitter), content consumption (Spotify, Netflix), and commerce (Amazon). For different types of user engagement with each type of platform, respondents were asked how much each of these behaviors is based on careful consideration ("carefully consider"), and how much time the user spends thinking about each behavior before deciding whether to take that action ("speed"). See Appendix H for more details on the survey and survey sample.

These questions get at the fundamental building block for our analysis: the automaticity of behavior in training data. The results, shown in Figure 7, illustrate several points. First, the different behaviors vary in their automaticity. Much as scrolling past posts was more automatic than deliberately choosing a friend on Facebook, many of these behaviors appear far more automatic than others. This is particularly easy to see with our measure of how much consideration the user gives to each behavior before they choose to carry it out. (Our measure of speed shows the same qualitative pattern but with much larger confidence intervals around each point estimate). This implies that the various algorithms used across these platforms, which rely on behaviors like this for their training data, may plausibly

differ in their automaticity as well. Second, because automaticity seems to vary so much across different behaviors *within* (not just *between*) different online platforms, our finding from Facebook may apply more broadly: some algorithms on these platforms may inherit the biases of automaticity while others may not. At a minimum, Figure 7 would seem to motivate additional in-depth audits akin to ours across these platforms.

A very different piece of evidence also suggests the importance of our findings. Neither in our audit study of Facebook nor in Figure 7 do we have explicit access to the underlying algorithm. But a recent *New York Times* report about the inner workings of TikTok's recommender algorithm provides insight into the exact mechanics of a specific, widely-used algorithm in another context. That report shows that videos are ranked according to a weighted average of several predictions: predicted user likes, predicted user comments, and predicted length of playing a TikTok.[28] Revealingly, the formula predicts behaviors that appear to span a range of automaticity, from letting a video play (relatively more automatic) to commenting on a video (relatively less automatic). That the recommender is essentially some simple weighted average of these predictions also suggests that there is no particular attention paid to the automaticity of these underlying behaviors.

## 6.2 Constructive Responses

So how *would* we fix this problem? Our psychologically informed perspective suggests two additional solutions beyond those already identified by the field:[29] (1) gathering data on explicit *preferences*; and (2) gathering data about the *automaticity* of different candidate behavioral measures.

First, in our audit studies, we diagnosed bias by contrasting rankings with self-reported user preferences directly (e.g., how much users liked posts). That approach can be applied more broadly. Many platforms already survey users about their preferences on a regular basis. As Milli et al. (2021) argue, such data can be used to build algorithms that better align recommendations with user value and not just measured engagement. In our case, we can build a solution that further relies on the particular psychology we have highlighted.

Using the notation from Section 2, suppose data on user preferences directly measure $u$, the utility from a post, and the platform measures $c$, the clicks or choices. Given data where we measure both $u$ and $c$, we can predict $u - c$ given $x$. That is, we can predict the types

[28] https://www.nytimes.com/2021/12/05/business/media/tiktok-algorithm.html

[29] While the literature in this area is vast, for examples see the excellent discussions (and additional papers cited therein) of Dwork et al. (2012); Zemel et al. (2013); Bellamy et al. (2019); Corbett-Davies et al. (2017); Chouldechova and Roth (2020);Raghavan et al. (2020); Vasileva (2020); and Wang et al. (2022).

of content where behavior systematically deviates from preference. Such predictions are in effect the more granular component parts underlying the predictions of what we show in Figure 4. If we refer to that as *predicted bias*, that gives us a prediction of bias for each type of content. Call the average bias predicted in this way bias($x$). We can then adjust rankings to account for this bias. So if the original ranking comes from predictions of $c$, the new bias-corrected ranking adds in bias($x$).

Importantly, this can be done when one has only a small sample of explicit user preference data. Our procedure only uses these data to predict which *kinds* of user-content combinations (defined by $x$) generate automaticity bias. Note also that this procedure does not require the platform to *ex ante* specify the relevant own-groups and out-groups. It can be estimated even if the dataset does not include explicit or direct measures of out-group status, since the algorithm could only express bias against out-groups if the $x$ variables available are able to at least partially reconstruct out-group status.

Our second proposed approach begins with the observation (as in Figure 7) that often platforms measure more than one behavior. TikTok, for example, measures commenting, viewing and liking. As another example, even in News Feed, some of the behaviors measured might include things like hover time over different posts, or whether the posts were read, or whether the user left a comment. A platform could, therefore, rely on well-developed survey measures of automaticity from the psychology literature (of the sort we have used in our own analysis here) to quantify how automatic different behavioral measures are. Specifically, they could collect data on each of the behaviors their algorithms use, as we did in Figure 3a. For example, they might find that time hovering over a particular post as they scroll through content may be more automatic than decisions about whether to share content that has just been read all the way through by the user.

If a platform finds that it is using behaviors that vary in their automaticity, that presents an opportunity. Suppose that a platform has two measures of behavior $c_1$ and $c_2$, where $c_1$ is more automatic than $c_2$. Specifically, following our earlier formalism, suppose that

$$Pr(c_j(s) = 1) = \frac{e^{u(s)(1-\beta_j g_s)}}{1 + e^{e^{u(s)(1-\beta_j g_s)}}},$$

where $\beta_j$ is the bias in behavior $j$ and $\beta_1 > \beta_2$. For a piece of content, when behavior differs systematically between behavioral measures $c_1$ and $c_2$, it suggests a problem due to automaticity. So we can introduce a new learning problem: predict, given $x$, $Pr(c_1 - c_2|x)$. Notice that this prediction is approximately $bias(x)$ but scaled by $\beta_2 - \beta_1$: the more similar

the behaviors are, the smaller the difference in bias. This predicted bias can now be effectively be rescaled by $\beta_2 - \beta_1$. Content can now be ranked by combining $c_1$, $c_2$ predictions with the prediction of bias. Notice this procedure can be considered a generalization of the procedure above: self-reported preference data are simply the case where we have at least one behavioral outcome where bias is known (assumed) to be zero. In many applications, there may be no behavior for which we are sure automaticity is not an issue at all. But so long as there is an "automaticity gradient" across behaviors, we may be able to use this gradient to extrapolate out to deliberate preferences.

These suggestions could surely be improved upon: they merely illustrate how knowing the structure of human bias can suggest the kind of data to collect, and how to use those data to build better algorithms.

# 7 Conclusion

We have provided evidence from two complementary sources—laboratory experiments and audits of one of the world's most widely-used social media platforms carried out in that product's two largest markets—consistent with the idea that curation algorithms (recommender systems) can learn and propogate biases that users themselves do not want. That is, we have shown that the extent of algorithmic bias in rankings is larger when the training data comes from contexts in which user choices are relatively more automatic. And we have shown that the bias in these algorithmic rankings can be pronounced even in situations where bias in actual human choices is hard to detect.

These findings are not only important in their own right but generative as well, suggesting additional new hypotheses that might be explored in future work. For example, biased algorithms may compound the user's own bias (or bias of users like her) by showing the user or prioritizing for the user a choice set that over-represents own-group items. After all, recommender systems are useful to consumers because the set of options is too vast for consumers to consider everything by themselves. So we would expect the magnitude of the bias in user choices to be larger when users choose from content ranked by biased algorithms than from randomly ranked choice sets. And we would expect biases in user behavior to be even larger than the bias we see in the algorithm rankings, given that users are choosing from an already-biased set of options. In other words, biased algorithms affect $P(seen|race)$, and biased humans affect $P(chosen|seen, race)$. For out-group content, the algorithm creates a "double penalty." We provide some initial, suggestive evidence that is consistent with this idea in both our laboratory experiment and our Facebook audits.

In the lab setting, we carried out a third experiment that started with the rank-orderings from an algorithm built using training data on user's "fast" choices. We then randomly assigned some study subjects to select content from a randomly-ranked set of options, while others were asked to select from algorithmically-ranked content. Since users typically scan lists from top to bottom, if users are willing to browse only a limited quantity of content, we would expect user choices to be more biased if made from a list that an algorithm has caused to be over-represented at the top with own-group content. That is in fact what we see in the lab data (additional details and results are in the appendix).

In the Facebook context we collected data on the 10 News Feed posts users most recently *interacted* with, either via reactions or comments. This is not a perfect measure of user behavior with News Feed, since there are other behavioral dimensions that we cannot measure in our setting (like the time the user had spent looking at each post, which we cannot capture in our lab setting). With this caveat in mind, when we ask study subjects to report their explicit preferences about News Feed posts on our surveys (using our Likert scale measures), among the "top 10" most preferred, 44.7% are own-group. Among the first 10 posts that News Feed showed to users, the share of own-group content is 50.5%—that is, algorithmic bias. Finally, when we look at the 10 most recent News Feed posts the user interacted with, fully 59.9% are own-group. By inadvertently learning our unconscious behaviors rather than our explicit preferences, the News feed algorithm (just like the algorithm in our lab experiment) seems to be creating a "double penalty" against out-group content, especially for choices made in rushed decision-making contexts.[30]

Taking all of our findings together, our study has two key implications: one practical, one conceptual.

The practical implication relates to the growing role of algorithms in mediating consumer choices, including with whom to interact online. In principle, these new technologies create new opportunities for us to connect within and across groups that are no longer limited by geography, which could reduce prejudice.[31] Yet despite the promise for these connections to reduce the prejudices in our world, our results suggest that in practice at least some of the

[30]These results are shown in Appendix Figure 5 for the sub-sample of respondents for which we have this behavioral measure. Given our use of a Likert scale that has fewer response options than the number of posts we ask users to rate, there will be many posts that share the same Likert scale value, and so for some subjects we cannot cleanly identify a "top 10" most preferred. Our results are not qualitatively sensitive to how we handle this problem, for example, by randomly selecting a top 10 from the top of the Likert scale distribution for users who have more than 10 posts grouped together with the same values. Each of the pairwise contrasts between the own-group versus out-group shares (user preferences versus News Feed rankings, and rankings versus user behavior) are statistically significant at the usual 5% threshold.)

[31]For a discussion of the literature on the inter-group contact hypothesis see Appendix A.

social network algorithms connecting billions of people seem to be actively reducing contact. This is especially troubling since these reductions occur even after people have made the effort to friend someone from another group.

Our conceptual contribution follows from the widespread recognition that human bias is typically responsible for most algorithmic bias. For example, many of today's large-scale human-facing algorithms are trained on data generated by people who are themselves biased. Yet despite the central role played by human bias, existing research on algorithmic bias has not fully capitalized on the large body of psychological research that exists.

Our analysis, we hope, demonstrates the value in doing so. A specific behavioral science insight has produced a richer understanding of the shape of algorithmic bias and possibly what to do about it. But behavioral science is a rich field that has many additional insights about people, beyond the one we focus on. Our over-arching hope is that this paper stimulates interest in "technology transfer" from the science of people to the science of algorithms.

# References


Abele, A. E., N. Ellemers, S. T. Fiske, A. Koch, and V. Yzerbyt (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review 128*(2), 290–314.

Agan, A. and S. Starr (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics 133*(1), 191–235.

Allport, G. W. (1954). *The nature of prejudice*. Reading, MA: Addison-Wesley.

Alvídrez, S., V. Piñeiro-Naval, M. Marcos-Ramos, and J. L. Rojas-Solís (2015). Intergroup contact in computer-mediated communication: The interplay of a stereotype-disconfirming behavior and a lasting group identity on reducing prejudiced perceptions. *Computers in Human Behavior 52*, 533–540.

Antheunis, M. L., M. M. P. Vanden Abeele, and S. Kanters (2015). The impact of Facebook use on micro-level social capital: A synthesis. *Societies 5*(2), 399–419.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership and men*, New York, NY, pp. 177–190. Carnegie Press.

Baer, M. (2010). The strength-of-weak-ties perspective on creativity: a comprehensive examination and extension. *Journal of applied psychology 95*(3), 592–601.

Bargh, J. A. (1994). The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In R. Wyer, Jr. and T. Srull (Eds.), *Handbook of social cognition: Basic processes; Applications*, pp. 1–140. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Barocas, S., M. Hardt, and A. Narayanan (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Bellamy, R. K., K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, et al. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development 63*(4/5), 4–1.

Bertrand, M. and S. Mullainathan (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review 94*(4), 991–1013.

Bobo, L. D., C. Z. Charles, M. Krysan, and A. D. Simmons (1972). The real record on racial attitudes. In P. V. Marsden (Ed.), *Social trends in American life: Findings from the General Social Survey since 1972*, pp. 38–83. Princeton, NJ: Princeton University Press.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016). Stereotypes. *The Quarterly Journal of Economics 131*(4), 1753–1794.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull and R. S. Wyer, Jr. (Eds.), *A dual process model of impression formation*, pp. 1–36. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues 55*(3), 429–444.

Buolamwini, J. and T. Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91.

Caliskan, A., J. J. Bryson, and A. Narayanan (2017). Semantics derived automatically from language corpora contain human-like biases. *Science 356*(6334), 183–186.

Caporael, L. R. (1997). The evolution of truly social cognition: The core configurations model. *Personality and Social Psychology Review 1*(4), 276–298.

Chaiken, S. and Y. Trope (1999). *Dual-process theories in social psychology*. Guilford Press.

Charles, K. K. and J. Guryan (2008). Prejudice and wages: an empirical assessment of Becker's *The Economics of Discrimination*. *Journal of political economy 116*(5), 773–809.

Chen, J., H. Dong, X. Wang, F. Feng, M. Wang, and X. He (2020). Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240*.

Chetty, R., M. O. Jackson, T. Kuchler, J. Stroebel, N. Hendren, R. B. Fluegge, S. Gong, F. Gonzalez, A. Grondin, M. Jacob, et al. (2022). Social capital I: measurement and associations with economic mobility. *Nature 608*(7921), 108–121.

Chouldechova, A. and A. Roth (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Chouldechova, A. and A. Roth (2020). A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM 63*(5), 82–89.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806.

Dai, W. and D. Albarracín (2022). It's time to do more research on the attitude–behavior relation: A commentary on implicit attitude measures. *Wiley Interdisciplinary Reviews: Cognitive Science 13*(4).

Datta, A., M. C. Tschantz, and A. Datta (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies 2015*(1), 92–112.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology 56*(1), 5–18.

Dovidio, J. F., K. Kawakami, C. Johnson, B. Johnson, and A. Howard (1997). On the nature of prejudice: Automatic and controlled processes. *Journal of experimental social psychology 33*(5), 510–540.

Dwork, C., T. P. Moritz Hardt, O. Reingold, and R. Zemel (2012). Fairness through awareness. *ITCS 2012 Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Eberhardt, J. L., P. G. Davies, V. J. Purdie-Vaughns, and S. L. Johnson (2006). Looking deathworthy: Perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychological science 17*(5), 383–386.

Eberhardt, J. L., P. A. Goff, V. J. Purdie, and P. G. Davies (2004). Seeing black: race, crime, and visual processing. *Journal of personality and social psychology 87*(6), 876–893.

Ellison, N. B., C. Steinfield, and C. Lampe (2007). The benefits of Facebook "friends:" social capital and college students' use of online social network sites. *Journal of computer-mediated communication 12*(4), 1143–1168.

Ferguson, M. J. and J. A. Bargh (2004). How social perception can automatically influence behavior. *Trends in cognitive sciences 8*(1), 33–39.

Fiske, S. T. and S. L. Neuberg (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology*, Volume 23, pp. 1–74. Elsevier.

Frank, M. G. and T. Gilovich (1988). The dark side of self- and social perception: black uniforms and aggression in professional sports. *Journal of personality and social psychology 54*(1), 74–85.

Gawronski, B. and L. A. Creighton (2013). Dual process theories. In D. E. Carlston (Ed.), *The Oxford handbook of social cognition*, pp. 282–312. Oxford University Press.

Gilbert, D. T. and J. G. Hixon (1991). The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and social Psychology 60*(4), 509–517.

Granovetter, M. S. (1973). The strength of weak ties. *American journal of sociology 78*(6), 1360–1380.

Greenwald, A. G., D. E. McGhee, and J. L. K. Schwartz (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology 74*(6), 1464–1480.

Greenwald, A. G., B. A. Nosek, and M. R. Banaji (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology 85*(2), 197–216.

Hamilton, D. L., S. J. Sherman, and C. M. Ruvolo (1990). Stereotype-based expectancies: Effects on information processing and social behavior. *Journal of Social Issues 46*(2), 35–60.

Henrich, J. (2015). *The secret of our success*. Princeton University Press.

