Received: 28 July 2022 | Accepted: 30 March 2023

DOI: 10.1002/jcpy.1351

RESEARCH REPORT

Algorithms propagate gender bias in the marketplace—with consumers' cooperation

Shelly Rathee[1] | **Sachin Banker**[2] | **Arul Mishra**[2] | **Himanshu Mishra**[2]

[1]Department of Marketing & Business Law, Villanova School of Business, Villanova, Pennsylvania, USA

[2]David Eccles School of Business, University of Utah, Salt Lake City, Utah, USA

Correspondence
Shelly Rathee, Department of Marketing & Business Law, Villanova School of Business, 3044 Bartley Hall, Villanova, PA, USA.
Email: shelly.rathee@villanova.edu

Abstract

Recent research shows that algorithms learn societal biases from large text corpora. We examine the marketplace-relevant consequences of such bias for consumers. Based on billions of documents from online text corpora, we first demonstrate that from gender biases embedded in language, algorithms learn to associate women with more negative consumer psychographic attributes than men (e.g., associating women more closely with *impulsive* vs. *planned* investors). Second, in a series of field experiments, we show that such learning results in the delivery of gender-biased digital advertisements and product recommendations. Specifically, across multiple platforms, products, and attributes, we find that digital advertisements containing negative psychographic attributes (e.g., impulsive) are more likely to be delivered to women compared to men, and that search engine product recommendations are similarly biased, which influences consumer's consideration sets and choice. Finally, we empirically examine consumer's role in co-producing algorithmic gender bias in the marketplace and observe that consumers reinforce these biases by accepting gender stereotypes (i.e., clicking on biased ads). We conclude by discussing theoretical and practical implications.

KEYWORDS

customer segmentation, digital advertising, gender bias, natural language processing, word embedding

INTRODUCTION

It was not until the latter half of the twentieth century that women in the United States were guaranteed equal access to holding bank accounts and credit cards in their own names, applying for mortgages and personal loans, and even obtaining a college education. Internationally, women still have on average only three-fourths of the legal rights afforded to men (World Bank, 2020). Persistent gender bias in societal beliefs has been widely documented in social psychological research, such that women are consistently associated with less positive and more negative attributes than men (Allport, 1954; Devine, 1989; Ellemers, 2018; Greenwald et al., 1998). Today, as algorithms guide many marketing decisions for firms—from product recommendations, customer segmentation, ad targeting, and new product development—the biases that algorithms learn from human language could inadvertently reintroduce prejudice into the marketplace. Recent research in computer science has demonstrated that the large text corpora used by algorithms to gain insights into human thoughts, opinions, and preferences can indeed reflect gender biases (Caliskan et al., 2017; Charlesworth et al., 2021; Garg et al., 2018). Building on this work, we examine whether (1) algorithms learn gender bias for psychographic attributes that influence marketplace outcomes, (2) whether such algorithmic biases learned from language have any downstream influence on consumers in the digital marketplace, and (3) whether consumers accelerate or impede the algorithmic propagation of gender bias through their online interactions with them.

Because business applications of language models are widespread, the algorithmic learning of societal biases embedded in language can be propagated through a variety of consumer-facing processes. Word embedding

Accepted by Grant Packard, Sarah Moore and Jonah Berger, Editors.

algorithms are routinely used to mine large text corpora for consumer preferences and thoughts because these corpora are a repository of millions of human thoughts over years. Algorithms learn consumer preferences as rules, which are then used to offer customized products on multiple digital market applications including ad-targeting and recommendation systems (Liu et al., 2015; Ozsoy, 2016; Zheng et al., 2017). The learnings of algorithms are incorporated into automated filters, providing customized offerings to users (Caselles-Dupré et al., 2018; Liu et al., 2015; O'Neil, 2016; Ozsoy, 2016; Zheng et al., 2017). If, as past research shows, algorithms learn gender bias, then the bias will be considered by the algorithm to be a rule that can be used to customize offerings to customers—in other words, biasing the items that enter consumer consideration sets in the first place.

While gender-based preferences within certain product categories (e.g., clothing, health and beauty care) are well characterized, we focus our analysis on more insidious forms of gender bias where there is no theoretical basis for gender-based differences. Specifically, extensive research in marketing offers insights into how psychographic attributes are associated with preferences and behaviors (Ailawadi et al., 2001; Raju, 1980; Steenkamp & Maydeu-Olivares, 2015), providing advertisers with dimensions that are leveraged in marketing materials to reach appropriate customer segments. On most psychographic dimensions commonly used to segment consumers, research holds that men and women are more alike than they are different (Epstein, 1988; Hyde & Plant, 1995; Kimball et al., 1995). Meta-analytical evidence supports this gender-similarities hypothesis, indicating that men and women are similar on measures of most psychological variables, with 95% of reported differences being near zero or small (Hyde, 2016; Zell et al., 2015). These psychographic attributes are not uniquely associated with a gender and do not hold differing base rates; that is, the current study extends prior findings documenting gender bias in the marketplace that can be explained by differences in base rates (e.g., women associated with the profession of nurse because 86% of nurses identify as women; Bolukbasi et al., 2016). Instead, we examine marketplace-relevant psychographic attributes in which there is no basis for gender differences due to base rates. Thus, from a theoretical perspective, firms would have no reason to expect ads targeting *impulsive* investors to be delivered to a greater share of women compared to those targeting *planned* investors—gender biases that may not be empirically true in the marketplace and which algorithms may be propagating because of information learned from large text corpora. Therefore, we build on past research in several ways by focusing on marketplace-relevant psychographic attributes to examine how gender biases learned by algorithms from large text corpora can have consequences for consumers in the digital marketplace.

