I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17282

He Said, She Said: Who Gets Believed When Spreading (Mis)Information

Nuzaina Khan
David Rand
Olga Shurchkov

SEPTEMBER 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17282

He Said, She Said: Who Gets Believed When Spreading (Mis)Information

Nuzaina Khan
Oxford University

David Rand
Massachusetts Institute of Technology

Olga Shurchkov
Wellesley College and IZA

SEPTEMBER 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

He Said, She Said: Who Gets Believed When Spreading (Mis)Information*

We design an online experiment that mimics a Twitter/X "feed" to test whether (perceived) poster gender influences users' propensity to doubt the veracity of a given post. On average, posts by women are less likely to be flagged as concerning than identical posts by men. Heterogeneity analysis reveals that men are more likely to flag female-authored posts as the post's topic domain becomes more male-stereotyped. Female users do not exhibit the same bias. Actual post veracity, user ideology, and user familiarity with Twitter do not explain the findings. Flagging behavior on Twitter's crowdsourced fact-checking program is consistent with these findings.

JEL Classification: C90, D9, J16, L86
Keywords: gender differences, misinformation, economic experiments

Corresponding author:
Olga Shurchkov
Department of Economics
Wellesley College
106 Central St.
Wellesley, MA
USA
E-mail: olga.shurchkov@wellesley.edu

* We wish to thank Jennifer Allen, Tatyana Deryugina, Jakina Guzman, Eric Hilt, Benjamin Ho, Cameron Martel, Danila Serra, and the participants of the Economics Research Seminar at Wellesley College, the New England Experimentalist Workshop 2023, Bocconi University's AXA Research Lab on Gender Equity seminar, the ASSA 2024 ESA session "Experiments on Gender Differences in Behavioral Traits, " and Amherst College Economics Department Seminar for helpful discussions and feedback. We thank Emily Brydges, Stella Gu, and Julia Xie for excellent research assistance and gratefully acknowledge financial support from Wellesley College Faculty Awards and the Human Cooperation Laboratory at MIT. The experiment in this paper was approved by the Institutional Review Board and preregistered at the AsPredicted 125111. All remaining errors are our own.

1 Introduction

Social media platforms, such as Twitter,[1] have emerged as the modern arenas where political elections are fought and decided, reputations are built and tarnished, and careers are either advanced through self-promotion or impeded by damaging allegations (see Allcott and Gentzkow 2017 for a discussion of the 2016 US presidential election). Mastering the art of persuasion on social media has become the contemporary method of influencing opinions and swaying public perceptions. In this important sphere, women appear to have fallen behind, as evidenced by significant gender gaps in Twitter use and influence.[2] For example, women receive significantly fewer likes and attract fewer followers than their male counterparts in academic medicine (Zhu et al., 2019). Our own analysis of observational data from Twitter's "Community notes" shows that tweets posted by women are significantly more likely to be flagged as harmful. The underrepresentation of women in social media discourse is particularly alarming, since men have been found to be more likely to knowingly spread misinformation (Buchanan 2020).

One explanation for the gender gap in social media influence could be that women and men post on different topics and use different communication styles. For example, Beltran et al. (2021) show that male and female politicians use different words and emojis in their statements on Twitter, conforming to their gender stereotypes. In a more representative sample, Hu et al. (2020) find that women focus on family- and home-related issues to a greater extent than men do in their posts. Another explanation relates to network dynamics: the incentive to publicly undermine women comes from the desire to enhance one's own status in the network and/or to strategically signal in-group identity (van der Does et al. 2022). Finally, it is possible that, even conditional on the post's content and style and absent any network effects, women are still considered less knowledgeable about a given topic, particularly when posting in a stereotypically male domain, such as military and finance matters (Huddy and Terkildsen 1993). Observational studies cannot distinguish between these explanations.