Imana, B., A. Korolova, and J. Heidemann (2021). Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the Web Conference 2021*, pp. 33–44.

Insko, C. A., J. Schopler, and C. Sedikides (1998). Personal control, entitativity, and evolution. In C. Sedikides, J. Schopler, and C. A. Insko (Eds.), *Intergroup cognition and intergroup behavior*, pp. 109–120. Lawrence Erlbaum Associates Publishers.

Kahneman, D. (2011a). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kahneman, D. (2011b). *Thinking, fast and slow*. Macmillan.

Kahneman, D., O. Sibony, and C. R. Sunstein (2021). *Noise: A flaw in human judgment*. Little, Brown.

Kawachi, I., L. Berkman, et al. (2000). Social cohesion, social capital, and health. *Social epidemiology 174*(7), 290–319.

Kay, M., C. Matuszek, and S. A. Munson (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3819–3828. ACM.

Kearns, M. and A. Roth (2019). *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press.

Klare, B. F., M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain (2012, December). Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security 7*(6), 1789–1801.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2022). The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *arXiv preprint arXiv:2202.11776*.

Kline, P., E. K. Rose, and C. R. Walters (2022). Systemic discrimination among large us employers. *The Quarterly Journal of Economics 137*(4), 1963–2036.

Lambrecht, A. and C. E. Tucker (2019). Algorithmic bias? an empirical study into apparent gender-based discrimination in the display of STEM career ads. *Management Science 65*(7), 2966–2981.

Lane, K. A., M. R. Banaji, B. A. Nosek, and A. G. Greenwald (2007). Understanding and using the implicit association test: IV: What we know (so far) about the method. In B. Wittenbrink and S. N. (Eds.), *Implicit measures of attitudes*, pp. 59—-102. The Guilford Press.

LeVine, R. A. and D. T. Campbell (1972). *Ethnocentrism: Theories of conflict, ethnic attitudes, and group behavior.* John Wiley & Sons.

Lowe, M. (2021). Types of contact: A field experiment on collaborative and adversarial caste integration. *American Economic Review 111*(6), 1807–1844.

Lowery, B. S., C. D. Hardin, and S. Sinclair (2001). Social influence effects on automatic racial prejudice. *Journal of personality and social psychology 81*(5), 842–855.

Lueke, A. and B. Gibson (2015). Mindfulness meditation reduces implicit age and race bias: The role of reduced automaticity of responding. *Social Psychological and Personality Science 6*(3), 284–291.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts (2011, June). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, pp. 142–150. Association for Computational Linguistics.

Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54*(6), 1–35.

Milkman, K. L., M. Akinola, and D. Chugh (2012). Temporal distance and discrimination: An audit study in academia. *Psychological science 23*(7), 710–717.

Milli, S., L. Belli, and M. Hardt (2021). From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 714–722.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application 8*(1), 141–163.

Mousa, S. (2020). Building social cohesion between christians and muslims through soccer in post-isis iraq. *Science 369*(6505), 866–870.

Mullainathan, S. and E. Shafir (2013). *Scarcity: Why having too little means so much.* Macmillan.

Mussen, P. H. (1950). Some personality and social factors related to changes in children's attitudes toward negroes. *The Journal of Abnormal and Social Psychology 45*(3), 423–441.

Paluck, E. L. and D. P. Green (2009). Prejudice reduction: What works? a review and assessment of research and practice. *Annual review of psychology 60*(1), 339–367.

Paluck, E. L., R. Porat, C. S. Clark, and D. P. Green (2021). Prejudice reduction: Progress and challenges. *Annual review of psychology 72*(1), 533–560.

Patacchini, E. and Y. Zenou (2008). The strength of weak ties in crime. *European Economic Review 52*(2), 209–236.

Payne, B. K., A. J. Lambert, and L. L. Jacoby (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology 38*(4), 384–396.

Pettigrew, T. F. and L. R. Tropp (2006). A meta-analytic test of intergroup contact theory. *Journal of personality and social psychology 90*(5), 751–753.

Putnam, R. D. (2000). Bowling alone: America's declining social capital. In L. Crothers and C. Lockhart (Eds.), *Culture and politics: A reader*, pp. 223–234. Palgrave Macmillan.

Raghavan, M., S. Barocas, J. Kleinberg, and K. Levy (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 469–481.

Richeson, J. A. and N. Ambady (2003). Effects of situational power on automatic racial prejudice. *Journal of Experimental Social Psychology 39*(2), 177–183.

Sandstrom, G. M. and E. W. Dunn (2014). Social interactions and well-being: The surprising power of weak ties. *Personality and Social Psychology Bulletin 40*(7), 910–922.

Schwab, A. K., C. Sagioglou, and T. Greitemeyer (2019). Getting connected: Intergroup contact on facebook. *The Journal of social psychology 159*(3), 344–348.

Sherif, M. (1988). *The robbers cave experiment: Intergroup conflict and cooperation*. Wesleyan University Press.

Sherman, J. W. and S. A. Klein (2021). The four deadly sins of implicit attitude research. *Frontiers in Psychology 11*, 604340.

Sherman, J. W., A. Y. Lee, G. R. Bessenoff, and L. A. Frost (1998). Stereotype efficiency reconsidered: Encoding flexibility under cognitive load. *Journal of personality and social psychology 75*(3), 589–606.

Sweeney, L. (2013). Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue 11*(3), 10–29.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social science information 13*(2), 65–93.

Takagi, E. (1996). The generalized exchange perspective on the evolution of altruism. In *Frontiers in social dilemmas research*, pp. 311–336. Springer.

Talhelm, T., X. Zhang, S. Oishi, C. Shimin, D. Duan, X. Lan, and S. Kitayama (2014). Large-scale psychological differences within China explained by rice versus wheat agriculture. *Science 344*(6184), 603–608.

Tassier, T. (2006). Labor market implications of weak ties. *Southern Economic Journal 72*(3), 704–719.

Todorov, A., A. N. Mandisodza, A. Goren, and C. C. Hall (2005). Inferences of competence from faces predict election outcomes. *Science 308*(5728), 1623–1626.

Todorov, A., C. Y. Olivola, R. Dotsch, P. Mende-Siedlecki, et al. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual review of psychology 66*(1), 519–545.

Unkelbach, C., J. P. Forgas, and T. F. Denson (2008). The turban effect: The influence of Muslim headgear and induced affect on aggressive responses in the shooter bias paradigm. *Journal of Experimental Social Psychology 44*(5), 1409–1413.

Vargas, P. T. (2004). The relationship between implicit attitudes and behavior: Some lessons from the past, and directions for the future. In *Contemporary perspectives on the psychology of attitudes*, pp. 275–298. Psychology Press.

Vasileva, M. I. (2020). The dark side of machine learning algorithms: how and why they can leverage bias, and what can be done to pursue algorithmic fairness. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3586–3587.

Voigt, R., N. P. Camp, V. Prabhakaran, W. L. Hamilton, R. C. Hetey, C. M. Griffiths, D. Jurgens, D. Jurafsky, and J. L. Eberhardt (2017). Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences 114*(25), 6521–6526.

Wang, C., B. Han, B. Patel, and C. Rudin (2022). In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, 1–63.

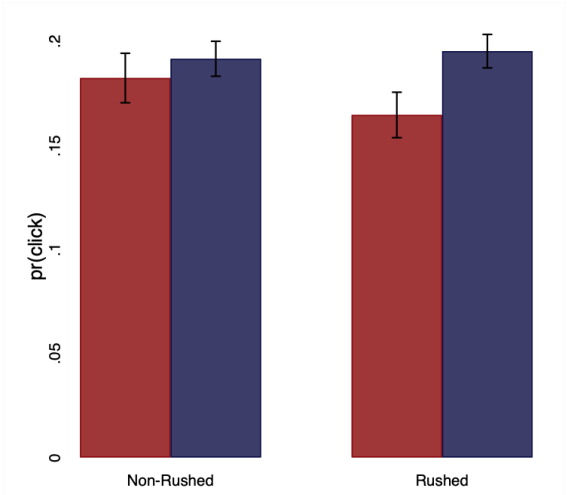
Williams, D. H. (1948). The effects of an interracial project upon the attitudes of negro and white girls within the YWCA. In A. Rose (Ed.), *Studies in Reduction of Prejudice*. American Council of Race Relations.

Wittenbrink, B., C. M. Judd, and B. Park (2001). Spontaneous prejudice in context: variability in automatically activated attitudes. *Journal of personality and social psychology 81*(5), 815.

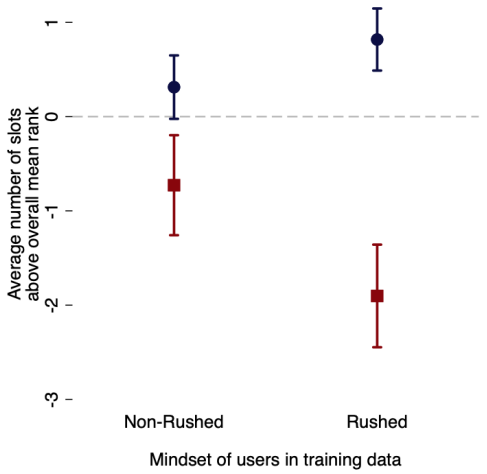
Wühr, P., B. P. Lange, and S. Schwarz (2017). Tears or fears? comparing gender stereotypes about movie preferences to actual preferences. *Frontiers in psychology 8*, 428.

Zemel, R., Y. Wu, K. Swersky, T. Pitassi, and C. Dwork (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning 28*(3), 325–333.

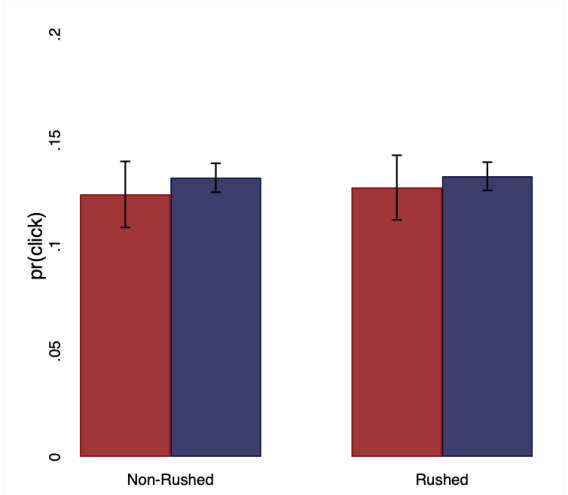
Figure 1: Movie Recommendation Lab Experiment Results



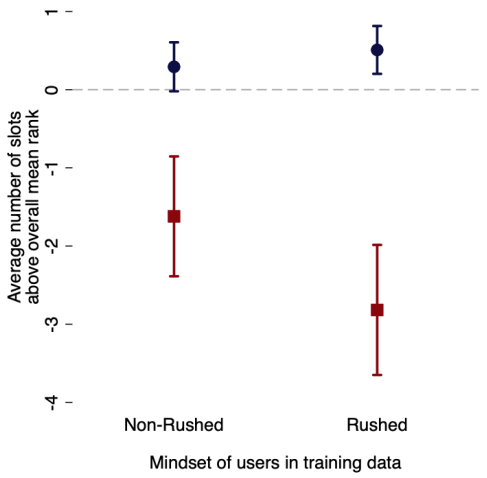

(a) Probability Choose Movie by Treatment and Recommender Type: Experiment 1




(b) Algorithmic Ranking Trained on Experiment 1 Choices



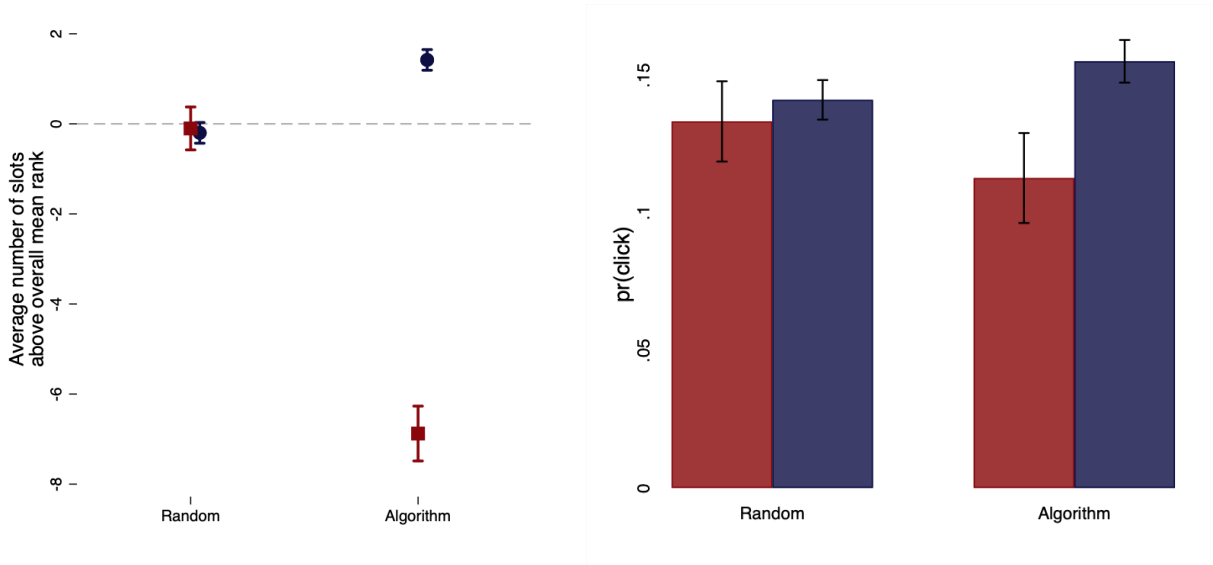

(c) Probability Choose Movie by Treatment and Recommender Type: Experiment 2




(d) Algorithmic Ranking Trained on Experiment 2 Choices


**Note:** Panels (a) and (b) show the results of Lab Experiment 1, which enlists a representative sample of US residents into an on-line choice task where we randomize whether choice options are signaled to be recommended by own-group vs. out-group members by randomizing the name of the fictive recommender, and also randomize whether subjects make choices about which content to engage with in a rushed vs. more deliberate decision (non-rushed) context. Panel (a) shows the engagement patterns for own-group vs. out-group recommended content, by rushed and non-rushed conditions. Panel (b) shows the results of using the data from the rushed and non-rushed conditions separately to build two separate algorithmic predictions that rank content by subject engagement choices. We then present the rank-orderings (relative to the overall mean rank) for content separately for own-group and out-group recommended content for the algorithm built using data made in a more deliberate, non-rushed context (left) or rushed context (right). Panels (c) and (d) replicate the analysis using data from Lab Experiment 2, which enrolled a more homogenous sample of white US males. Own-group is defined as same race or gender.

Figure 2: Movie Recommendation Lab Experiment Results: Random versus Algorithmic Ranking

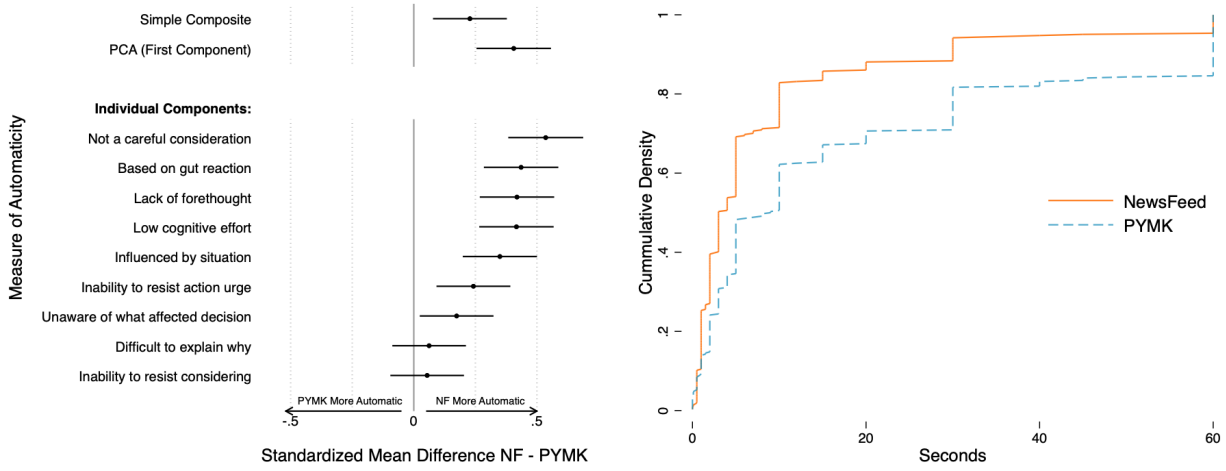



(a) Ranking of posts in Random vs. Algorithmic Ranking Treatments

(b) Probability Choose Movie by Treatment and Recommender Type: Experiment 3

**Note:** This figure shows the results from Lab Experiment 3, in which we asked a new sample of subjects (different from those enrolled in lab experiments 1 and 2) to make choices about engaging in content when those content options are either ranked randomly, or ranked by a machine-learning algorithm. Panel (a) at left shows the relative ranking of own-group and out-group content compared to the overall mean rank when we rank randomly; as expected, there is little difference in the average rankings for own-group and out-group recommended content when ranked randomly. At the right in panel (a), we show the rank-ordering relative to overall mean of own-group and out-group recommended content from a recommender algorithm that we built using data from the rushed condition in Lab experiment 2 as the training dataset; we can see there is now a sizable disparity in the rank-orderings of content recommended by own-group versus out-group members. Panel (b) shows at the left that subjects have a preference for own-group recommended content when choice options are randomly ranked (the "single penalty" for out-group content created by human bias), while at the right we show that the rate at which own-group content is favored increases substantially when the content is now rank-ordered by the recommender algorithm that up-ranks own-group content (the 'double penalty' that comes on top of the single penalty of implicit bias in the subject choices). Own-group is defined as same race or gender.