To examine whether algorithms learn gender-biased psychographic associations, we first applied natural language processing (NLP) methods known as word embeddings. Recent findings indicate that large text corpora reflect implicit societal beliefs. Algorithms learn to associate men with positive words such as *love, cheer, peace*, while women are instead associated with negative words such as *murder, filth, evil* (Boghrati & Berger, 2020; Caliskan et al., 2017; Charlesworth et al., 2021; Crawford, 2017; DeFranza et al., 2020; Garg et al., 2018; May et al., 2019). Other work suggests algorithms learn to associate men and women with different stereotypical job roles, connecting women with the words *homemaker*, *nurse*, *receptionist* and men with *maestro*, *skipper*, *protégé* (Bolukbasi et al., 2016). Building on this work, in Study 1 we examine whether algorithms learn gender-biased associations even for marketplace-relevant psychographic attributes that have no basis to be uniquely associated with a specific gender (e.g., base rates would not suggest that women are less *rational* or *loyal* than men).

While prior research has documented biases that algorithms learn from text corpora, very little work has examined downstream consequences for consumers. Therefore, we next aimed to understand how algorithmic learning of gender bias can materially affect consumer choices. Gender-biased associations learned by algorithms could, for example, bias delivery of services to men and women when unsuspecting advertisers target "*irresponsible* investors" versus "*disciplined* investors," even when such psychographic attributes are not uniquely associated with a gender. We empirically document a gender-biased pattern of ad delivery in a series of field experiments across platforms, product categories, and marketplace-relevant psychographic attributes (Appendices S7–S9), and we further illustrate in Study 2 that these algorithmic gender biases can have material consequences for consumer consideration-set formation and choice. Based on these findings, we present a debiasing strategy that advertisers can apply to check for and reduce gender-biased delivery prior to the launch of their offering (Appendix S12).

Furthermore, we gain additional insight into the role that consumers themselves play in co-producing gender bias online. Because of certain ad-targeting rules algorithms adopt based on user interactions, consumer "acceptance" or "rejection" of implicit gender stereotypes could further amplify or attenuate existing algorithmic biases. Algorithmic biases in ad delivery are reflected when product offerings are delivered to consumers in a gender-biased manner: men and women encounter gender bias through restricted choice sets (with women more likely to receive ads targeting "*irresponsible* investors" vs. those targeting "*disciplined* investors"). However, consumer responses to algorithmically biased offerings can influence the dynamics of gender bias in the marketplace. On one hand, there are reasons to believe that consumers will accept the gender stereotypes perpetuated by algorithmic learning from language. Prior research suggests that more moderate gender stereotypes exhibit assimilation, whereas more extreme stereotypes exhibit contrast (Kray et al., 2001; Manis et al., 1988). Thus, for marketplace-relevant psychographic attributes, relatively subtle gender-biased associations could lead consumers to interpret offerings as familiar, fluent, and identity-consistent rather than overtly stereotypical in nature (Reed et al., 2012; Susser et al., 2016), facilitating "acceptance" and amplification of gender bias. On the other hand, if algorithmic bias in ad targeting and recommendations results in presenting consumers with more extreme and overtly gender-stereotyped offerings, this could yield contrastive effects. As a consequence, consumers may instead exhibit reactance (Brehm, 1966) and respond by "rejecting" the gender stereotypes, thereby attenuating the propagation of gender bias. Evaluating these possibilities in Study 3, we find that consumers tend to accept gender stereotypes to co-produce algorithmic gender biases in the marketplace.

STUDY 1: GENDER-BIASED PSYCHOGRAPHIC ASSOCIATIONS LEARNED FROM LARGE TEXT CORPORA

We first examined whether algorithms learn gender-biased customer psychographic associations from Common Crawl, a large text corpus consisting of billions of webpages. This study extends prior findings documenting gender bias in the marketplace that can be explained by differences in base rates (e.g., women associated with *homemaker*, *nurse*, *receptionist*; Bolukbasi et al., 2016) by focusing on marketplace-relevant psychographic attributes in which there is no basis for gender differences due to base rates (e.g., *honest*, *reasonable*, *hedonistic*).

Methods

To examine whether algorithms link women and men in a gender-biased manner to psychographic attributes used in customer segmentation and ad targeting, we compiled a list of 59 customer psychographic attributes studied in the marketing literature (see Table 1). These attributes were drawn from prior research (e.g., Aaker, 1997; Anderson, 1968; Berry & McArthur, 1985; Briggs, 1992; Eysenck, 1982; Hofstee et al., 1992; Roberts et al., 2005; Steenkamp & Maydeu-Olivares, 2015); please see Appendix S2 for further details. We categorized these attribute words into positive (desirable) and negative (undesirable) groups based on Garg et al. (2018) and supported by a pretest (Appendix S3).

We test for gender bias by comparing the similarity of target words (e.g., *he, she, her, him*; see Table 2 for full list) with positive and negative customer psychographic attributes (e.g., *innovative*, *planned*, *conformist*, *impulsive*). To do so, we apply GloVe word embeddings (Pennington et al., 2014) pretrained on the Common Crawl text corpus. These word embeddings provide a 200-dimensional vector for each word, where the cosine distance between vectors captures the semantic similarity/dissimilarity that the algorithm learned from billions of webpages. We can compute gender bias as the net similarity of female (or male) word vectors to positive- and negative-attribute word vectors (Caliskan et al., 2017; Garg et al., 2018), summarized in Equation 1 below. We apply a nonparametric permutation test from Caliskan et al. (2017) to evaluate the significance of the gender bias measured (Appendix S1 presents technical details).