We design an online survey experiment that simulates engagement on Twitter to provide a clean explanation for the observed gender gap in influence on social media. In particular, we examine the circumstances under which identical statements made by women and men receive differential treatment from users. Our main finding is that men are more likely to flag female-authored posts as the post's topic domain becomes more male-stereotyped. Female users do not exhibit the same bias. These results hold true regardless of the statement's truthfulness, the user's political

[1] The social media platform Twitter was recently rebranded by its owner, Elon Musk, including a name change to X (Corse, Alexa, Collin Eaton, and Newley Purnell. 2023. "Elon Musk Replaces Twitter's Blue Bird With an 'X.'" The *Wall Street Journal*, July 24, 2023. Accessed online on August 1, 2023 at https://www.wsj.com/articles/elon-musk-says-twitter-will-change-its-logo-to-x-5f73c349). Our experiment took place prior to this change. Throughout this paper, we will refer to the social media platform by its former name "Twitter."

[2] Dixon, Stacey Jo. 2024. "Gender distribution of social media audiences worldwide as of January 2023, by platform." *Statista*. Accessed online on April 2, 2024 at https://www.statista.com/statistics/274828/gender-distribution-of-active-social-media-users-worldwide-by-platform/ Social media use is associated with well-documented individual-level and societal harms (see for example, Allcott et al. 2020). Thus, there is reason to select away from engaging online. This paper focuses on the potential gains from social media visibility, conditional on having selected into making public statements via posts.

ideology, or their familiarity with Twitter. We also validate these findings in Twitter field data of "Community Notes" flags of statements made by male and female politicians on Twitter (Allen et al., 2022).[3] The findings are consistent with the explanation that biased beliefs (stereotypes) about differential abilities of men and women in certain tasks lead to discrimination and result in gender gaps in outcomes (Bordalo et al. 2019; Bohren et al. 2019).

The study contributes to a growing literature exploring the role of gender stereotypes in perpetuating disparities in wages and career progression within traditionally male-dominated fields, such as finance (Goldin et al., 2017), economics (Lundberg and Stearns 2019), and STEM (Michelmore and Sassler 2016). Recent experimental research shows that these gaps vanish in similar situations characterized by female stereotypes (see Shurchkov 2012 for the willingness to compete, Coffman et al. 2021a for discrimination in hiring, and Coffman et al. 2021b for bias in team leader selection). This paper highlights that perceived credibility of women on social media is also affected by gender stereotypes.

The second contribution of this study is to investigate the demand-side of spreading information (truthful and false). Much is known in the literature about the determinants of the production side of misinformation (see Abeler et al. 2019 for meta-analysis of experimental literature on lying and cheating; Vosoughi et al 2018 on the spread of misinformation on social media).[4] The demand-side, however, is to our knowledge, much less well-understood. By exogenously manipulating the gender of the source of information, as well as its veracity, we test whether receivers of information respond to the identity of the person sharing it: who gets believed and when.

2 Methodology

To study the impact of gender on the likelihood of being believed, we designed an online survey experiment that closely mimics interactions on Twitter (pre-registered on AsPredicted 125111). Twitter presents a particularly compelling setting to study stereotypes because of its fast-paced interactions and low-attention environment, conducive to revealing biases and stereotypical thinking that might remain hidden or unobserved in more controlled or less spontaneous settings.

2.1 Participants

The study was programmed in Qualtrics and conducted online in March 2023 via Lucid, an online survey platform that aggregates responses from many survey providers, relying on quota sampling to provide researchers with nationally representative samples (Coppock and McClellan 2019). The platform was chosen particularly because of our social media context which is characterized by low attention rates, and because studies of misinformation conducted on Lucid produce robust replicable results with effect sizes that are typically smaller relative to other platforms (Pennycook and Rand 2022).

---

[3] "Community Notes" (previously called Birdwatch) is Twitter's crowdsourced fact-checking program, where we can directly observe judgments of whether certain tweets are considerably harmful, and whether the associated free-text evaluations (notes) of tweets are deemed helpful.

[4] On average, women are significantly less likely to lie than men (Abeler et al 2019), and the gender gap in honesty seems to hold across cultures (Cohn et al 2019).