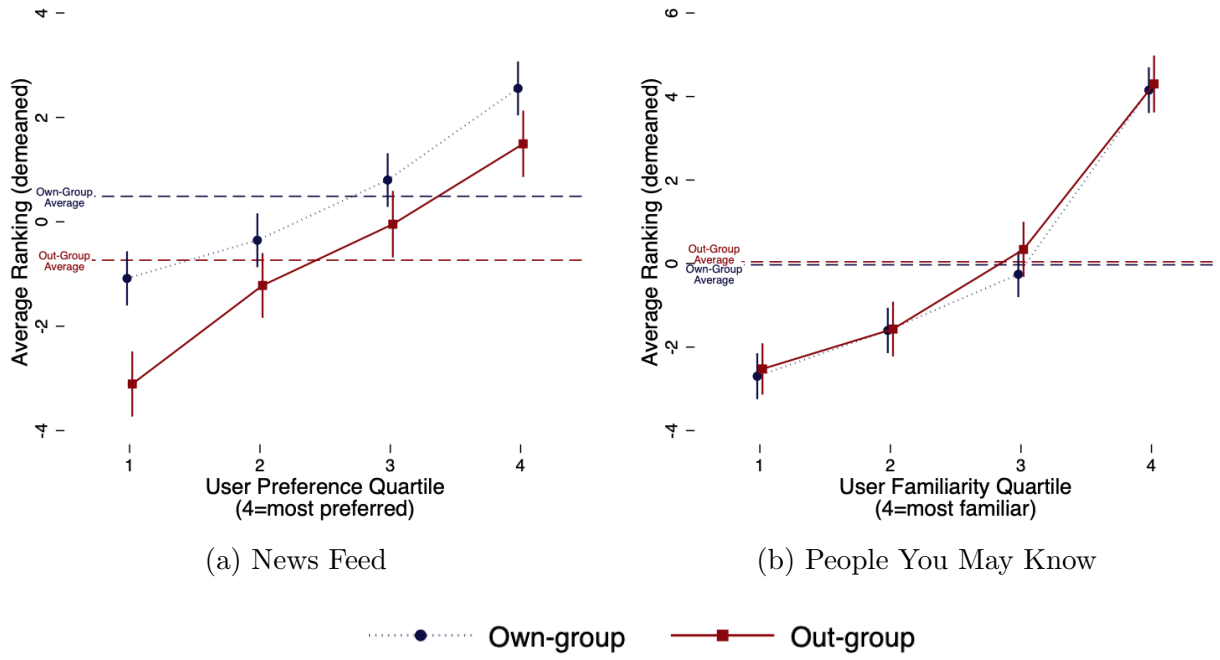
Figure 3: How Automatic are Users When Choosing to Engage with News Feed Content versus People You May Know Suggestions




(a) Standardized Effect Sizes (b) Speed (time to decide in seconds) CDF

**Note:** In a survey, we collected 10 measures of deliberateness in making decisions to engage in content on the News Feed and the People You May Know recommendations. See Appendix C for details on the sample and the questions. In Panel (a), the standardized effect size is the mean response for the News Feed and the mean response for PYMK divided by the pooled standard deviation. These are shown for NF − PYMK for the various measures of automaticity described in Appendix C. Higher values indicate more automatic for all questions except time where more seconds indicates more such that values $> 0$ imply more automatic decision making in News Feed versus PYMK. The first two measures are a simple composite average of the underlying individual components and the first principal component of those underyling individual measures. Panel (b) shows the CDF of self-reported typical time to decide to take an action on deciding to engage in content on News Feed and PYMK (top-coded at 60 seconds). A Kolmogorov-Smirnov test rejects that these two CDFs are the same with $p < 0.01$.

Figure 4: Relationship between News Feed Algorithmic Ranking and Own-group Status Conditional on Subject Explicit Preference

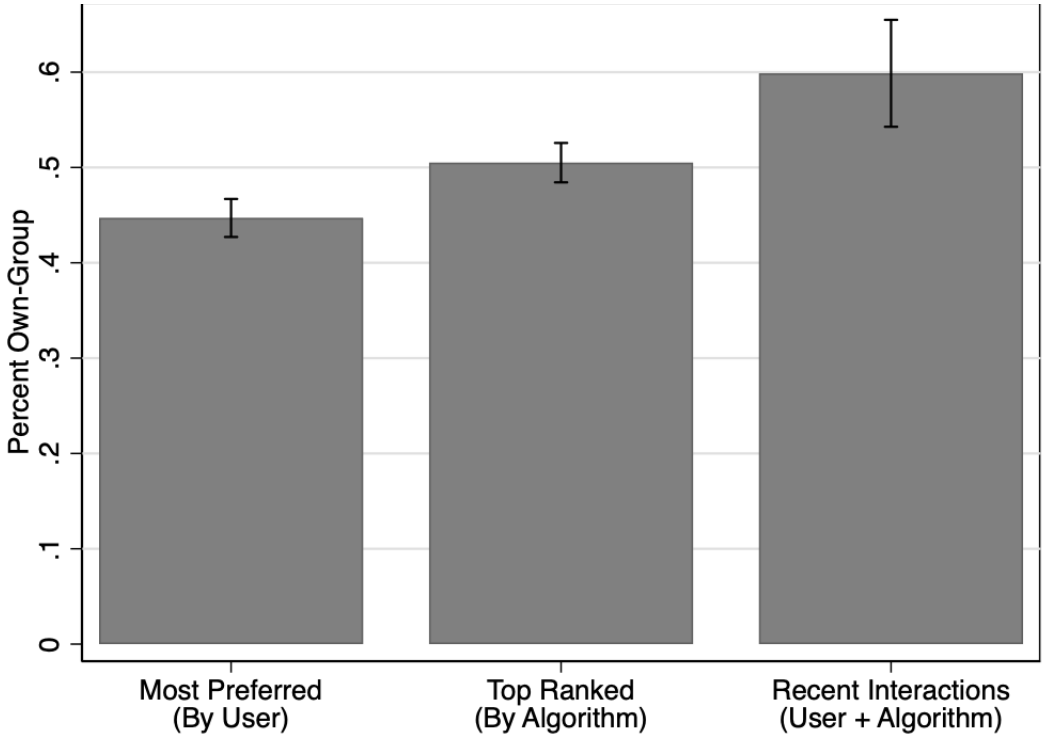



(a) News Feed

(b) People You May Know

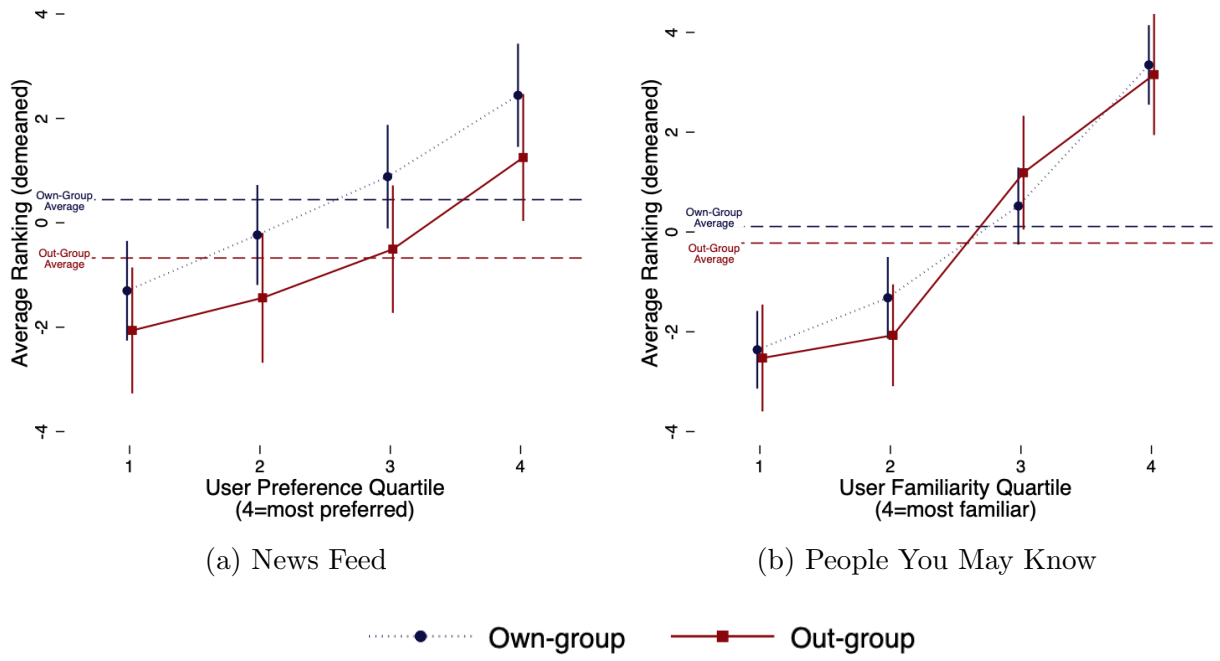
**Note:** On the left of each panel, we show the mean ranking of own-group and out-group posts/recommendations above the overall mean, and then we show this by subject stated preference/familiarity. The normalized subject explicit preference/familiarity quartile is the across-subject quartile of within-subject z-scores for stated preference for a post/familiarity with a suggested friend. Each subject's ratings were mean-centered and then divided by the subject's standard deviation of responses. The resulting distribution was then split into four equally sized bins. Own-group is defined as same race.

Figure 5: Share Own-Group by User Preference, Algorithmic Ranking, and Recent Interactions




**Note:** For a subset of individuals, we collected information on 10 posts users had recently interacted with on News Feed (this does not necessarily have to be any of the posts currently on their News Feed). Recent interactions include the 10 most recent "likes," reactions, and/or comments. This figure shows the percent of those interactions that are on own-group posts (far right bar). We also show the the percent own-group in the first 10 posts as ranked by the algorithm ("Top Ranked (By Algorithm)"). We further show the percent own-group for the posts that are most preferred by the user. To define most-preferred, we sorted respondents' posts by their raw stated preference; we then defined as most-preferred those posts whose preference rating was the same or larger than the post ranked 10 by this ranking (this results in more than 10 "most preferred" for most respondents; in Appendix Figure F.6 we repeat this analysis but choosing only the top 10 most preferred based on a random ranking of ties). The interactions analysis is based on 102 participants who were asked to show their recent activity (interactions) in Waves 4 and 5. The most preferred and top ranked analysis is based on the 654 participants for whom we have News Feed data (from Waves 1-2, and 4-6). Appendix Figure F.5 repeats this analysis restricted to only the 102 participants asked about recent activity.

Figure 6: Relationship between News Feed Algorithmic Ranking and Own-group Status Conditional on Subject Explicit Preference: India Sample




(a) News Feed

(b) People You May Know

**Note:** This figure recreates Figure 4 for our India sample. Own-group is defined as same religion. On the left of each panel, we show the mean ranking of own-group and out-group posts/recommendations above the overall mean, and then we show this by subject stated preference/familiarity. The normalized subject explicit preference/familiarity quartile is the across-subject quartile of within-subject z-scores for stated preference for a post/familiarity with a suggested friend. Each subject's ratings were mean-centered and then divided by the subject's standard deviation of responses. The resulting distribution was then split into four equally sized bins.