Net gender bias:

$$\begin{aligned}\text{Bias} = &\left[\text{Similarity (male words to positive attributes)} - \text{Similarity (male words to negative attributes)}\right] \\ &- \left[\text{Similarity (female words to positive attributes)} - \text{Similarity (female words to negative attributes)}\right].\end{aligned} \quad (1)$$

Results

Our analysis indicated that algorithms learning from the large text corpus in Common Crawl form gender-biased associations of men and women with psychographic attributes, as supported by nonparametric permutation tests (d=1.057, p=0.006). Specifically, the algorithm

TABLE 1 Dictionary of attributes.

Positive attribute words	Honest, reasonable, independent, thorough, dependable, rational, relaxed, loyal, reliable, disciplined, patient, creative, innovative, planned, resolute, resistant, industrious, certain, determined, wise, tough, jolly, civilized, strong, enterprising, quick, logical, original, methodical, kind
Negative attribute words	Unfriendly, unkind, rigid, moody, intolerant, hedonistic, tempted, fragile, indulgent, irresponsible, instinctive, dissatisfied, conformist, impulsive, fickle, unreliable, emotional, vain, lazy, submissive, risky, irritable, frivolous, inhibited, sensitive, vindictive, complicated, changeable, sarcastic

TABLE 2 Dictionary of gender target words.

Female target words	She, hers, her, woman, female, herself, women, females, gal, girl
Male target words	He, his, him, man, male, himself, men, males, guy, boy

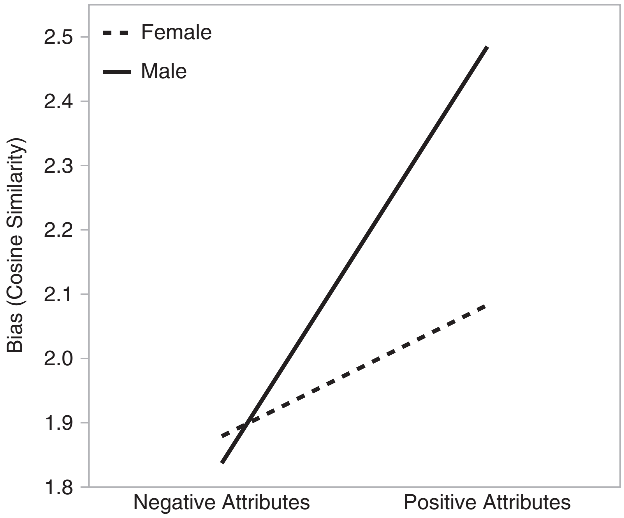


FIGURE 1 Gender-biased psychographic associations in common crawl corpus.

TABLE 3 Top attributes differentially associated with women versus men.

Associated with women vs. men	Associated with men vs. women
1. Submissive (0.077)	1. Wise (−0.080)
2. Inhibited (0.052)	2. Tough (−0.072)
3. Fragile (0.050)	3. Patience (−0.067)
4. Sensitive (0.038)	4. Certain (−0.061)
5. Frivolous (0.037)	5. Quick (−0.060)
6. Changeable (0.034)	6. Methodical (−0.059)
7. Irritable (0.023)	7. Kind (−0.058)
8. Indulgent (0.018)	8. Original (−0.053)
9. Emotional (0.018)	9. Loyal (−0.052)
10. Relaxed (0.013)	10. Logical (−0.052)

represented men as having significantly greater semantic similarity with positive psychographic attributes (2.485) than women did (2.083; $d=0.763$, $p<0.001$). Also, the algorithm represented women as having significantly greater semantic similarity with negative psychographic attributes (1.879) than men did (1.837; $d=0.154$, $p<0.001$; see Figure 1). In Table 3, we tabulated the top attributes differentially associated with women and men based on the magnitude of the bias for each attribute word.

These findings are remarkably robust. We find similar evidence of gender bias across other text corpora (see Appendix S5 for results from Amazon reviews), and even after expanding the attribute lists and taking thousands of independent random subsamples (reported in Appendix S6).

DOES ALGORITHMIC GENDER BIAS LEARNED FROM LANGUAGE HAVE CONSEQUENCES FOR CONSUMERS?

Study 1 demonstrated that algorithms detect and learn gender-biased psychographic associations present in large text corpora. To evaluate the consequences of such algorithmic learning within digital marketing contexts, we next studied whether ad-targeting platforms, which leverage these algorithms, could deliver biased product offerings to consumers in a manner consistent with the learned gender biases.

To evaluate the consequences of biased algorithmic learning for consumers, we conducted a series of field experiments on Facebook and Google's ad platforms (reported in Appendices S7–S9). In each study, we manipulated the psychographic attributes included in ad copy to feature either positive attributes or negative attributes. The dependent variable was the gender distribution of the users served the advertisement. This design allowed us to assess whether advertisements targeting negative psychographic attributes were more likely to be served to women.

We observed a consistent pattern of biased ad delivery across these experiments. For instance, partnering with an existing astrology business, we found that ads targeting negative psychographic attributes (e.g., "Are you *irresponsible*?") were significantly more likely to be served to women compared to ads targeting their positive attribute counterparts (e.g., "Are you *dependable*?"; see Appendix S7). We observed a similar pattern of findings within a financial investment domain. Ads targeting negative psychographic attributes (e.g., "Save money for a better life: Investing tips for the *impulsive* investor") were significantly more likely to be delivered to women by the ad targeting algorithm compared to those targeting positive attributes (e.g., "…for the *planned* investor") across multiple ad platforms and optimization goals (impressions optimization vs. clickthrough optimization; see Appendices S8 and S9). The key results from these experiments are summarized in Table 4. We present an additional field experiment evaluating a debiasing strategy for firms in Appendix S12.