Following our pre-registration, we aimed to recruit 1,000 adult (18 or older) US residents, in line with previous research on misinformation and online behavior (Pennycook et al. 2020). The final sample is restricted to subjects who provided informed consent and passed both attention checks, resulting in 818 respondents. Most of our demographic variables closely track Census data with 51.5% of participants identifying as female; 71% identifying as White, 13% Black, and 5% Asian; 14% identifying as Hispanic (slightly below the national average of 19%; U.S. Census 2022); and the mean age of 46 years (which is slightly older than the national average of 39 years; U.S. Census 2023). See SI appendix Table C1 for summary statistics.

2.2 Procedures

After answering a few questions aimed at subtly priming gender (as in Shurchkov and van Geen 2019), participants viewed a series of 24 posts, randomly drawn from a pool of 96 total. The posts appeared on a stylized Twitter "feed," each containing text and a profile icon (see Figure 1 for example posts). All posts were sourced from online non-partisan fact-checking organizations (Politifact.com, FactCheck.org, or Snopes.com) and represented original content from political speeches, various social media posts, and news headlines (full list of posts and their sources is provided in the SI Appendix B).

[Figure 1 about here]

Subjects could interact with each post by flagging, liking, and/or retweeting via designated buttons. The instructions clearly explained that flagging was an unambiguously negative reaction: if a subject came across a post they would normally flag on social media out of concern or dislike, they were to click the 'flag' button.[5] Although the determinants of the decision to flag was our primary interest, the possibility of liking and retweeting posts was added for the sake of ecological validity. Liking was framed as an unambiguously positive reaction: if a subject came across a post they would like if they were on social media, they were to click the 'like' (or 'heart') button. The instructions only indicated to click the 'retweet' button for posts that they would normally share on social media. Since sharing may happen for many reasons – some negative (flagging posts as outrageous or fake) and some positive (sharing interesting information) – we expect that clicking the 'retweet' button may have been an ambiguous reaction. Subjects were able to practice with all three buttons prior to starting the main experimental block.

While it was possible not to engage with a given post, subjects were encouraged to "read the tweet carefully and interact with it if they wished" (see the SI appendix A for a full set of experimental instructions). At a random point during the main experimental block, participants encountered two attention-check tweets for which they were asked to perform a specific action (like the first and flag the second). Subjects who failed both attention checks were shown error messages but could still complete the survey; these observations are excluded in the analysis.

[5] "Flagging" is used to signal to users that the content of information is objectionable or in violation of terms of service (Crawford and Gillespie 2016). The primary motive for flagging content on social media has been to identify and reduce the spread of misinformation (Allcott and Gentzkow 2017). In addition to explaining the purpose of this button explicitly, we expect the negative connotation to be salient to the subjects from own experiences online.

After viewing all 24 tweets, subjects were taken to a demographic questionnaire. The experiment ended with debriefing, where participants were informed about the true purpose of the experiment, in order to minimize potential risks associated with reading misleading or false information.

2.3 Design Considerations

We designed the user experience to mimic scrolling through a Twitter feed with the aim of maximizing the ecological validity. For example, by not forcing subjects to engage with tweets, we are able to measure "lurking" behavior and get more accurate effect estimates, which is difficult to do in observational data. As is customary in the literature studying misinformation on social media (Pennycook and Rand 2020), subjects were not incentivized based on their reactions to posts but rather were paid a base payment in accordance with Lucid compensation mechanism.[6]