Figure 7: Cross-Platform Automaticity Survey

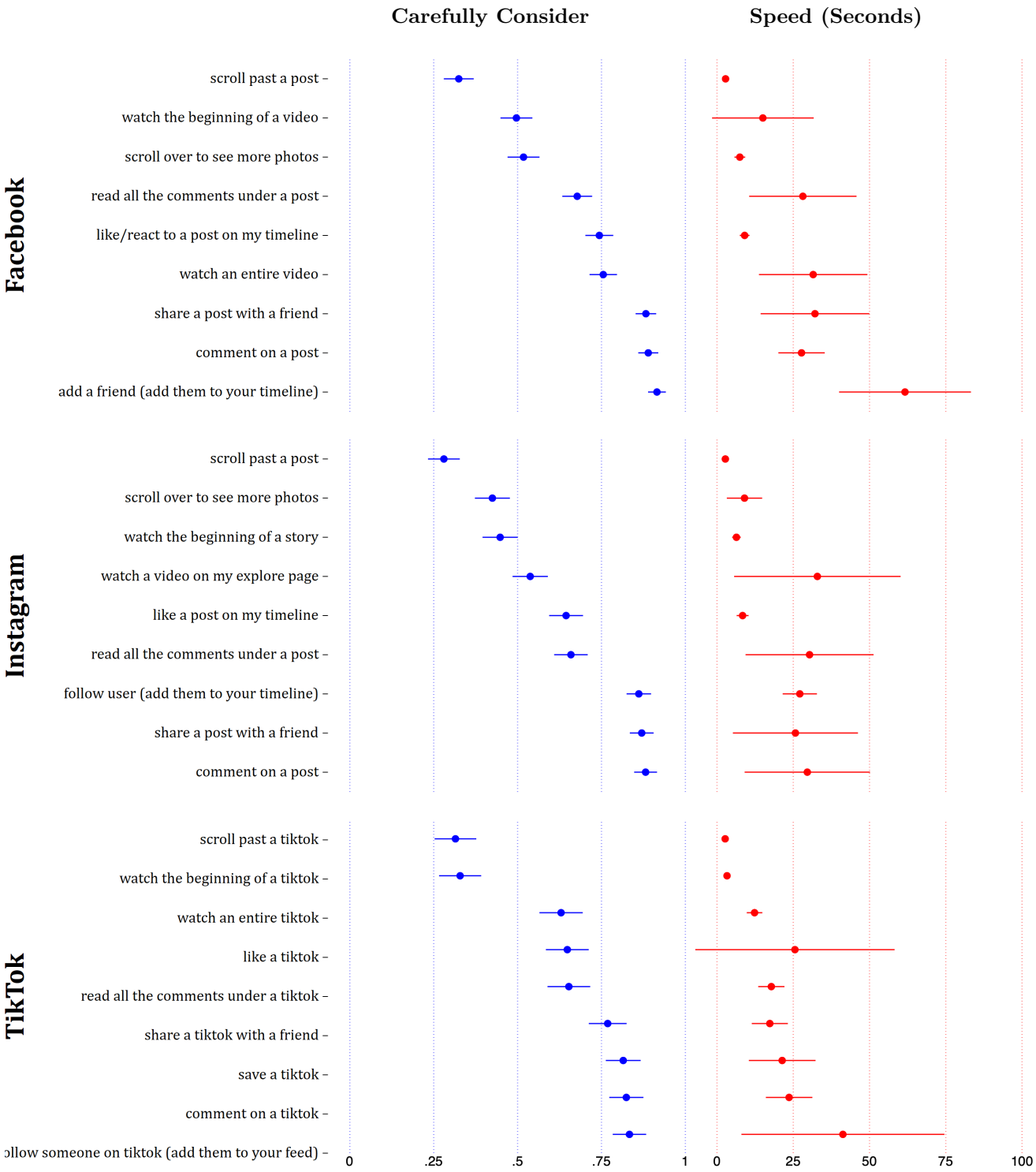
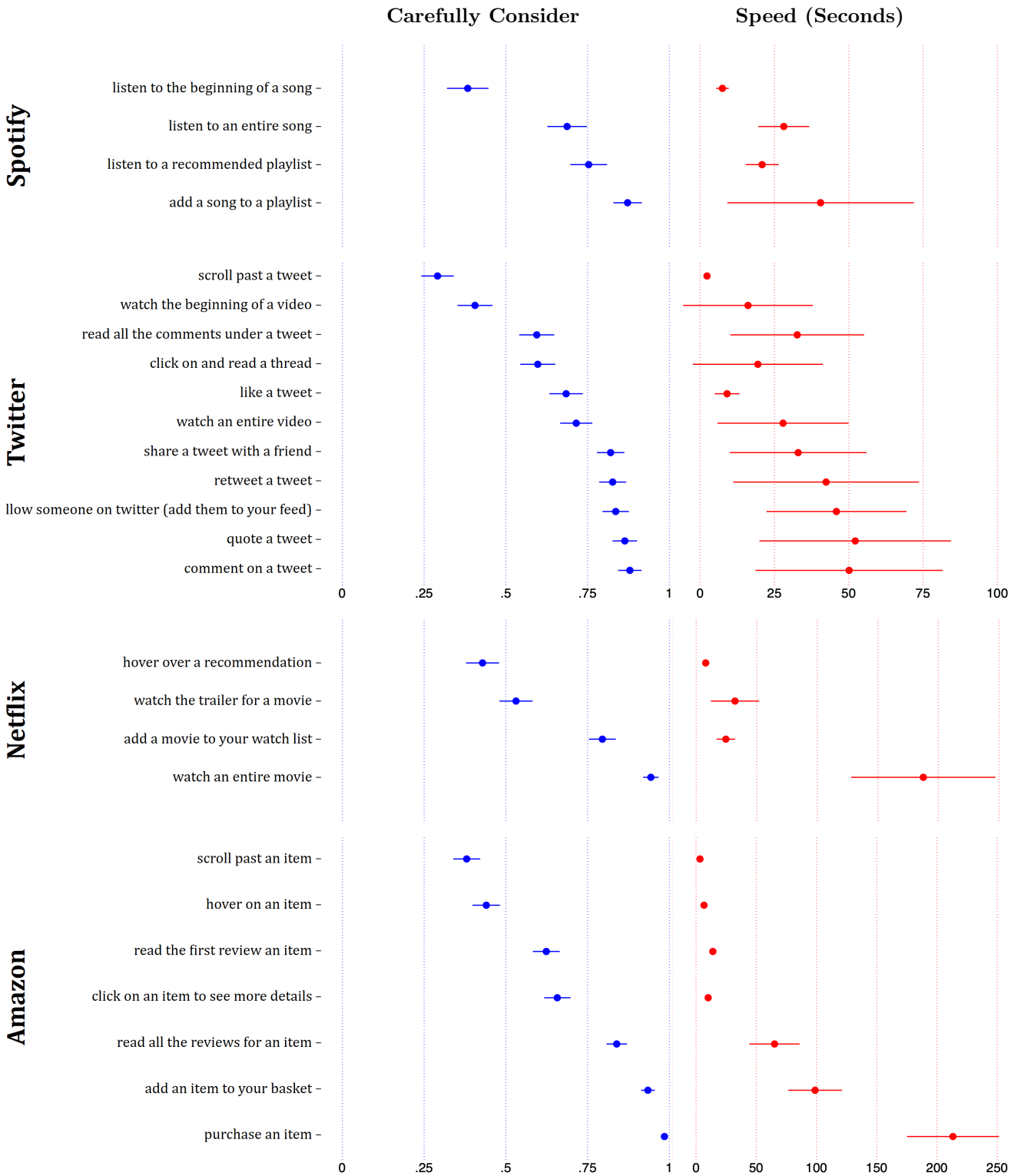


Figure 7 (cont.): Cross-Platform Automaticity Survey




**Note:** Survey sample details can be found in Appendix H. For any platform respondents said they used in the past month, they were asked two questions. "Carefully Consider" is the average response to "My decision to {behavior} is usually based on careful consideration"; "Speed (Seconds)" is the average response to "How much time do you usually spend thinking about it before you take the following action".

Table 1: Simulation Results: Average p-Values for Test of Own- vs. Out-group Differences for Clicks & Predicted Clicks

| Sample Size | (1) Clicks | (2) Predicted Clicks |
|---|---|---|
| 50 | 0.422 | 0.092 |
| 100 | 0.370 | 0.080 |
| 500 | 0.122 | 0.015 |
| 1,000 | 0.030 | 0.002 |

**Note:** We simulated 1,000 samples of size N, as listed in the table, equally split between own- and out-group. Simulated click rates follow $c_1 = 0.5 + \delta + x_1 + z_1$ for own-group posts and $c_0 = 0.5 + x_0 + z_0$ for out-group; here $\delta = 0.01$, $z_i \sim N(0, 0.05)$, and $x_i \sim N(0, 0.01)$. Columns (1) and (2) show the average p-value across the 1,000 simulated datasets at each specified sample size from a 2-tailed test of equal mean between own- and out-groups for the "raw" simulated click data (column 1) and the algorithm's prediction of clicks (column 2), which comes from a linear regression of clicks on group status and $x$.

Table 2: Lab Experiment Regression Results

| | Experiment 1 | | Experiment 2 | | Experiment 3 | |
|---|---|---|---|---|---|---|
| | (1) Human Choice | (2) Slots above Mean Ranking | (3) Human Choice | (4) Slots above Mean Ranking | (5) Slots above Mean Ranking | (6) Human Choice |
| Own-group | 0.009 | 1.040*** | 0.008 | 1.913*** | -0.098 | 0.008 |
| | (0.007) | (0.321) | (0.009) | (0.422) | (0.268) | (0.008) |
| Rushed | -0.018** | -1.175*** | 0.003 | -1.197** | | |
| | (0.008) | (0.388) | (0.011) | (0.577) | | |
| Rushed x Own-group | 0.021** | 1.680*** | -0.003 | 1.413** | | |
| | (0.010) | (0.456) | (0.012) | (0.619) | | |
| Algorithmic | | | | | -6.775*** | -0.021* |
| | | | | | (0.378) | (0.011) |
| Algorithmic x Own-Group | | | | | 8.395*** | 0.035*** |
| | | | | | (0.413) | (0.012) |
| Constant | 0.182*** | -0.728*** | 0.124*** | -1.621*** | -0.102 | 0.133*** |
| | (0.006) | (0.271) | (0.008) | (0.391) | (0.241) | (0.007) |
| Observations | 26268 | 10656 | 22833 | 10482 | 20952 | 20952 |
| Mean Dep Var | 0.187 | 0 | 0.131 | 0 | 0 | 0.144 |

**Note:** This table reports regression results for our three lab experiments. These mirror the results seen in Figures 1 and 2. Human Choice is a binary variable = 1 for whether the participant chose the movie or not. As described in Appendix B, human choices were used as an input into a machine learning algorithm to rank recommendations according to their likelihood of being selected. Columns (2), (4), and (5) have as a dependent variable the rank-ordering of a movie/recommendation relative to the overall mean rank, such that a 1-unit increase implies 1 slot above the overall mean rank (closer to the top), where in Columns (2) and (4) this is what is predicted by the algorithm but not actually shown, and in Column (5) is the rank ordering above the overall mean rank shown to the user. Subjects in Lab Experiments 1 and 2 were randomized between making movie choices in a "rushed" or "non-rushed" condition, with the order of the movie recommendations being random. In Lab Experiment 3, subjects were randomized between seeing movies in a random order or an algorithmically determined order, but all made decisions in a "rushed" condition. Lab Experiment 1 used participants meant to be representative of the US population. Lab Experiments 2 and 3 focused on only white, male participants. See Appendix B for more details on the data collection.

* $p < 0.10$ **$p < 0.05$ ***$p < 0.01$

Table 3: US Facebook Subject Summary Statistics

| | (1) Newsfeed | (2) PYMK | (3) Activity |
|---|---|---|---|
| Race of subject (RA) | | | |
| Asian | 0.409 | 0.445 | 0.308 |
| Black | 0.091 | 0.087 | 0.115 |
| Hispanic | 0.091 | 0.073 | 0.125 |
| Other Race | 0.014 | 0.016 | 0.000 |
| White | 0.396 | 0.379 | 0.452 |
| Race of Subject (Self Identification) | | | |
| Asian | 0.385 | 0.429 | 0.260 |
| Black | 0.071 | 0.078 | 0.077 |
| Hispanic | 0.066 | 0.046 | 0.087 |
| Other Race | 0.024 | 0.023 | 0.029 |
| White | 0.353 | 0.336 | 0.423 |
| Two or more Races | 0.069 | 0.062 | 0.087 |
| Gender | | | |
| Male | 0.261 | 0.269 | 0.221 |
| Female | 0.681 | 0.676 | 0.721 |
| Non-binary | 0.027 | 0.027 | 0.019 |
| Age | 26.639 | 27.138 | 26.090 |
| Less than Bachelor's Degree | 0.379 | 0.358 | 0.385 |
| Bachelor's Degree | 0.346 | 0.347 | 0.317 |
| Graduate Degree | 0.224 | 0.240 | 0.260 |
| Average Facebook Usage | | | |
| Hourly | 0.119 | 0.123 | 0.125 |
| Daily | 0.517 | 0.534 | 0.481 |
| Weekly | 0.234 | 0.224 | 0.279 |
| Monthly | 0.074 | 0.073 | 0.067 |
| Yearly | 0.018 | 0.014 | 0.000 |
| Never | 0.008 | 0.005 | 0.010 |
| Within past hour | 0.473 | 0.486 | 0.423 |
| Last Facebook Log-in | | | |
| Within past day | 0.364 | 0.349 | 0.442 |
| Within past week | 0.098 | 0.112 | 0.067 |
| Within past month | 0.030 | 0.021 | 0.029 |
| Within past year | 0.005 | 0.005 | 0.000 |
| Total Facebook Friends | 810.386 | 836.215 | 864.524 |
| Mean NF Post Preference | 3.439 | 3.338 | 3.518 |
| Standard Dev of NF Post Preference | 1.640 | 1.625 | 1.698 |
| Mean PYMK Rec Familiarity | 2.351 | 2.351 | . |
| Standard Dev of PYMK Rec Familiarity | 1.636 | 1.636 | . |
| Observations | 662 | 438 | 104 |

**Note:** In each wave of the data collection in the US, we collected information about News Feed posts. Thus Column (1) shows summary statistics on all US participants. In Waves 1 and 2, we also collected information about recommendations on the People You May Know (PYMK) page; summary statistics for those respondents are in Column (2). And in Wave 6, we collected information on the 10 most recent interactions. Summary statistics for those respondents are in Column (3).

Table 4: Facebook Regression Results (US)

| | Newsfeed Posts | | PYMK Recommendations | |
|---|---|---|---|---|
| | (1) Slots Above Mean Ranking | (2) In Top 10 | (3) Slots Above Mean Ranking | (4) In Top 10 |
| Own-group (Race) | 1.192*** | 0.022*** | -0.236 | -0.006 |
| | (0.208) | (0.005) | (0.217) | (0.005) |
| Preference/Familiarity | 1.693*** | 0.030*** | 2.874*** | 0.052*** |
| | (0.105) | (0.002) | (0.108) | (0.003) |
| Constant | -0.653*** | 0.171*** | 0.138 | 0.172*** |
| | (0.161) | (0.004) | (0.166) | (0.004) |
| Observations | 28747 | 28747 | 25593 | 25593 |
| Dep Var Mean | 0 | 0.183 | 0 | 0.168 |

**Note:** These show regression versions of our main Facebook outcomes by own-group and user self-reported preference/familiarity. Slots above mean ranking is rank-ordering of a movie/recommendation relative to the overall mean rank, such that a 1-unit increase implies 1 slot above the overall mean rank (closer to the top). In Top 10 is a binary variable indicating whether the post or recommendation was among the first 10 shown to the user. User preference/familiarity is normalized within user.

Table 5: Summary Statistics on Collected Facebook Outcomes (US)

| | Newsfeed Posts | | | PYMK Recommendations | | |
|---|---|---|---|---|---|---|
| | (1) All | (2) Ingroup | (3) Outgroup | (4) All | (5) Ingroup | (6) Outgroup |
| Own-group (Race) | 0.601 | | | 0.585 | | |
| Race of Poster/Rec | | | | | | |
| Asian | 0.257 | | | 0.306 | | |
| Black | 0.102 | | | 0.086 | | |
| Hispanic | 0.100 | | | 0.089 | | |
| Other Race | 0.028 | | | 0.035 | | |
| White | 0.499 | | | 0.469 | | |
| Pref./Familiarity Ranking | | | | | | |
| 1st Quartile | 0.249 | 0.247 | 0.253 | 0.250 | 0.238 | 0.266 |
| 2nd Quartile | 0.251 | 0.249 | 0.253 | 0.250 | 0.249 | 0.251 |
| 3rd Quartile | 0.250 | 0.254 | 0.245 | 0.250 | 0.253 | 0.246 |
| 4th Quartile | 0.250 | 0.251 | 0.249 | 0.250 | 0.259 | 0.238 |
| Ranking | 29.265 | 28.777 | 30.000 | 30.350 | 30.380 | 30.307 |
| In First 10 Posts/Recs | 0.183 | 0.192 | 0.169 | 0.168 | 0.167 | 0.170 |
| Post to Group | 0.365 | | | | | |
| Days Ago Posted | 1.134 | 1.120 | 1.156 | | | |
| Mutual Friends | | | | 22.933 | 21.902 | 24.385 |
| Observations | 28747 | 17271 | 11476 | 25593 | 14969 | 10624 |

**Note:** This table shows summary statistics on News Feed Posts and friend recommendations from the PYMK algorithm for our US Facebook data collection. News Feed posts are restricted to posts from "humans" and not from companies/firms. The normalized subject-explicit preference/familiarity quartile is the across-subject quartile of within-subject z-scores for stated preference for a post/familiarity with a suggested friend.

# Supplemental Materials

for "Algorithmic Curation Creates Bias: Theory, Experiment, and Evidence From Facebook"

# A A Brief Primer on the Behavioral Science of Prejudice

In this Appendix, we connect some of the key behavioral science ideas discussed briefly in the main text of our paper to the voluminous literature in psychology that supports those points. Because the literature is vast, we focus here on summarizing the key ideas and references to some of the important papers in each sub-literature, which will in turn include pointers to the other excellent papers in the field.

## A.1 (In)groups are central to our lives

Many scholars believe that the ability of people to cooperate with one another in groups has been key to our success over the course of human evolution (LeVine and Campbell, 1972; Caporael, 1997; Talhelm et al., 2014; Henrich, 2015). That historically adaptive tendency remains a part of human psychology today as well, but when applied in the context of modern society, it can create a number of maladaptive outcomes—like own-group favoritism and out-group disfavor or prejudice (Takagi, 1996; Insko et al., 1998; Brewer, 1999).

The definitions of "own-group" and "out-group" depend partly on what categorizations are most salient in a given context. Our automatic (System 1) cognition seems to be particularly sensitive to key demographic features like gender, age, and race or ethnicity (Todorov et al., 2015). Further, we do not merely favor people based on their (perceived) group identity, but also draw inferences about what other people are like based on these characteristics (Brewer, 1988; Fiske and Neuberg, 1990), (Abele et al., 2021). Importantly, while race, age and ethnicity are often central to our sense of identity, far more subtle and arbitrary cues can cause our minds to draw distinctions between "us" and "them." For example, in the seminal "Robbers Cave" study, Sherif (1988) shows that even superficial conflict is sufficient to generate out-group hostility by randomly assigning middle school boys to two groups (the Eagles and the Rattlers), who were pitted against each other in a few small competitions. After just a couple weeks of meeting their group mates and competing with the other group, the boys exhibited increasingly negative views about the trustworthiness, integrity, and athletic skill of members of the other group. Taking this idea further, Tajfel (1974) shows that the mere presence of groups (without any real import to the members) is sufficient to generate own-group favoritism. In his famous *minimal group paradigm*, he randomly assigns teenage boys to a "Kandinsky group" and "Klee group," telling them that the assignment was based on their apparent preferences for abstract art. The boys were then tasked with allocating rewards between two unnamed participants, one in their own group and one in the other group. Tajfel first finds that when given the choice between maximizing the profit for all and maximizing the profit for their own group, participants chose the latter. More strikingly, when given the choice between maximizing the profit for their own group and maximizing the difference between their own group and the other group, participants chose the latter, indicating a preference against the out-group.