TABLE 4 Summary of gender bias in ad delivery field experiments.

Platform	Product domain	Psychographic attributes	Positive attributes	Negative attributes	Test statistics
Facebook (Field Experiment: Appendix S7)	Astrology	[+]: strong, relaxed, jolly, dependable, tough [−]: fragile, moody, irritable, irresponsible, sensitive	9.5% ads delivered to women	21.9% ads delivered to women	$N=26,720$ $\chi^2=792.1$ $p<0.001$
Google (Field Experiment: Appendix S8)	Investment/Financial Services	[+]: planned, disciplined, creative [−]: impulsive, dissatisfied, irresponsible	13.3% ads delivered to women	19.9% ads delivered to women	$N=11,260$ $\chi^2=86.2$ $p<0.001$
Facebook (Field Experiment: Appendix S9)	Investment/Financial Services	[+]: planned, disciplined, creative [−]: impulsive, dissatisfied, irresponsible	8.8% ads delivered to women	10.8% ads delivered to women	$N=18,332$ $\chi^2=20.4$ $p<0.001$

STUDY 2: GENDER BIAS IN CONSIDERATION AND CHOICE

Building on the initial field experiments, Study 2 was designed to examine how gender bias learned by algorithms can bias the consumer consideration set and choice in online marketplaces. For this, we used product search recommendations on shopping portals to evaluate how product recommendations differed for men and for women on the targeted psychographic attributes. Like display ads, product search recommendations are delivered to users algorithmically depending on relevance to search keywords, leveraging quantitative text representations to match the results to users.

Methods

We recruited 87 participants from TurkPrime (46 women, 41 men) who created new accounts on an online shopping platform and searched for different products. In the first phase of the study, participants were asked to create new accounts on Google or Bing (at random, to generalize across shopping platforms). The following day, men and women were asked to search on desktop/laptop computers for identical keywords (health, snacks, magazines, vacation, birthday gifts) and upload screenshots of the first 20 products delivered by the product recommendation algorithm (i.e., the consideration set). From each consideration set, participants were asked to choose one item that they liked most.

In the second phase, we recruited a separate sample of raters ($N=140$) who evaluated products from the consideration sets delivered to men and women by the recommendation algorithms. To evaluate gender bias in product recommendations, these independent raters evaluated the degree to which products matched positive and negative psychographic attributes, using a 6-question, 7-point scale with anchors corresponding to attributes (*rational–emotional, industrious–lazy, innovative–conformist, determined–vain, logical–frivolous, loyal–fickle*) drawn from the previous studies. Raters were given the meaning of the attributes and then asked to rate each product on these attribute scales. For instance, if a phase-one participant found Wheat Thins recommended as a snack, phase-two participants would evaluate whether the snack would be more suitable for *rational* vs. *emotional* consumers, *industrious* vs. *lazy* consumers, and so on.

Ratings were averaged to obtain a positive–negative attribute score for each product in each of the consideration sets. This served as the dependent variable—which we refer to as the bias of the consideration set—in which higher values corresponded to the consideration set being more closely associated with negative attributes.

We assessed whether bias existed in delivery of products to male and female consumers by examining the bias of the consideration set and chosen product.

Results

Consideration set

We conducted a hierarchical linear regression analysis in which consideration set ratings were nested within each attribute dimension. With consideration set ratings as the outcome variable, we entered the user gender (of the phase-one participant), website, and rater gender as fixed effects, and intercepts for each attribute as random effects.

Estimates revealed a significant main effect of user gender ($b=0.13$, SE$=0.041$, $p=0.0001$), suggesting that the product consideration set delivered to female consumers by the recommendation algorithm was negatively biased (a higher score means more negative ratings). We also observed a significant main effect of website ($b=0.10$, SE$=0.040$, $p=0.01$), indicating product recommendations from Google were more negatively biased than those from Bing; we did not observe any influence of the rater gender ($b=-0.04$, SE$=0.041$, $p=0.30$).

Product choice

Evaluating whether product search algorithms also biased consumer choice, we examined the bias of the items that male and female participants selected. As in the previous analysis, choice ratings were nested within each attribute dimension; we used choice ratings as the outcome variable, with user gender, website, and rater gender as fixed effects and random effect intercepts for each attribute.

We observed a significant main effect of user gender ($b=0.15$, SE$=0.064$, $p=0.02$), indicating that women also chose more negatively biased products relative to men from the consideration sets delivered to consumers by the recommendation algorithm. We also observed a significant main effect of website ($b=0.21$, SE$=0.064$, $p=0.001$), with choices made on Google more negatively biased; we did not observe any influence of rater gender ($b=-0.05$, SE$=0.063$, $p=0.40$).

Further analysis also confirmed that bias in the consideration set mediated effects on choice (see Appendix S10). The findings from this study illustrate material consequences for consumers who interact with gender-biased algorithms in the digital marketplace: Algorithms that learn to associate women with negative psychographic attributes from language subsequently deliver more negatively biased product recommendations to female users that bias consideration and choice.

STUDY 3: CONSEQUENCES OF GENDER-BIASED PSYCHOGRAPHIC ASSOCIATIONS IN AD TARGETING

The previous studies documented the role that ad targeting algorithms play in propagating gender biases learned from large text corpora to consumers in the digital marketplace. However, they do not inform us of the role that consumers may play in amplifying or impeding existing algorithmic gender biases. Do consumers accept stereotypes due to increased fluency and liking of identity-consistent offerings (Reed et al., 2012; Susser et al., 2016), thus guiding adaptive ad- targeting algorithms toward magnifying gender biases? Or do consumers instead reject stereotypes due to reactance against explicit stereotypes (Kray et al., 2001), thus preventing the propagation of gender biases? A key aim of Study 3 was to evaluate the role that consumers may (inadvertently) play in co-producing gender-bias in digital ad targeting.