That not much is required for people to perceive others as "other" renders those judgements quite dependent on superficial details of the context and situation. Take for example another

well-known experiment: Frank and Gilovich (1988) measure levels of interest in aggressive activities (e.g., chicken fights and dart gun duels) between members of two groups at two moments in time: once before assigning the groups to wear different color jerseys and once after the jerseys were assigned and worn. Before the jerseys were worn, there was no detectable difference in aggressive-activity responses between the two groups. But once the jerseys were worn, the participants wearing black jerseys sought out more aggressive activities with the other group.

Given what is known from behavioral science about the susceptibility of people to treating own-group and out-group members differently in real life, there would seem to be reason to worry the same tendencies may manifest themselves in online social environments as well.

## A.2 Automaticity and prejudice

Over time, self-reported measures of prejudice in the US have fallen dramatically (see for example Bobo et al. 1972; Charles and Guryan 2008). These measures speak to the conscious choices and attitudes that people endorse for themselves; they capture our System 2 preferences. However, the automatic System 1 preferences that have been passed on to us through evolution retain the same instinctual tendency towards own-group favoritism described above. So despite the fact that our conscious choices have become less biased, our subconscious choices retain the biases that have been historically functional (Gilbert and Hixon, 1991; Hamilton et al., 1990; Sherman et al., 1998; Unkelbach et al., 2008).

The tension between our conscious and subconscious attitudes is moderated by the extent to which we can actively inhibit our gut responses (Dovidio et al., 1997; Devine, 1989). That cognitive effort is required to reign in our worst selves bestows a central role to the effect of automaticity on how prejudices are revealed. When we think slowly, we may be biased; but when we think quickly, we are even more biased. The logic that our automatic choices will be more biased than our more deliberate choices is supported by a vast body of empirical work across several domains inside and outside the lab (see for example Todorov et al. 2005; Richeson and Ambady 2003; Lowery et al. 2001; Eberhardt et al. 2004, 2006; Voigt et al. 2017). In short, thinking fast seems to facilitate many of the prejudiced behaviors we see in the world. This is important because the specifics of our social interactions are not usually a series of deliberate, controlled choices that we consciously decide. Instead, they are often the result of situational factors that affect our subconscious thinking (see for example Ferguson and Bargh 2004; Wittenbrink et al. 2001).

Implicit attitudes and biases are one component of the difference between thinking fast and slow, which has received tremendous attention in psychological research seeking to understand prejudiced behavior. This is usually measured via the Implicit Associations Test (IAT), which seeks to measure the strength of association between concepts (such as own-group/out-group affiliations) with normative evaluations (good/bad) or stereotyped attributions (see for example Greenwald et al. 1998, 2003). However, despite evidence that more automatic behaviors exhibit more bias than more deliberate behaviors, there remains ongoing debate about whether specific IAT measures are good predictors of behavior (see for example Vargas

2004; Lane et al. 2007; Dai and Albarracín 2022; Sherman and Klein 2021).

## A.3 Intergroup contact, a way forward

The behavioral science insights into prejudice mean that our results have direct implications for society. To see how, consider what is thought to be one of the most important ways to reduce prejudice: intergroup contact. Inspired by early studies suggesting that housing and workplace desegregation in the United States reduced prejudice toward Black people (Williams 1948; Mussen 1950), Allport (1954) argued that interactions between people of different groups will allow them to know each other as individuals and learn their true nature; that we are not so different after all, despite the histories and stereotypes associated with our various groups.[32] Since then, hundreds of studies have investigated the idea that intergroup interactions reduce prejudice (e.g., Pettigrew and Tropp 2006; Paluck and Green 2009; Paluck et al. 2021). For example, in recent work, Mousa (2020) randomly assigns Iraqi Christians displaced by the Islamic State of Iraq and Syria (ISIS) to an all-Christian soccer team or to a team mixed with Muslims and finds that the intervention improved behaviors toward Muslim peers. (See also Lowe 2021 for a similar design and result in India.)

[32] Allport also laid out four conditions (equal status in the situation, common goals, cooperation between groups, and support of norms and laws in the environment) for the interaction which he suspected were necessary conditions for the interactions to result in prejudice reduction.

# B Lab Experiment Materials and Methods

## B.1 Data Collection for Experiments 1 & 2

Each experiment has the following participant flow:

1. **Data consent**: Participants provide informed consent as approved by the University of Chicago IRB (21-0412).

2. **User profile**: Race/ethnicity (participants could select as many categories as desired from the 7 US Census categories, including "Other"), Gender (Male, Female, Non-binary), Age, Education, "How often do you watch movies?" (Every day, Several times a week, Several times a month, Once a month, A few times a year, Once a year, Never), "What is your favorite movie genre?" (Action and adventure, Animation, Comedy, Drama, Historical, Horror, Science Fiction, Other).

3. **Overview of task**: All subjects are given the following overview of the task.

   Thank you for joining our study! We are building a movie recommendation algorithm and we'd like you to help by completing the following task. In this task you will:

   - choose **four (4)** movies
   - receive a link to watch one of your selected movies
   - choose only **movies you want to watch**

   To help you choose we've given you real user reviews for each movie

4. **Experimental instructions**: Participants in the deliberate condition were given the instructions, "You will have 15 minutes to complete the task. This is plenty of time, so please read the reviews carefully and do your best!" Participants in the rushed condition were given the instructions "You will have 5 minutes to complete the task. This is not much time, so please read the reviews carefully and do your best!"

5. **Choice task**: Participants were shown randomized real movie recommendations (taken from the public dataset used in Maas et al. (2011)), though the names of the recommenders were randomly assigned to signal race and gender.[33] The task was set up such that only three movies could be seen at a time. Participants could click a "See more" button and scroll to see the remaining options. There were 42 options in total. Participants did not need to view all the options, but they did need to select four movies before moving on. For each recommendation, participants could also click to open up a pop-up box showing the full text of the recommendation. Participants in the deliberate condition are shown a countdown of minutes and seconds on the left panel of the screen; participants in the rushed condition are shown a countdown of minutes and seconds *and milliseconds*. See Figure B.1 for screenshots that show the

[33] Names were taken from Bertrand and Mullainathan (2004), Agan and Starr (2018), Milkman et al. (2012) and include, for example, John E., Ryan S., Juan R., Juanita L., Meredith H., Darnell P., Jamal F., and Gabriella S.

general instructions, and the differences in instructions and the countdown clock for the rushed and non-rushed conditions; panels (c) and (e) also show how the movie recommendations appeared to the participant.

6. **Endline**: Participants were asked the following Likert scale (1-7) questions:

   - How would you rate the selection of movies available to you?
   - How satisfied were you with the movies you chose?
   - How much did you rely on the recommendations to make your choice?
   - How likely are you to use your earnings to rent the movie?

   As well as the following free-response questions:

   - What do you think this study is about?
   - Anything else you'd like to share with us?

7. **Reward**: Finally, participants receive the code to input into Prolific to receive payment and are simultaneously given a link to watch one of their selected movies. It is not possible to check how often people watch the movie provided, but 72% of people clicked the link.

Each experiment thus generates a dataset in which approximately half of the movies were browsed in a rushed context and the rest in a more deliberate context.

## B.2 Building the Algorithm

There are 10 inputs to the algorithm:[34]

1. Genre match: An indicator set to 1 if a movie is in the user's favorite genre and 0 otherwise.
2. Favorite genre: Favorite genre taken as given from the participant.
3. Frequency: Movie watching frequency taken as given from participant.
4. Rating: The IMDB rating of the movie (note that this rating is never seen by participants).
5. Recommender gender: Gender signalled by the randomly assigned name.
6. Recommender race: Race signalled by the randomly assigned name.

[34]Here we explicitly include race and gender as features. This practice may be uncommon in real-world settings. However, omitting race and gender variables does not preclude the algorithm from creating racial disparities if included features (e.g., education, location, preferences, language, etc.) covary with race, as is usually the case.

7. Simplified education: All users were coded into one of three education categories: Less than bachelor's, Bachelor's, and More than bachelor's.

8. Simplified user race: All users were coded into one of four race categories: Black, Hispanic, White, and Other. Any user who listed Hispanic as one of their races/ethnicities was categorized as Hispanic. Any user who listed Black as one of their races/ethnicities and did not list Hispanic, was categorized as Black. Any user who listed White as one of their races/ethnicities and did not list either Black or Hispanic, was categorized as White. Everyone else was categorized as Other.

9. User age: Age taken as given from participant.

10. User gender: Gender taken as given from participant.

These features are used to train a random forest to predict an indicator for whether a movie was chosen by a user. Two hyperparameters for the random forest (number of trees to grow and the number of variables to sample as candidates at each split) were tuned with cross validation. R defaults are used for all other hyperparameters: whether samples are drawn with replacement (they are); size of each sample (in our case, they are equal to the overall sample size); and the minimum size of terminal nodes for each tree (5 cases).

## B.3 Design and Data Collection for Experiment 3

Experiments 1 and 2 demonstrate that algorithms up-rank own-group content, particularly when training data comes from contexts where decision-making is done in a rushed context and so is more likely to be automatic rather than deliberate. We also carried out a third lab experiment, which provides at least suggestive evidence that the use of such an algorithm to curate choice options for subjects creates a "double penalty" against out-group content, particularly in settings of rushed decision-making.

For Experiment 3, we enrolled a total of (N=757) U.S. white male study subjects on the Prolific platform (additional demographics are presented in Table E.1). We replicated the rushed condition from Experiments 1 and 2, but now randomized subjects to two conditions:

- A *randomly ranked* condition in which subjects are shown candidate movie recommendations that are randomly ranked.

- An *algorithmically ranked* condition in which subjects are shown movie recommendations that are ranked on the screen using the algorithm that we built using data from the rushed condition for Experiment 2, which, as shown above, up-ranks own-group recommendations and down-ranks out-group recommendations.

That is, the data collection procedure for experiment 3 is nearly identical to that of experiments 1 and 2 but with two key differences:

1. **Ordering in the choice task**: Participants in the "random" condition are shown movies in a randomized order. Participants in the "algorithm" condition are shown movies according an algorithmic ranking. Specifically, the algorithm described in A.2

was applied to the movies shown to each user, which produces a probability that each recommendation will be selected. Movies were shown in descending order of their predicted probability of being selected.[35]

2. **Experimental instructions:** Participants in the "random" condition are instructed: "You will have 5 minutes to complete the task. This is not much time, so please choose quickly and do your best!" Participants in the "algorithm" condition are instructed: "You will have 5 minutes to complete the task. This is not much time, so please choose quickly and do your best! Note that these posts are algorithmically ranked by what we think you'd like most."

## B.4 Experiment 3 Results

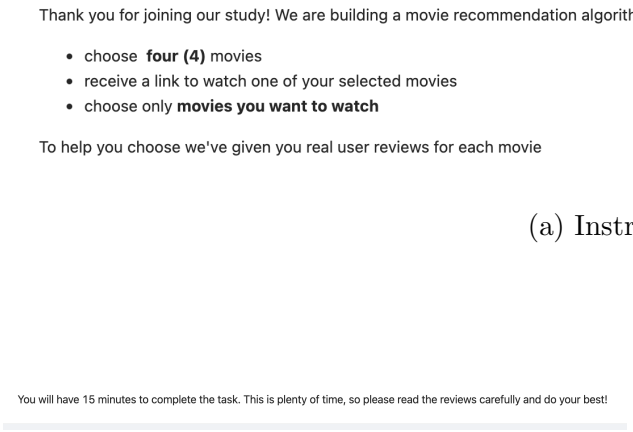
The left-hand panel of Appendix Figure 2 shows that random ranking, as expected, ranks own-group and out-group recommended content very similarly on average (the difference is −0.10 slots, standard error 0.27). In contrast, we can see on the right figure that the curation algorithm built using data from the "rushed" subjects in Lab experiment 2 substantially up-ranks posts from own-group recommenders versus out-group recommenders (the difference is 8.30 ranking slots, standard error 0.31). By way of comparison, this is about as large an impact on the algorithm's rankings as the movie genre being the study subject's favorite genre (equal to 8.43 ranking slots, standard error 0.22).

The panel at right shows that when choices are randomly ranked there is a 0.9 pp difference in favor of own-group content that is not statistically significant (standard error of 0.8 pp), which stems from the single penalty of implicit bias. But when we add the double penalty of the curation algorithm's tendency to down-ranking out-group content, the own-group vs. out-group difference in the chances subjects engage with the content increases to 3.6 pp (standard error 1.3 pp). See Appendix Table 2 Columns (5) and (6) for regression versions of the results shown in these figures.

While these data are from the lab setting, not real-world choices, the results are at least consistent with the idea that algorithms, especially in rushed decision-making contexts, can exacerbate the problem of people's own biases by reducing the chances people see out-group content—thereby creating a "double penalty" against out-group members.

[35] All participants were shown the same countdown on the left panel.

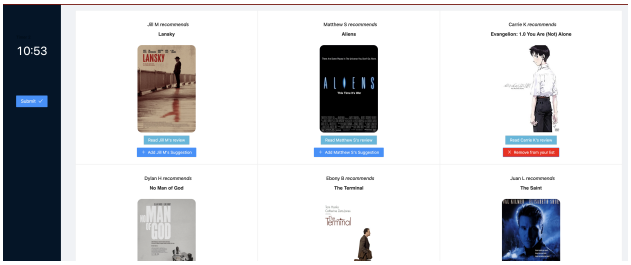
Figure B.1: Lab Experiment Screenshots



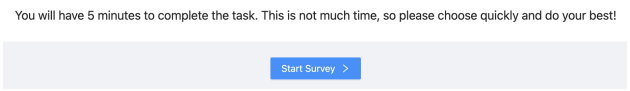

(a) Instructions



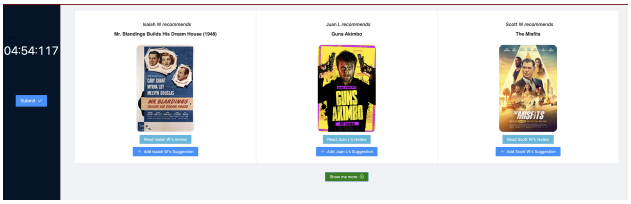

(b) Nonrushed Time Instructions




(c) Nonrushed Screen (After Selecting "Load More")




(d) Rushed Time Instructions




(e) Rushed Screen

**Note:** These are screenshots taken directly from the websites that participants used to complete the experiments.

# C Automaticity Survey Details

We recruited online subjects via Prolific (n=300) and asked a series of questions that capture various dimensions of automaticity. All subjects provided informed consent and the study was approved by the University of Chicago IRB (20-0131).

All subjects were given the following instructions:

"Thank you for participating in our research study. We'd like to know a little more about the decisions you make on Facebook. On the following screens, you'll be asked a series of questions comparing how you decide to react[36] to a post on your News Feed and how you decide to accept a friend recommended to you by the 'People You May Know' tab. We thank you in advance for answering thoughtfully and honestly."

Then subjects were asked 11 pairs of questions. For each question, subjects answered about both News Feed and People You May Know. We randomized both the order in which the pairs are shown and whether News Feed is the first in each pair.

The questions were as follows:

- **Access**: When you decide to (react to a post/add a friend from the 'People You May Know' recommendation) on Facebook, how easy would it be to explain why you decided to do so? [1–7; Very difficult–Not very difficult]
- **Awareness**: I have a good sense of all the things that affect my decision to (react to a post/add a friend from the 'People You May Know' recommendation). [1–7; Strongly disagree–Strongly agree]
- **Careful consideration**: Do you agree or disagree with the following statement: My decision to (react to a post/add a friend from the 'People You May Know' recommendation) is usually based on a careful consideration. (Binary; Agree/Disagree)
- **Cognitive effort**: When you're on Facebook, how much thought do you put into deciding to (react to a post/add a friend from the 'People You May Know' recommendation)? [1–7; I don't think about it–I put a great deal of thought into it]
- **Controllability**: I feel like I can alter or resist the immediate urge to (react to a post/add a friend from the 'People You May Know' recommendation) when I want to. [1–7; Strongly disagree–Strongly agree]
- **Efficiency**: I put a lot of mental effort into deciding to (react to a post/add a friend from the 'People You May Know' recommendation) [1–7; Strongly agree–Strongly disagree]
- **Free text**: When you're on Facebook, how do you decide whether to (react to a post/add a friend from the 'People You May Know' recommendation)? What goes through your mind? [Free response]

[36] We explain that "reactions" refer to the six animated emotions: Wow, Haha, Love, Sad, Angry, and the classic Like.