We examined the role of consumer co-production by manipulating the ad campaign objectives to be more or less adaptive to consumer input. Ad-targeting platforms provide advertisers with two key options: whether to optimize clickthroughs (ad delivery is updated based on consumer clickthrough interactions) or optimize impressions (ad delivery is *not* updated based on consumer clickthrough interactions). Because ad-targeting algorithms exhibit comparatively greater adaptation to consumer input under clickthrough optimization relative to impressions optimization, differences in ad delivery between campaign objectives are indicative of the role consumers play in amplifying or impeding existing algorithmic gender biases. That is, if consumers co-produce gender bias by "accepting" gender stereotypes (e.g., negative attribute ads receiving greater clickthrough rates from women vs. men), then gender biases would be magnified in the more adaptive clickthrough-optimization campaign relative to the impressions-optimization campaign (i.e., a greater proportion of negative ads delivered to women). On the other hand, if consumers instead "reject" gender stereotypes (e.g., negative attribute ads receiving lower clickthrough rates from women vs. men), then gender biases would be reduced in the clickthrough-optimization campaign relative to the impressions-optimization campaign.

Methods

We collaborated with an astrologer who had been in practice for over two decades in the United States and wanted to leverage online advertising to reach a wider target audience. To test whether targeting algorithms display gender-biased delivery and whether consumers also co-produced the gender bias, we adopted a 2 (ad targeting attribute: negative vs. positive psychographic attribute) × 2 (campaign objective: impressions optimization vs. clickthrough optimization) experimental design.

We manipulated the psychographic attribute within the advertisement to be either positive (*strong*) or negative (*fragile*), selecting psychographic attributes from our Study 1 analysis that revealed large gender-biased associations. The gender distribution of users who were delivered the ad served as the dependent variable (i.e., the "subject" in these causal tests was the ad-targeting algorithm, and the dependent variable was how it behaved in terms of delivering ads to men and women; see Figure 2 for example stimuli).

The advertisements were simultaneously deployed on a major advertising platform (Facebook). We used keywords such as "healing," "astrology," and "chakra healing" to select the audience, with the location limited to the United States. Each campaign had a budget of $50 and ran for one day. An important note: We did not specify any gender in the targeting settings. To probe whether the ad-targeting platform updated gender distribution of users in response to consumer interaction, we took snapshots at 15-min intervals.

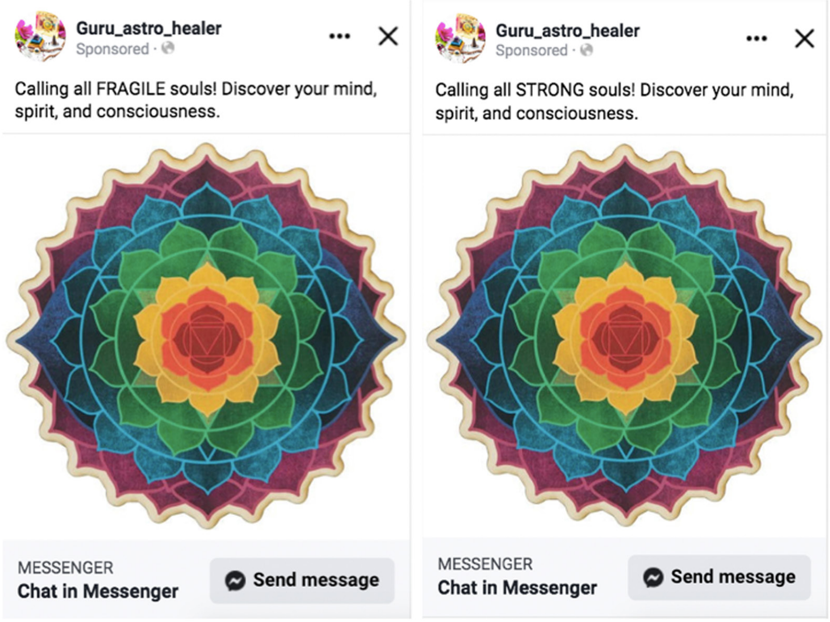
Results

Gender-biased ad delivery

We first tested whether ad delivery was gender biased. Note that we obtained impression measures across both the impression-optimization and clickthrough-optimization conditions ($N=16{,}282$). The analysis in the impression-optimization campaign indicated that the negative-attribute ad was delivered to a significantly greater percentage of women (38.9%) compared to the positive-attribute ad (36.5%, $\chi^2(1, N=8819)=5.15$, $p=0.023$). This result suggests that varying only the psychographic attribute in an advertisement can bias the resulting gender distribution of users who are shown the ad, consistent with our findings from additional ad platforms, product domains, and psychographic attributes reported in Appendices S7, S8, and S9.

Consumer co-production of gender bias

To test consumers' role in co-producing the bias, we compared campaign objectives. We found that under clickthrough optimization, the negative-attribute ad was delivered to a significantly greater percentage of women (70.5%) compared to the positive-attribute ad (54.8%, $\chi^2(1, N=7463)=195.2$, $p<0.001$). Most notable, submitting gender to a logistic regression on attribute valence, campaign objective, and their interaction revealed a significant interaction effect ($\chi^2(1, N=16{,}252)=72.8$, $p<0.001$), indicating that the gender bias in clickthrough optimization was significantly greater than that in impressions optimization (70.5% negative vs. 54.8% positive ads delivered to women in clickthrough optimization compared to 38.9% negative vs. 36.5% positive ads delivered to women in impressions optimization). Main effects of attribute valence ($\chi^2(1, N=16{,}252)=133.7$, $p<0.001$) and campaign objective ($\chi^2(1, N=16{,}252)=1005.3$, $p<0.001$) were also significant, showing that negative ads and ads in the clickthrough-optimization campaign were



Calling all FRAGILE souls! Discover your mind, spirit, and consciousness.