- **Gut reaction**: Do you agree or disagree with the following statement: My decision to (react to a post/add a friend from the 'People You May Know' recommendation) is usually based on a gut reaction. (Binary; Agree/Disagree)
- **Intentionality**: As soon as I see a (News Feed post/friend recommendation), I can't help but think about whether to (react on/accept) it or not. [-1-7; Strongly disagree–Strongly agree]
- **Outside influence**: When I'm on Facebook, my decision to (react to a post/add a friend from the 'People You May Know' recommendation) is often affected by what's happening around me at the time. [1–7; Strongly agree–Strongly disagree]
- **Speed**: How much time do you usually spend thinking about it before you (react to a post/add a friend from the 'People You May Know' recommendation)? (Please answer in seconds. For example, if you take two seconds, write 2; if you take half a second, write 0.5.) [numeric]

Then the respondents were asked a series of questions to record their demographic information.

# D Facebook Materials and Methods

We collected data from 662 subjects over six sequential waves between March, 2020 and October, 2020. Over four waves, we recruited 466 subjects through the CDR (US). In a single wave, we recruited 196 subjects through HDSL (US). In a single wave, we recruited 198 subjects in India through the The Centre for Social and Behaviour Change at Ashoka University, which recruited participants through colleges across India. All waves share the same basic structure in which 1) subject privately completes a self-assessment, 2) enumerator guides each subject through the News Feed (and for some subjects the PYMK or Recent Activity) while recording information about each post, and 3) enumerator guides subject through some additional data collection.

Data from each subject was collected in a single one-on-one Zoom session with an enumerator which lasted approximately one hour on average. After the data collection was completed, subjects were sent a link to access their payment of $20 ($10 in India).

## D.1 Wave Overview

- U.S. Data Collection:
  - Wave 1 - CDR, NF + PYMK, 242
  - Wave 2 - HDSL, NF + PYMK, 196
  - Wave 4 - CDR, NF + Recent Activity, 54
  - Wave 5 - CDR, NF (connectedness), 120
  - Wave 6 - CDR, NF (about) + Recent Activity, 50
- India Data Collection:
  - Wave 3 - India, NF + PYMK, 198

## D.2 Waves 1–3

Waves 1–3 were nearly identical, but each wave was on a different population. Because waves 1 and 2 were in the US, the group membership was based on perceived race, whereas wave 3 in India collected group membership based on perceived religion.

### D.2.1 Part I: Subject Categorization

After joining a Zoom call with an RA, subjects were asked to fill out a Qualtrics survey. In the survey, subjects were asked to describe their demographics and Facebook usage. As a main variable in our study, the assessment of the own-group is paramount. US subjects were shown the seven race and ethnicity categories used in the US Census and were given the option to check as many boxes as they like. Indian subjects were asked to report their religion.

While the subject filled out the survey, the enumerator made her best assessment of the subject's own-group (race in the US; religion in India), using up to two categories. Neither the subject nor the enumerator was aware of the assessment that the other has made. This protocol has the advantage of allowing us to observe how much alignment there is between how subjects self-identify and how they are perceived.

#### D.2.2 Part II: News Feed

Users opened their Facebook account and shared their screen with the enumerator. Then, scrolling sequentially through each post in the News Feed, the subject answered exactly one question about each post: "There are more posts than Facebook can possibly show you. How would you rate this post on a scale from 1-7 where 1 means 'can skip' and 7 means 'definitely want to see?'" In addition to recording the explicit preference, the enumerator assessed and recorded the perceived race of the poster of the content as well as some other details of the post, such as how long ago it was posted and whether it was posted to a group. The exact data being recorded by the enumerator were unknown to the subject. This continued for the first 60 non-sponsored posts.

#### D.2.3 Part III: People You May Know

Subjects then navigated to the Facebook recommender for new friends, entitled "People You May Know" (PYMK). The procedure for this section was similar to that in Part II. The subject scrolled down the list and for each recommended user the subject answered one question: "How familiar are you with this person on a scale from 1-7?" In addition to recording the familiarity, the enumerator assessed and recorded the perceived race of the recommended user as well as the number of mutual friends. This continued for 60 recommendations.

### D.3 Wave 4

Wave 4 differed slightly from the waves before it. Parts I and II were identical, but for part III, instead of scrolling through the PYMK recommendation, participants navigated to and scrolled through their 'recent activity' as follows.

#### D.3.1 Part I: Subject Categorization

Identical to waves 1–3.

#### D.3.2 Part II: News Feed

Identical to waves 1–3.

#### D.3.3 Part III: Recent Activity

Subjects then navigated to the Facebook activity log, which is sorted in reverse chronological order. The enumerator instructed the subject to scroll down until identifying the first post

with a reaction or comment. Then the enumerator recorded perceived race and gender for the identified poster. This process repeated for the 10 most recent comments/reactions to posts. As with Newsfeed data collection, if the race or gender of a user was not discernible from the post, the enumerator recorded the name in a separate list, and came back to the list after collecting all 10 posts.

## D.4 Wave 5

Wave 5 sought to collect richer data on the relationship between the subject and the author behind each News Feed post. There was no Part III in this wave.

### D.4.1 Part I: Subject Categorization

Identical to waves 1–4.

### D.4.2 Part II: News Feed

Users opened their Facebook accounts and shared their screens with the RA. Then, scrolling sequentially through each post in the News Feed, the subject answered exactly *three* questions about each post:

1. "There are more posts than Facebook can possibly show you. How would you rate this post on a scale from 1-7 where 1 means 'can skip' and 7 means 'definitely want to see?' "
2. How well do you know the person who posted this content? (1–7)
3. How do you know this person? [Family, Friend, Acquaintance, Don't know personally]

In addition to recording the explicit preference, the enumerator assessed and recorded the perceived race of the poster of the content as well as some other details of the post, such as how long ago it was posted and whether it was posted to a group. The exact data being recorded by the enumerator were unknown to the subject. This continued for the first 60 non-sponsored posts.

## D.5 Wave 6

Finally, Wave 6 sought to elicit subject perceptions on the purpose of the study.

### D.5.1 Part I: Subject Categorization

Identical to waves 1–5.

### D.5.2 Part II: News Feed

Identical to wave 5.

### D.5.3 Part III: Recent Activity

Almost identical to wave 4, collecting 30 recent activity items instead of 10.

### D.5.4 Part IV: Study Purpose

Enumerator asked the subject, "What do you think this study is about?" and transcribed the answer as close to verbatim as possible.

# E Additional Tables and Figures for Lab Experiments

Table E.1: Lab Experiment Participant Summary Statistics

| | Experiment 1 | | | Experiment 2 | | | Experiment 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| | All | Rushed | Non-Rushed | All | Rushed | Non-Rushed | All | Random | Algorithm |
| **Participant Characteristics** | | | | | | | | | |
| Male | 0.39 | 0.40 | 0.37 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Age | 29.55 | 30.08 | 28.96 | 38.83 | 38.83 | 38.95 | 33.61 | 33.06 | 34.20 |
| Black | 0.14 | 0.14 | 0.13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| White | 0.62 | 0.61 | 0.63 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| Hispanic | 0.22 | 0.24 | 0.21 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Race Other | 0.02 | 0.01 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Less than Bachelor's | 0.37 | 0.38 | 0.36 | 0.44 | 0.44 | 0.44 | 0.48 | 0.49 | 0.48 |
| Bachelor's | 0.31 | 0.30 | 0.33 | 0.37 | 0.37 | 0.37 | 0.37 | 0.37 | 0.37 |
| More than Bachelor's | 0.31 | 0.32 | 0.29 | 0.19 | 0.19 | 0.19 | 0.15 | 0.14 | 0.15 |
| **Outcomes** | | | | | | | | | |
| Number Movies Seen | 26.48 | 26.64 | 26.31 | 30.32 | 30.32 | 30.56 | 27.68 | 28.58 | 26.73 |
| Time Spent (min) | 4.71 | 3.84 | 5.67 | 4.37 | 4.37 | 5.22 | 3.84 | 3.84 | 3.85 |
| N Participants | 992 | 517 | 475 | 753 | 388 | 365 | 757 | 389 | 368 |

**Note:** This table reports summary statistics on participants in the three lab experiments. In each experiment, subjects needed to choose four movies from a set of recommendations. The recommendations were shown three at at time, and the participant could choose to "see more" multiple times, though most did not see all 42 available movies. Subjects in Lab Experiments 1 and 2 were randomized between making movie choices in a "rushed" or "non-rushed" condition, with the order of the movie recommendations being random. In Lab Experiment 3, subjects were randomized between seeing movies in a random order or an algorithmically determined order, but all made decisions in a "rushed" condition. Lab Experiment 1 used participants meant to be representative of the US population. Lab Experiments 2 and 3 focused on only white, male participants. See Appendix B for more details on the data collection. The only difference between rushed and non-rushed conditions that is statistically significant at the 10% level is Age in Experiment 1.

Table E.2: Lab Experiment Movie Summary Statistics

| | Experiment 1 | | | Experiment 2 | | | Experiment 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| | All | Rushed | Non-Rushed | All | Rushed | Non-Rushed | All | Random | Algorithm |
| *Panel A: All Movies Seen* | | | | | | | | | |
| Chosen | 0.19 | 0.19 | 0.19 | 0.13 | 0.13 | 0.13 | 0.14 | 0.14 | 0.15 |
| Own-group | 0.68 | 0.68 | 0.68 | 0.85 | 0.84 | 0.85 | 0.83 | 0.81 | 0.85 |
| **Recommender is:** | | | | | | | | | |
| Male | 0.67 | 0.67 | 0.67 | 0.66 | 0.66 | 0.66 | 0.51 | 0.49 | 0.54 |
| Black | 0.16 | 0.15 | 0.16 | 0.14 | 0.14 | 0.14 | 0.13 | 0.14 | 0.12 |
| White | 0.62 | 0.62 | 0.62 | 0.63 | 0.63 | 0.63 | 0.65 | 0.62 | 0.68 |
| Hispanic | 0.22 | 0.22 | 0.23 | 0.23 | 0.23 | 0.23 | 0.22 | 0.23 | 0.20 |
| Observations | 26268 | 13773 | 12495 | 22833 | 11679 | 11154 | 20953 | 11118 | 9835 |
| *Panel B: Amongst All Chosen Movies* | | | | | | | | | |
| Own-group | 0.70 | 0.71 | 0.69 | 0.85 | 0.85 | 0.85 | 0.85 | 0.82 | 0.89 |
| **Recommender is:** | | | | | | | | | |
| Male | 0.66 | 0.67 | 0.65 | 0.67 | 0.67 | 0.67 | 0.54 | 0.49 | 0.59 |
| Black | 0.15 | 0.15 | 0.15 | 0.14 | 0.14 | 0.13 | 0.12 | 0.14 | 0.10 |
| White | 0.62 | 0.63 | 0.61 | 0.64 | 0.64 | 0.65 | 0.68 | 0.63 | 0.74 |
| Hispanic | 0.23 | 0.22 | 0.24 | 0.22 | 0.22 | 0.22 | 0.20 | 0.23 | 0.17 |
| Observations | 4911 | 2554 | 2357 | 3001 | 1541 | 1460 | 3014 | 1551 | 1463 |

**Note:** This table reports summary statistics on outcomes in our three lab experiments. In all three experiments, movie recommendations were shown three at at time and the participant could choose to "see more" multiple times, though most did not see all 42 available movies. Panel A shows outcomes and recommender characteristics for the set of movies a participant saw on their screen (after clicking "see more" as many times as they liked). Panel B shows characteristics of the movies the participants *actually chose*. Subjects in Lab Experiments 1 and 2 were randomized between making movie choices in a "rushed" or "non-rushed" condition, with the order of the movie recommendations being random. In Lab Experiment 3, subjects were randomized between seeing movies in a random order or an algorithmically determined order, but all made decisions in a "rushed" condition. Lab Experiment 1 used participants meant to be representative of the US population. Lab Experiments 2 and 3 focused on only white, male participants. See Appendix B for more details on the data collection.