Calling all STRONG souls! Discover your mind, spirit, and consciousness.

FIGURE 2 Display advertisement stimuli.

delivered to a greater percentage of women overall (see Appendix S11). This pattern of findings (main effect and interaction) is consistent with the hypothesis that gender bias is co-produced due to algorithmic learning of gender biases from large text corpora that is amplified based on adaptive learning from affirmative consumer responses to biased ad offerings.

Temporal dynamics of co-produced gender bias

If it is true that algorithms update ad delivery based on user interactions, such as clickthroughs indicating interest and match, then the gender distribution of ad delivery should show an increasing pattern in the clickthrough campaign but not in the impression campaign. Specifically, if women click on the negative ads displayed to them, then they also (inadvertently) contribute to gender-biased algorithmic learning. We examine ad delivery over time to test for this co-production. Regression analyses of the percent of women ad recipients on campaign, timepoint, and their interaction revealed significant interactions when comparing across campaigns both positive attribute ads ($b=0.135\%$, $SE=0.011\%$, $t(76)=12.88$, $p<0.001$) and negative attribute ads ($b=0.154\%$, $SE=0.006\%$, $t(76)=22.45$, $p<0.001$), indicating that the rate of change in user gender distribution was significantly greater for the clickthrough-optimization campaign compared to the impressions-optimization campaign (see Figure 3). Further analyses within only the clickthrough-optimization condition also confirmed that the gender bias increased over time; moreover, we found consistent results in time-series analyses that account for potential autocorrelation in the data (Appendix S11).

We found that in the impressions-optimization campaign, there were no differences in clickthrough rates between men and women for either positive-attribute or negative-attribute ads (χ^2s<1.18). However, in the clickthrough-optimization campaign, women clicked on negative-attribute ads significantly more often than men (1.63% vs. 0.76%, $\chi^2(1, N=3578)=4.15$, $p=0.042$); women also clicked on positive-attribute ads marginally more often (1.08% vs. 0.52%, $\chi^2(1, N=3855)=3.32$, $p=0.057$). These findings indicate that clickthrough-optimization was more effective in eliciting clicks from women compared to men, with the negative-attribute advertisement eliciting significantly greater interest from women vs. men (consistent with the idea that consumers "accept" gender stereotypes).

These findings demonstrate that algorithmic learning of gender bias from language can result in biased delivery of advertisements to men and women consumers, who can amplify the algorithmic bias through acceptance of biased offerings when interacting with adaptive algorithms.

GENERAL DISCUSSION

Theoretical implications and future research

The findings in this research enhance our understanding of how gender bias learned by algorithms from large text corpora can have consequences for consumers in online marketplaces. In an analysis based on billions of documents from Common Crawl, we build on prior research to illustrate that algorithms can learn gender-biased associations with marketplace-relevant psychographic attributes even when there is no psychological basis for such gender associations (Study 1). Across a series of field experiments on Facebook, Google, and Bing, we document material consequences for consumers in the form of biased product offerings delivered to women through gender-biased ad targeting and product recommendation algorithms that have not been examined in prior work (Study 2; Appendices S7–S9). Finally, we unpack the role consumers can play in amplifying algorithmic gender bias through their acceptance of biased offerings, providing greater insight into the consumer–algorithm co-production of gender bias (Study 3).

By characterizing the consumer-relevant consequences of algorithmic gender bias in digital marketplaces, this work generates new directions for future research. Understanding the conditions under which consumers *reject* gender-biased algorithmic offerings can provide an important avenue to stem the propagation of gender biases in online marketplaces, given that consumer responses can inform adaptive algorithms. While advertisements targeting negative psychographic attributes may be subtle, efforts to increase consumer attention to and awareness of gender-biased offerings may lead to greater rejection of stereotypes as prior research has shown that people are more inclined to reject gender stereotypes when they are made explicit rather than implicit (Kray et al., 2001). Our findings also query the long-term downstream consequences of repeated exposure to gender-biased offerings in digital marketplaces. Because consumers can develop a self-identity and self-concept based on the environments they inhabit and the possessions they acquire (Belk, 1988; Berger & Heath, 2007), consistent gender biases in the items consumers are offered, consider, and purchase could recultivate historical prejudices by influencing consumer psychology in insidious ways. In the current work, we identify a general tendency for word-embedding algorithms to associate negative psychographic traits more closely with women than men. Further research could explore whether there exist contexts in which algorithms may

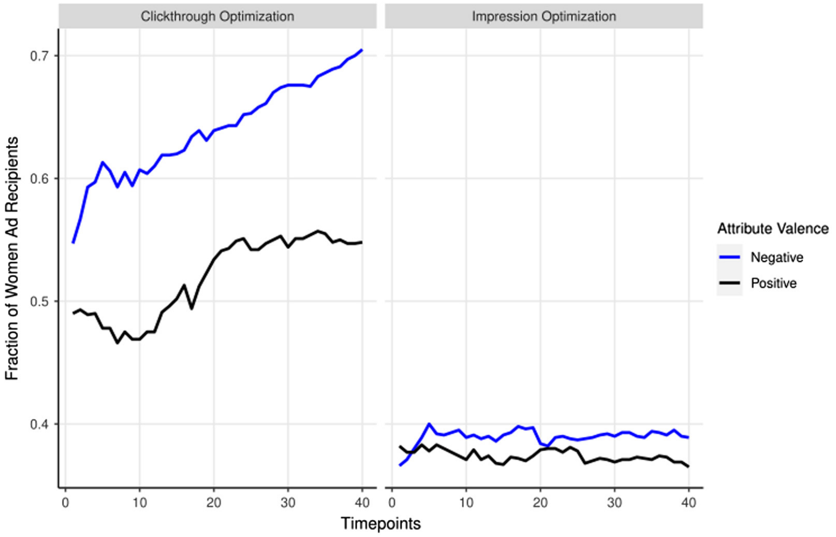


FIGURE 3 Percentage of ads delivered to women across the duration of the campaign.

learn more positive gender associations. Traditional positive stereotypes (such as women displaying more warmth than men) are often not learned by embedding algorithms (DeFranza et al., 2020) because of the propensity (or bias) to associate many positive traits more with men than with women. Although our research focuses exclusively on gender bias, relationships between algorithmic gender bias and other forms of bias such as racial or age-based bias may be important to understand how they interact and compound each other in the online marketplace when interacting with different types of algorithms (Agan et al., 2023).

Future studies should explore not just ways in which algorithms can learn and propagate gender bias (Boghrati & Berger, 2022; Caliskan et al., 2017) but also the more important downstream consequences of such a bias on various consumer decisions such as satisfaction, segmentation, loyalty etc. Studies could investigate if online algorithms are susceptible to biases such as misogyny and other diversity, equity, and inclusion related constructs. Second, research has suggested that the way men and women decide and work with information may be different (Bristor & Fischer, 1993; Spielmann et al., 2021). Hence, acknowledging and understanding how biases may be propagated because of these differences is another area of research. Third, future studies should investigate how different biases and stereotypes (e.g., it is assumed that female consumers are not good at financial planning; Lee et al., 2011) such as gender vs. racial vs. age could result in downstream consequences in the marketplace. Fourth, research should further examine how consumers can be taught to safeguard against biased recommendations and, as a first step be able to recognize the presence of bias. For instance, as our research demonstrates, consumers may inadvertently assist in letting biased recommendation stay because they are not identifying the bias and allowing it perpetuate. In other words, methods in which consumers can be educated to spot and reject biased recommendations are an important area of future research. Fifth, such an ability to spot bias is not just helpful in the marketplace but also in other areas of business such as hiring or employee evaluation (Mohr & Henson, 1996) in which we are seeing an increased use of algorithms. Sixth, much research in consumer psychology can be leveraged in understanding how different personality traits can affect people's ability to spot and reject or accept biased recommendations – the interactive examination would provide important insights. Seventh, an important moderator in today's time-strapped world is the ability to detect bias when people are multi-tasking, stressed or working under cognitive load. Moreover, new research can examine how recommendations provided not just algorithmically but also through social media platforms can result in spreading of gender-biased stereotypes.

Practical implications

In order to mitigate the negative consequences of gender-biased algorithms in the marketplace, greater awareness of the existence of these biases would encourage consumers to search for alternative sources of information. Since consumers tend to trust and rely on personalized recommendations, enabling consumers to understand biases in online ad delivery can foster healthy skepticism. However, advertisers and marketers should also be careful to pick attributes in marketing materials that minimize gender-biased delivery to consumers. As our findings illustrate, algorithms learn and propagate gender biases from language; thus, evaluating the degree of gender bias in marketing materials prior to launching ad campaigns can provide advertisers with a practical solution for averting gender-biased delivery of their offerings to users.

We conducted an additional field experiment to evaluate this strategy (presented in Appendix S12). Collaborating with an astrology business, we manipulated the key psychographic attribute featured in the advertisement based on gender-bias estimates from our Common Crawl word-embedding analysis. Comparing three different advertisements (including a strongly biased attribute, weakly biased attribute, and a no-attribute control condition), we found that while the strongly biased attribute resulted in gender-biased ad delivery, when the firm selected a weakly biased attribute we did not observe significant gender bias in delivery relative to the control. This study offers a practical way for advertisers to anticipate and avoid gender-biased delivery of ad campaigns based on psychographic attributes included in the ad copy language.

In addition to contributing to recent research on linguistics in consumer research (Kronrod et al., 2020; Packard & Berger, 2023), our findings also have important ethical, legal, and policy implications. Firms routinely gather insights by using algorithms and then base their decisions on these insights. If algorithms are influencing so many aspects of decision making, they should not come with the risk of caveat emptor. Because consumers are largely unaware of how ad-targeting algorithms can limit their access to goods and services and because of limited human processing capacity, they may be unable to independently discover such biased offerings themselves, it is important for firms to consider fairness objectives in conjunction with ad efficiency and optimization objectives.

ORCID

Shelly Rathee https://orcid.org/0000-0002-4403-7775

REFERENCES

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of Marketing Research*, *34*(3), 347–356.

Agan, A., Davenport, D., Ludwig, J., & Mullainathan, S. (2023). *Automating automaticity: How the context of human choice affects the extent of algorithmic bias*. NBER Working Paper.

Ailawadi, K. L., Neslin, S. A., & Gedenk, K. (2001). Pursuing the value-conscious consumer: Store brands versus national brand promotions. *Journal of Marketing*, *65*, 71–89.

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

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of Personality and Social Psychology*, *9*(3), 272–279.

Belk, R. W. (1988). Possessions and the extended self. *Journal of Consumer Research*, *15*(2), 139–168.

Berger, J., & Heath, C. (2007). Where consumers diverge from others: Identity signaling and product domains. *Journal of Consumer Research*, *34*(2), 121–134.