# F Additional Tables and Figures for US Facebook Data Collection

Figure F.1: Distributions of Raw Preference Ratings for News Feed Posts by Race

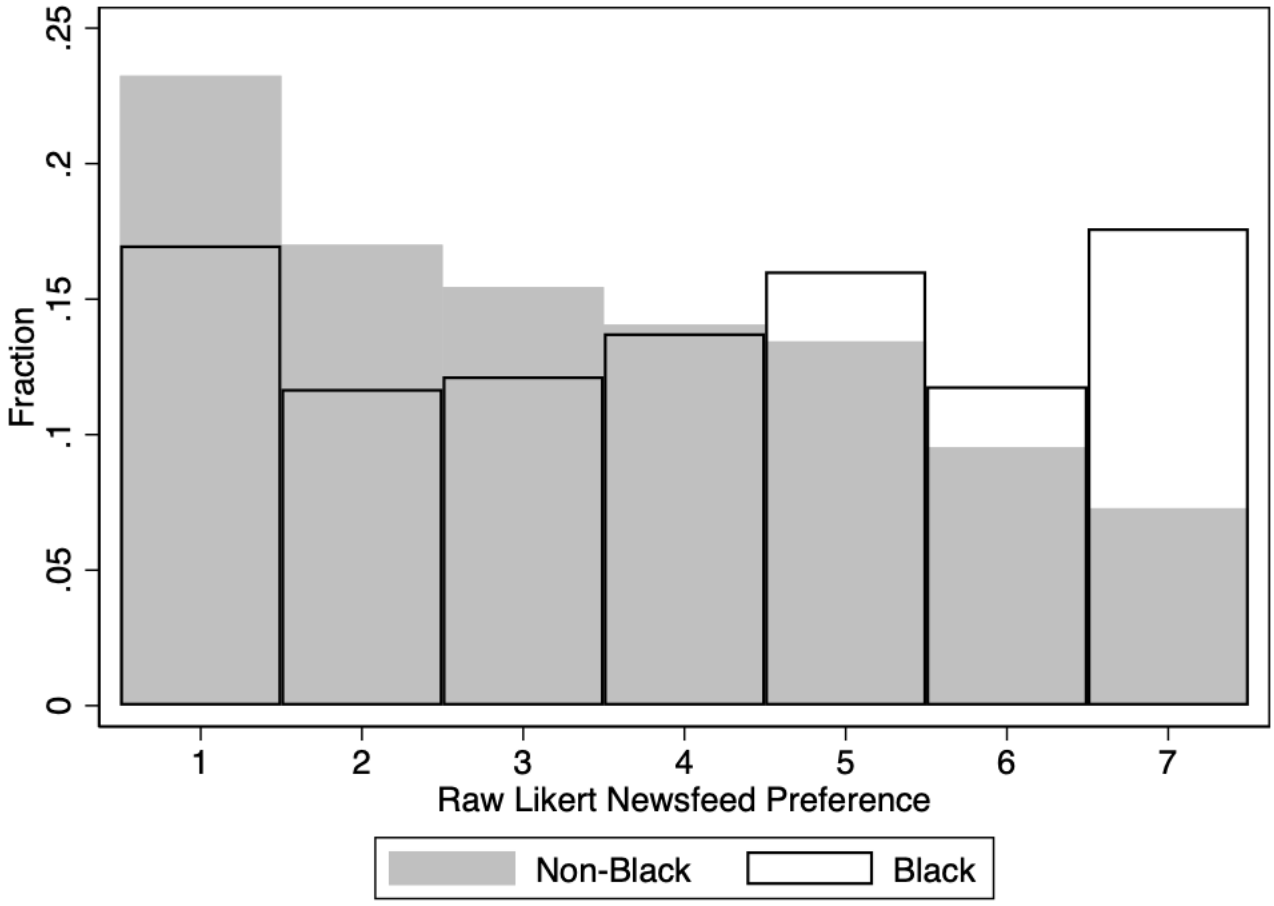


Figure F.2: CDF of Normalized Preferences by Own-group

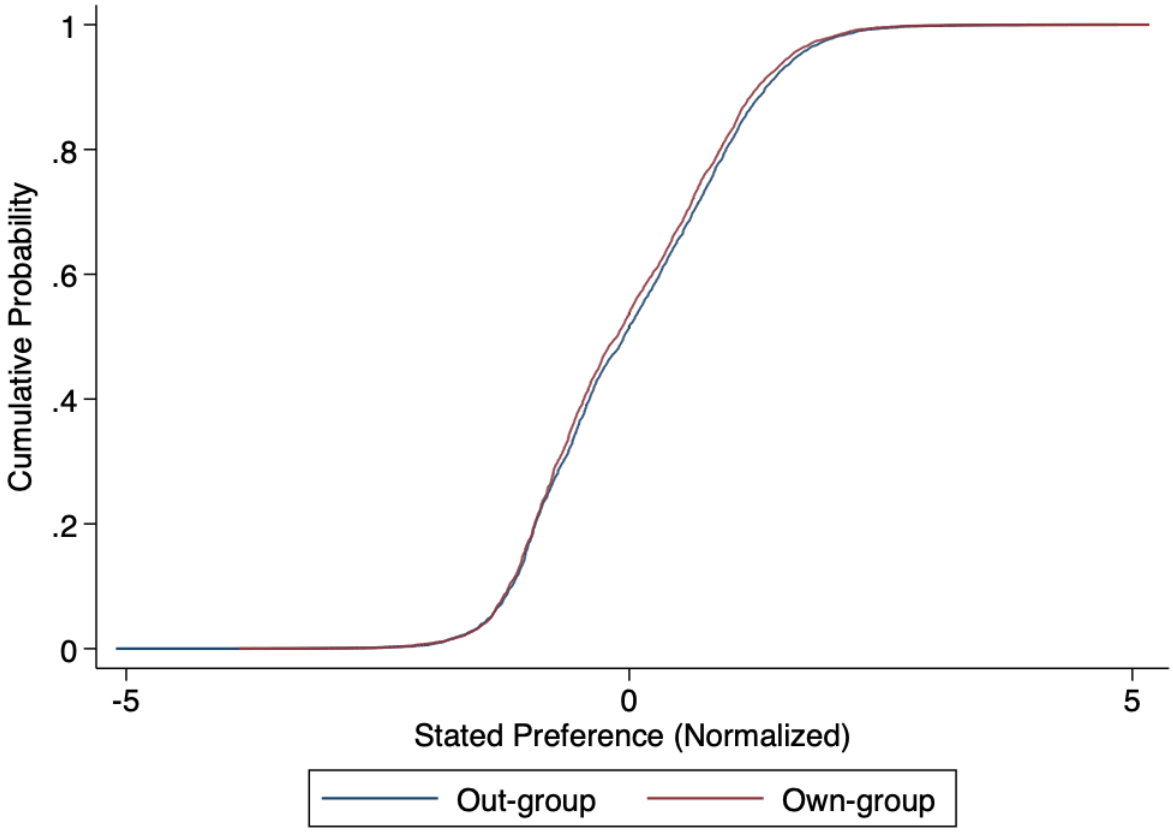
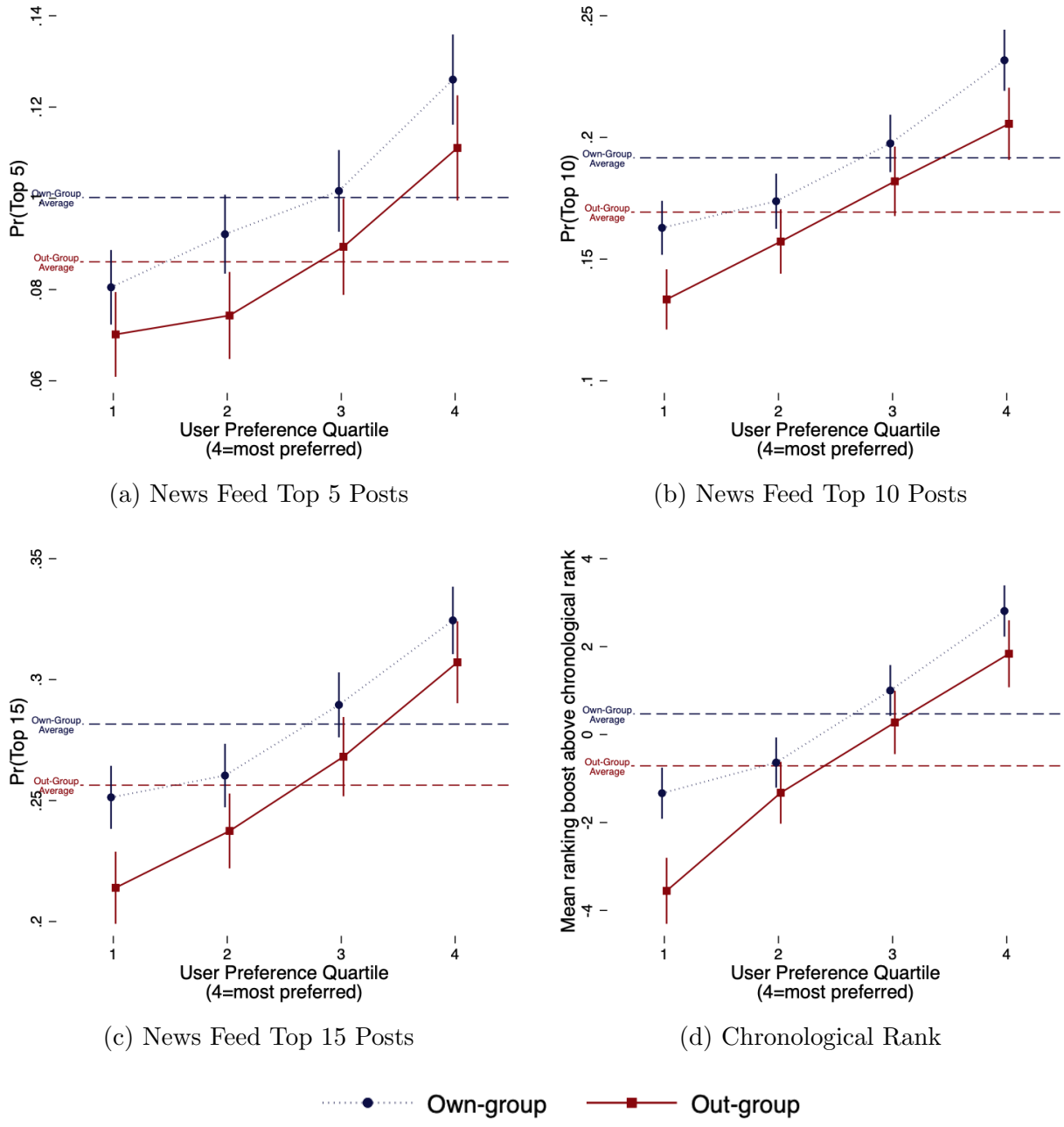


Figure F.3: Relationship between News Feed Top 5, 10, 15 and Chronological Ranking by Own-group Status Conditional on Subject Explicit Preference




(a) News Feed Top 5 Posts

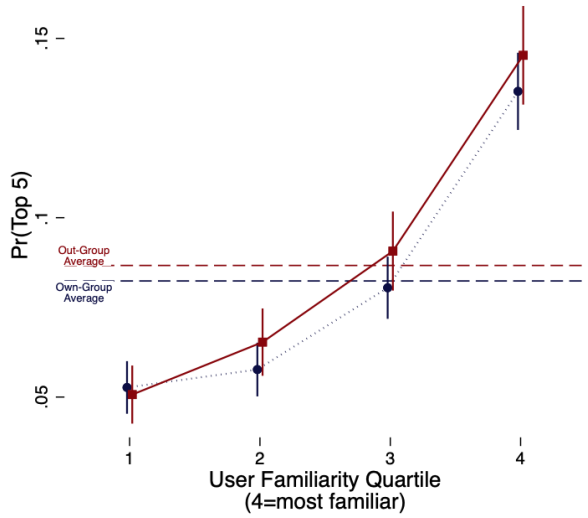
(b) News Feed Top 10 Posts

(c) News Feed Top 15 Posts

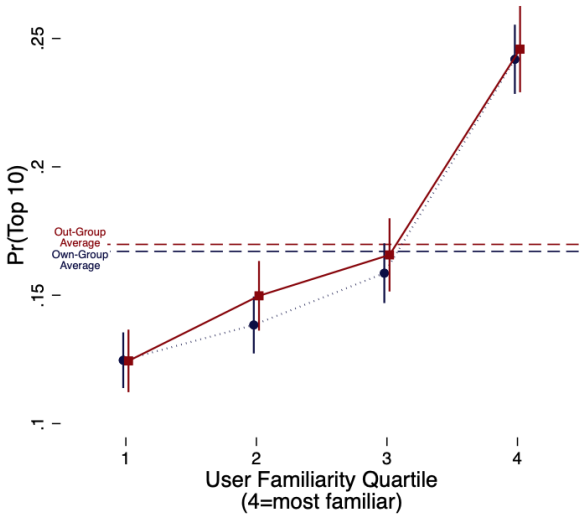
(d) Chronological Rank

**Note:** This figure mimics Figure 4 for News Feed; however, Panels (a)–(c) use Pr(Top X) where this is the probabiltiy the post is in the Top X of posts on the individuals News Feed. Chronological ranking imagines re-ranking News Feed posts in reverse chronological order such that the most recent post shows up on top and asks what is the mean ranking above chronological ranking for a post on the user's News Feed. The normalized subject-explicit preference/familiarity quartile is the across-subject quartile of within subject z-scores for stated preference for a post/familiarity with a suggested friend. Each subject's ratings were mean-centered and then divided by the subject's standard deviation of responses. The resulting distribution was then split into four equally sized bins. Own-group is defined as same race.

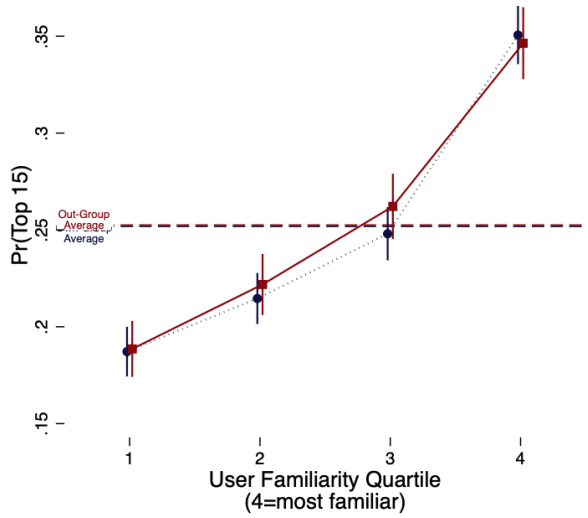
Figure F.4: Relationship between PYMK Top 5, 10, 15 and Mutual Friend Percent Rankingby Own-group Status Conditional on Subject Familiarity



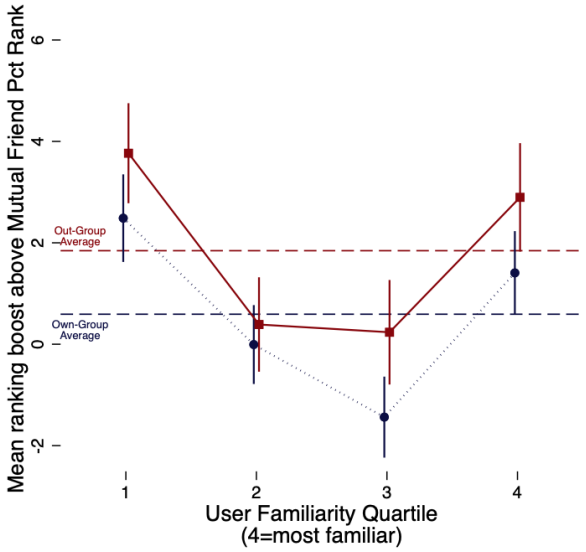

(a) Top 5 Friend Recommendations




(b) Top 10 Friend Recommendations



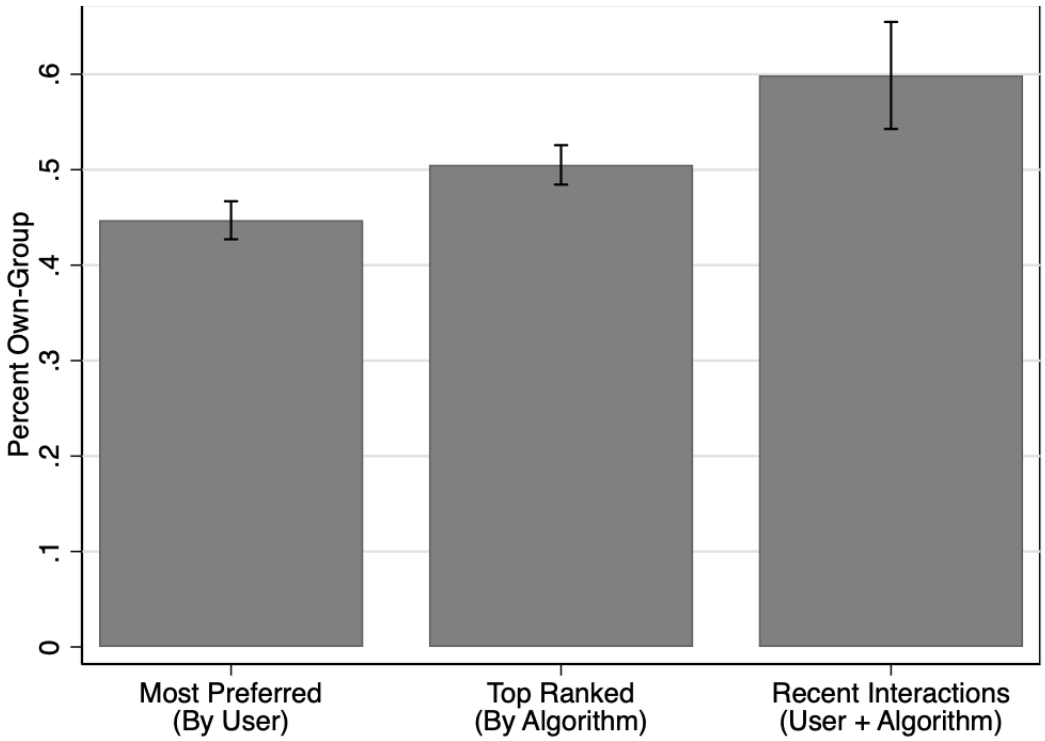

(c) Top 15 Friend Recommendations




(d) Vs Mutual Friend Rank

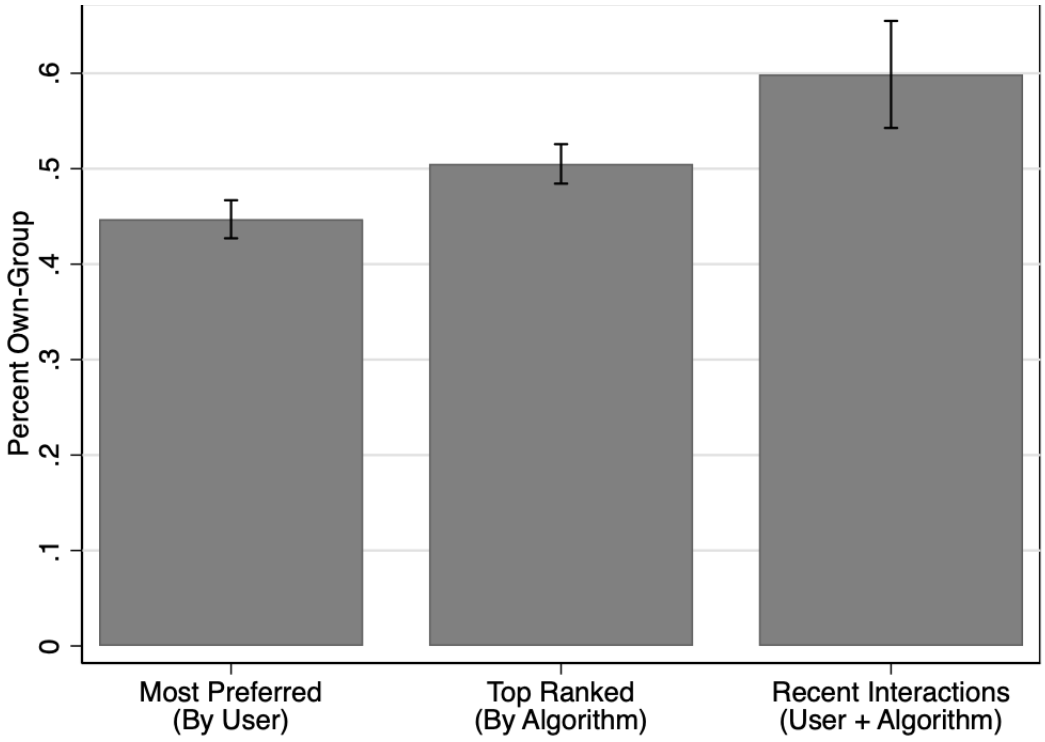

**Note:** This figure mimics Figure 4 for PYMK; however, Panels (a)–(c) use Pr(Top X) where this is the probability the post is in the Top X of recommendations as shown to the user in the PYMK section. Mutual friend percent rank imagines re-ranking PYMK by how many mutual friends you have in common, with those with the most mutual friends showing up first, and asks what the mean ranking above chronological ranking is for a post on the user's PYMK recommendations. The normalized subject-explicit preference/familiarity quartile is the across-subject quartile of within-subject z-scores for stated preference for a post/familiarity with a suggested friend. Each subject's ratings were mean-centered and then divided by the subject's standard deviation of responses. The resulting distribution was then split into four equally sized bins. Own-group is defined as same race.