Berry, D. S., & McArthur, L. Z. (1985). Some components and consequences of a babyface. *Journal of Personality and Social Psychology*, *48*(2), 312–323.

Boghrati, R., & Berger, J. (2020). *Quantifying 50 years of misogyny in music*. Working paper.

Boghrati, R., & Berger, J. (2022). *Quantifying gender bias in consumer culture*. *arXiv* preprint arXiv:2201.03173.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, Y., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, *29*, 4349–4357.

Brehm, J. W. (1966). *A theory of psychological reactance*. Academic Press.

Briggs, S. R. (1992). Assessing the 5-factor model of personality description. *Journal of Personality*, *60*(2), 253–293.

Bristor, J. M., & Fischer, E. (1993). Feminist thought: Implications for consumer research. *Journal of Consumer Research*, *19*(4), 518–536.

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

Caselles-Dupré, H., Lesaint, F., & Royo-Letelier, J. (2018). Word2vec applied to recommendation: Hyperparameters matter. In *Proceedings of the 12th ACM Conference on Recommender Systems* (pp. 352–356).

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240.

Crawford, K. (2017). *The trouble with bias*. Speech at Conference on Neural Information Processing Systems.

DeFranza, D., Mishra, H., & Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology*, *119*(1), 7–22.

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

Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology*, *69*, 275–298.

Epstein, C. F. (1988). *Deceptive distinctions: Sex, gender, and the social order*. Yale University Press.

Eysenck, H. J. (1982). The biological basis of cross-cultural differences in personality: Blood group antigens. *Psychological Reports*, *51*(2), 531–540.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3635–E3644.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480.

Hofstee, W. K., De Raad, B., & Goldberg, L. R. (1992). Integration of the big five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*(1), 146–163.

Hyde, J. S. (2016). Sex and cognition: Gender and cognitive functions. *Current Opinion in Neurobiology*, *38*, 53–56.

Hyde, J. S., & Plant, E. A. (1995). Magnitude of psychological gender differences: Another side to the story. *American Psychologist*, *50*(3), 159–161.

Kimball, M. M., Cole, E., & Rothblum, E. D. (1995). *Feminist visions of gender similarities and differences*. Psychology Press.

Kray, L. J., Thompson, L., & Galinsky, A. (2001). Battle of the sexes: Gender stereotype confirmation and reactance in negotiations. *Journal of Personality and Social Psychology*, *80*(6), 942–958.

Kronrod, A., Packard, G., Moore, S. G., Berger, J., Inman, J., Meyer, R., Shrum, L. J., Humphreys, A., Lurie, N., Luangrath, A. W., & Lee, J. (2020). Where consumer behavior meets language: Applying linguistic methods to consumer research. In J. Argo, T. M. Lowrey, & H. J. Schau (Eds.), *NA – Advances in* *consumer research* (Vol. *48*, p. 1247). Association for Consumer Research.

Lee, K., Kim, H., & Vohs, K. D. (2011). Stereotype threat in the marketplace: Consumer anxiety and purchase intentions. *Journal of Consumer Research*, *38*(2), 343–357.

Liu, P., Joty, S., & Meng, H. (2015). Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1433–1443).

Manis, M., Nelson, T. E., & Shedler, J. (1988). Stereotypes and social judgment: Extremity, assimilation, and contrast. *Journal of Personality and Social Psychology*, *55*(1), 28–36.

May, C., Wang, A., Bordia, S., Bowman, S. R., & Rudinger, R. (2019). *On measuring social biases in sentence encoders*. arXiv preprint arXiv:1903.10561.

Mohr, L. A., & Henson, S. W. (1996). Impact of employee gender and job congruency on customer satisfaction. *Journal of Consumer Psychology*, *5*(2), 161–187.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Ozsoy, M. G. (2016). *From word embeddings to item recommendation*. arXiv preprint arXiv:1601.01356.

Packard, G., & Berger, J. (2023). The emergence and evolution of consumer language research. *Journal of Consumer Research*. https://doi.org/10.1093/jcr/ucad013

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. (pp. 1532–1543).

Raju, P. S. (1980). Optimum stimulation level: Its relationship to personality, demographics, and exploratory behavior. *Journal of Consumer Research*, *7*(3), 272–282.

Reed, A., II, Forehand, M. R., Puntoni, S., & Warlop, L. (2012). Identity-based consumer behavior. *International Journal of Research in Marketing*, *29*(4), 310–321.

Roberts, B. W., Wood, D., & Smith, J. L. (2005). Evaluating five factor theory and social investment perspectives on personality trait development. *Journal of Research in Personality*, *39*(1), 166–184.

Spielmann, N., Dobscha, S., & Lowrey, T. M. (2021). Real men don't buy "Mrs. Clean": Gender bias in gendered brands. *Journal of the Association for Consumer Research*, *6*(2), 211–222.

Steenkamp, J. B. E., & Maydeu-Olivares, A. (2015). Stability and change in consumer traits: Evidence from a 12-year longitudinal study, 2002–2013. *Journal of Marketing Research*, *52*(3), 287–308.

Susser, J. A., Jin, A., & Mulligan, N. W. (2016). Identity priming consistently affects perceptual fluency but only affects metamemory when primes are obvious. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(4), 657–662.

World Bank. (2020). *Women, business and the law 2020*. The World Bank.

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. *American Psychologist*, *70*(1), 10–20.

Zheng, L., Noroozi, V., & Yu, P. S. (2017). Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (pp. 425–434).

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rathee, S., Banker, S., Mishra, A., & Mishra, H. (2023). Algorithms propagate gender bias in the marketplace—with consumers' cooperation. *Journal of Consumer Psychology*, *33*, 738–631. https://doi.org/10.1002/jcpy.1351