Figure F.5: Share Own-Group by User Preference, Algorithmic Ranking, and Recent Interactions: Same Sample for all Bars




**Note:** This figure recreates Figure 5, but the "most preferred" analysis is restricted to only the top 10 most preferred based on a random ranking of tied posts. Recent interactions include the 10 most recent "likes," reactions, and/or comments. This figure shows the percent of those interactions that are on own-group posts (far-right bar). We also show the the percent own-group in the first 10 posts as ranked by the algorithm ("Top Ranked (By Algorithm)"). We also show the percent own-group for the posts that are most preferred by the user. Samples are the same as in Figure 5.

Figure F.6: Share Own-Group by User Preference, Algorithmic Ranking, and Recent Interactions: Only 10 Most Preferred




**Note:** This figure recreates Figure 5 but all bars are limited to the 102 individuals for whom we have recent activity data. Recent interactions include the 10 most recent "likes," reactions, and/or comments. This figure shows the percent of those interactions that are on own-group posts (far-right bar). We also show the the percent own-group in the first 10 posts as ranked by the algorithm ("Top Ranked (By Algorithm)"). And we also show the percent own-group for the posts that are most preferred by the user. To define most-preferred, we sorted respondents' posts by their raw stated preference; we then defined as most-preferred those posts whose preference rating was the same or larger than the post ranked 10 by this ranking. The N for each bar is 102 participants who were asked to show their recent activity (interactions) in addition to News Feed in Waves 4 and 5.

Table F.1: Facebook Regression Results (US): Raw Preferences

| | Newsfeed Posts | | PYMK Recommendations | |
|---|---|---|---|---|
| | (1) Slots Above Mean Ranking | (2) In Top 10 | (3) Slots Above Mean Ranking | (4) In Top 10 |
| Own-group (Race) | 1.099*** | 0.020*** | -0.172 | -0.004 |
| | (0.208) | (0.005) | (0.217) | (0.005) |
| Preference/Familiarity (Raw) | 0.745*** | 0.012*** | 1.353*** | 0.023*** |
| | (0.053) | (0.001) | (0.058) | (0.001) |
| Constant | -3.167*** | 0.129*** | -3.079*** | 0.116*** |
| | (0.235) | (0.005) | (0.212) | (0.005) |
| Observations | 28747 | 28747 | 25593 | 25593 |
| Dep Var Mean | 0 | 0.183 | 0 | 0.168 |

**Note:** These show regression versions of our main Facebook outcomes by own-group and user self-reported preference/familiarity. This table mimics Table 4 except we use the raw likert scale preference or familiarity statement from the user rather than normalized. Slots above mean ranking is rank-ordering of a movie/recommendation relative to the overall mean rank, such that a 1-unit increase implies 1 slot above the overall mean rank (closer to the top). In Top 10 is a binary variable indicating whether the post or recommendation was among the first 10 shown to the user.

Table F.2: Facebook Regression Results (US): Family

| | Newsfeed Posts | |
|---|---|---|
| | (1) Slots Above Mean Ranking | (2) In Top 10 |
| Own-group (Race) | 1.161*** | 0.026** |
| | (0.448) | (0.010) |
| Preference | 1.433*** | 0.030*** |
| | (0.228) | (0.005) |
| Family | 5.371* | 0.175* |
| | (3.237) | (0.090) |
| Own-group x Family | -2.984 | -0.132 |
| | (3.358) | (0.093) |
| Preference x Family | 1.805** | 0.059*** |
| | (0.805) | (0.022) |
| Constant | 0.897** | 0.187*** |
| | (0.350) | (0.008) |
| Observations | 6525 | 6525 |
| Dep Var Mean | 0 | 0.183 |

**Note:** These regressions mimic columns (1) and (2) from Table 4, but restricted to the sample of individuals in Wave 5 whom we asked how they knew the poster ("Family," "Acquaintance," "Friend," "Don't know Personally"). Here, Family is a binary indicator for whether the person indicated the poster was family. Slots Above Mean Ranking is rank-ordering of a movie/recommendation relative to the overall mean rank, such that a 1-unit increase implies 1 slot above the overall the mean rank (closer to the top). In Top 10 is a binary variable indicating whether the post or recommendation was among the first 10 shown to the user. User preference/familiarity is normalized within user.

Table F.3: Facebook Regression Results (US): Reweighted to Match US Demographics

| | Newsfeed Posts | | PYMK Recommendations | |
|---|---|---|---|---|
| | (1) Slots Above Mean Ranking | (2) In Top 10 | (3) Slots Above Mean Ranking | (4) In Top 10 |
| *Panel A: Main* | | | | |
| Own-group (Race) | 1.192*** | 0.022*** | -0.236 | -0.006 |
| | (0.208) | (0.005) | (0.217) | (0.005) |
| *Panel: Reweighted to match U.S. Demographics* | | | | |
| Own-group (Race) | 1.193*** | 0.032*** | 0.354 | 0.012 |
| | (0.415) | (0.009) | (0.445) | (0.010) |
| Observations | 28747 | 28747 | 25413 | 25413 |

**Note:** The first panel repeats the regressions from Table 4 showing only the main coefficient of interest on "own-group." The second panel reweights our sample to match the demographics of the US population on gender, race, age, and education. The weights were calculated through an iterative proportional fitting procedure, also known as "raking." This was performed using `ipfweight` in Stata. The procedure was done with a maximum of 200 iterations. No weight trimming was required. When matching to US demographics, the demographics matched were: proportion female, Black, white, Hispanic (non-white and white), Asian, other race, age 18–25, age 26–34, age 35–54, age over 65, less than high school, high school, some college, college, and more than college; US demographics were taken from Census and CPS surveys.

# G Additional Tables and Figures for India Facebook Data Collection

Table G.1: India Facebook Subject Summary Statistics

| | (1) Newsfeed | (2) PYMK |
|---|---|---|
| Religion of subject (RA) | | |
| Hindu | 0.645 | 0.640 |
| Muslim | 0.301 | 0.308 |
| Other Religion | 0.044 | 0.041 |
| Religion of Subject (Self Identification) | | |
| Hindu | 0.634 | 0.628 |
| Muslim | 0.301 | 0.308 |
| Other Religion | 0.066 | 0.064 |
| Gender | | |
| Male | 0.607 | 0.610 |
| Female | 0.377 | 0.372 |
| Non-binary | 0.016 | 0.017 |
| Age | 22.579 | 22.738 |
| Less than Bachelor's Degree | 0.514 | 0.506 |
| Bachelor's Degree | 0.322 | 0.326 |
| Graduate Degree | 0.158 | 0.163 |
| Educ. Unknown | 0.060 | 0.047 |
| Average Facebook Usage | | |
| Hourly | 0.186 | 0.198 |
| Daily | 0.393 | 0.390 |
| Weekly | 0.251 | 0.256 |
| Monthly | 0.120 | 0.110 |
| Yearly | 0.016 | 0.017 |
| Never | 0.033 | 0.029 |
| Within past hour | 0.454 | 0.459 |
| Last Facebook Log-in | | |
| Within past day | 0.284 | 0.285 |
| Within past week | 0.164 | 0.163 |
| Within past month | 0.060 | 0.052 |
| Within past year | 0.038 | 0.041 |
| Total Facebook Friends | 789.891 | 777.523 |
| Mean NF Post Preference | 4.069 | 4.069 |
| Standard Dev of NF Post Preference | 1.728 | 1.724 |
| Mean PYMK Rec Familiarity | 3.033 | 3.033 |
| Standard Dev of PYMK Rec Familiarity | 2.060 | 2.060 |
| Observations | 183 | 172 |

**Note:** This table presents summary statistics on India data collection participants. There was only one wave of data collection.

Table G.2: Summary Statistics on Collected Facebook Outcomes (India)

| | Newsfeed Posts | | | PYMK Recommendations | | |
|---|---|---|---|---|---|---|
| | (1) All | (2) Ingroup | (3) Outgroup | (4) All | (5) Ingroup | (6) Outgroup |
| Own-group (Religion) | 0.603 | | | 0.668 | | |
| Religion of Poster/Rec | | | | | | |
| Muslim | 0.182 | | | 0.173 | | |
| Hindu | 0.600 | | | 0.693 | | |
| Other Religion | 0.089 | | | 0.076 | | |
| Pref./Familiarity Ranking | | | | | | |
| 1st Quartile | 0.249 | 0.245 | 0.256 | 0.249 | 0.249 | 0.251 |
| 2nd Quartile | 0.250 | 0.259 | 0.236 | 0.251 | 0.232 | 0.287 |
| 3rd Quartile | 0.251 | 0.249 | 0.254 | 0.250 | 0.259 | 0.231 |
| 4th Quartile | 0.250 | 0.247 | 0.255 | 0.250 | 0.260 | 0.232 |
| Ranking | 28.307 | 27.864 | 28.981 | 29.164 | 29.054 | 29.386 |
| In First 10 Posts/Recs | 0.200 | 0.203 | 0.195 | 0.187 | 0.189 | 0.183 |
| Post to Group | 0.218 | | | | | |
| Days Ago Posted | 1.849 | 1.853 | 1.843 | | | |
| Mutual Friends | | | | 17.818 | 19.094 | 15.250 |
| Observations | 7864 | 4745 | 3119 | 10882 | 7271 | 3611 |

**Note:** This table shows summary statistics on News Feed Posts and friend recommendations from the PYMK algorithm for our India Facebook data collection. News Feed posts are restricted to posts from "humans" and not from companies/firms. The normalized subject-explicit preference/familiarity quartile is the across-subject quartile of within subject z-scores for stated preference for a post/familiarity with a suggested friend.

Table G.3: Facebook Regression Results (India)

| | Newsfeed Posts | | PYMK Recommendations | |
|---|---|---|---|---|
| | (1) Slots Above Mean Ranking | (2) In Top 10 | (3) Slots Above Mean Ranking | (4) In Top 10 |
| Own-group (Religion) | 1.110*** | 0.007 | 0.152 | 0.002 |
| | (0.398) | (0.009) | (0.347) | (0.008) |
| Preference/Familiarity | 1.421*** | 0.029*** | 2.430*** | 0.048*** |
| | (0.196) | (0.005) | (0.166) | (0.004) |
| Constant | -0.670** | 0.195*** | -0.107 | 0.186*** |
| | (0.312) | (0.007) | (0.282) | (0.006) |
| Observations | 7864 | 7864 | 10880 | 10880 |

**Note:** These show regression versions of our main Facebook outcomes by own-group and user self-reported preference/familiarity. Slots Above Mean Ranking is rank-ordering of a movie/recommendation relative to the overall mean rank, such that a 1-unit increase implies 1 slot above the overall the mean rank (closer to the top). In Top 10 is a binary variable indicating whether the post or recommendation was among the first 10 shown to the user. User preference/familiarity is normalized within user.

Table G.4: Facebook Regression Results (India): Reweighted to Match India Demographics

| | Newsfeed Posts | | PYMK Recommendations | |
|---|---|---|---|---|
| | (1) Slots Above Mean Ranking | (2) In Top 10 | (3) Slots Above Mean Ranking | (4) In Top 10 |
| *Panel A: Main* | | | | |
| Own-group (Religion) | 1.110*** | 0.007 | 0.152 | 0.002 |
| | (0.398) | (0.009) | (0.347) | (0.008) |
| *Panel: Reweighted to match India Demographics* | | | | |
| Own-group (Religion) | 1.155*** | 0.003 | 0.081 | 0.006 |
| | (0.448) | (0.010) | (0.390) | (0.009) |
| Observations | 7352 | 7352 | 10044 | 10044 |

**Note:** The first panel repeats the regressions from Table G.3, showing only the main coefficient of interest on "own-group." The second panel reweights our sample to match the demographics of the India population on gender and religion. The weights were calculated through an iterative proportional fitting procedure, also known as "raking." This was performed using `ipfweight` in Stata. The procedure was done with a maximum of 200 iterations. No weight trimming was required. When matching to India demographics, the demographics matched were: proportion female, Hindu, and Muslim. Proportion female came from World Bank Data and proportion by religion came from Pew Research. We did not match on age or education as reliable statistics on those demographic breakdowns were not easily available and our sample size was smaller, making re-weighting on many demographics more difficult.

# H Cross-Platform Automaticity Survey Details

On December 10, 2022, we recruited a nationally representative sample of online subjects via Prolific (n=600) and asked a series of questions that captured two measures of automaticity for a variety of online platforms and behaviors. All subjects provided informed consent and the study was approved by the University of Chicago IRB (20-0131). 576 of those participants passed the attention checks and went on to answer the following protocol.

All subjects were given the following instructions:

"Thank you for participating in our research study. We'd like to know a little more about your online choices. On the following screens, you'll be asked a series of questions comparing how you decide to interact with various online platforms. We thank you in advance for answering thoughtfully and honestly."

Then subjects were asked which of the following platforms they have used in the past month. Next to each platform we include here how many participants selected the option.

- Amazon (534; 93%)
- Facebook (425; 74%)
- Instagram (346; 60%)
- Netflix (371; 64%)
- Spotify (227; 39%)
- TikTok (216; 38%)
- Twitter (323; 56%)

We then ask a series of questions based on the platforms used by the subject. For example, a subject who says she has used Facebook in the last month will see all of the Facebook questions; if she does not say she has used Spotify in the last month, then she will see none of those questions. For each platform, subjects answered two questions about a set of behaviors relevant to that platform:

(1) "My decision to {*behavior*} is usually based on a careful consideration" [Binary; Do you agree or disagree?], and

(2) "How much time do you usually spend thinking about it before you take each of the following actions? (Please answer in seconds. For example, if you take two seconds, write 2; if you take half a second, write 0.5.)".

The order of the platforms, the order of the questions within platforms, and the order of behaviors within a question were randomized.

The behaviors were as follows:

- **Instagram**: scroll past a post, watch a video on my explore page, like a post on my timeline, watch the beginning of a story, scroll over to see more photos, read all the comments under a post, comment on a post, share a post with a friend, follow user (add them to your timeline)
- **Facebook**: scroll past a post, watch an entire video, like/react to a post on my timeline, watch the beginning of a video, scroll over to see more photos, read all the comments under a post, comment on a post, share a post with a friend, add a friend
- **TikTok**: scroll past a tiktok, watch an entire tiktok, like a tiktok, watch the beginning of a tiktok, save a tiktok, read all the comments under a tiktok, comment on a tiktok, share a tiktok with a friend, follow someone on tiktok (add them to your feed)
- **Twitter**: scroll past a tweet, watch an entire video, like a tweet, watch the beginning of a video, quote a tweet, read all the comments under a tweet, comment on a tweet, share a tweet with a friend, follow someone on twitter (add them to your feed), retweet a tweet, click on and read a thread
- **Amazon**: scroll past an item, hover on an item, click on an item to see more details, read the first review for an item, read all the reviews for an item, add an item to your basket, purchase an item
- **Spotify**: listen to the beginning of a song, listen to an entire song, add a song to a playlist, listen to a recommended playlist
- **Netflix**: watch the trailer for a movie, watch an entire movie, hover over a recommendation, add a movie to your watchlist

Then the respondents were asked a series of questions to record their demographic information.