However, we also modified the setting in order to control the environment. Firstly, we made the flag button as accessible as the like and retweet buttons, in order to ensure that the differences in behavior were not driven by the differences in the salience of the retweet, like, and flag buttons. Secondly, we included two control questions with each post. One asked the participants to identify the hair color of the poster (choice from five options, including not applicable). The purpose of this question was to nudge subjects toward looking at the profile icon, making the gender of the post's author salient in a subtle way. The other question asked the participants to identify the broad topic of the tweet, in order to ensure that subjects read the post and paid attention to its topic domain. Finally, we blurred out two kinds of typically available information to ensure control and minimize experimenter demand effects. First, we blinded the username of the poster of the tweet to eliminate any potential confounds that might be associated with information revealed via one's name (including race, ethnicity, age, and other characteristics, see for example Elder and Hayes 2023). Second, we blurred out the post's timestamp, as well as the number of existing reactions, because individuals may use verification strategies, including the number of retweets, to ascertain the truth, relying on these external validations rather than the post's content or the identity of the poster (Dabbous et al. 2021).

2.4 Treatments

Gender of poster

Our main treatment randomization is implemented at the subject-tweet level. Each tweet can randomly appear as part of the 22-tweet feed (excluding attention checks) arranged in random order and accompanied by one and only one of the following kinds of profile images: an ostensibly female poster; an ostensibly male poster; an inanimate poster or no profile image (see Figure 1).[7] No male or female profile picture appeared more than once on a given subject's feed. Inanimate

[6] Lucid charges researchers a CPI (cost per completed interview). In our case, CPI was 1 USD and that money was then transferred from Lucid to its suppliers who in turn send a portion to the participants. Participants received this compensation in different varied forms including cash and loyalty rewards.

[7] Male and female pictures were obtained from publicly available free stock images on Pexel.com. We restricted the profile images to White-presenting individuals in order to abstract away from the effects of race and ethnicity and their interactions with gender. We leave the important analysis of intersectional effects for future research. For a full set of posts that include profile images, refer to the SI appendix H.

images and tweets with no profile picture (default Twitter logo) were included in order to enhance ecological validity and minimize experimenter demand effects. We relied on past research to identify images that have minimal gender associations for the inanimate profile pictures (Meagher 2017), but we nevertheless exclude these observations from our main analysis.

Tweet veracity

The second treatment dimension corresponds to whether the statement in a given post is objectively true or false, as determined by third-party professional fact-checking organizations. The 96 posts in our pool are equally divided into false and true statements.

Topic domains

The final dimension is the extent to which the tweet's subject aligns with stereotypically male areas of interest and expertise. Here, we rely on two exogenous ways of classifying tweets into topic domains.

The first way is a crude classification based on the broad categories assigned to the post by the fact-checking organizations (see the SI appendix B for an illustration of how this information is presented on these sites). Using this criterion, the posts were equally split between healthcare, education, finance, and defense/military topics, with the first two topics categorized as female-typed and the second two topics categorized as male-typed based on the previous literature (Huddy and Terkilson 1993).

However, closer inspection of the content of posts revealed that content with multiple subtopics requires a more nuanced classification. For example, the following post was classified under the broad category of education on Politifact.com:

"*The FBI is using its counterterrorism division to investigate and add 'threat tags' to parents who are protesting school boards*."

The post touches upon multiple themes, including law enforcement, civil rights, and free speech, in addition to the topic of schools and education. Thus, we took a systematic approach to classifying posts into topics and domains using generative AI technology, ChatGPT-4.[8] In particular, we applied a two-step algorithm to each of the 96 posts. The first step classified each tweet into broad categories by topic domain. We took the top three topics, and for each, asked the tool to determine whether women or men would more likely to be interested in the topic, based on the gender stereotypes literature. The responses were then coded as neutral (no clear consensus on whether women or men are more likely to be interested in the topic; maleness value of 0); female-typed (often assumed that women may have a higher interest or be more directly affected by the topic or pursue careers in related areas; maleness value of -0.5); or male-typed (there is a common perception that men are more likely to be interested in the topic or pursue careers in related areas; maleness value of 0.5). There is a small number of more ambiguous topics, between gendered and neutral, in which case we assigned them intermediate values of -0.25 or 0.25. Once

[8] Gilardi et al (2023) show that ChatGPT outperforms crowd workers on platforms such as Amazon Mechanical Turk for several annotation tasks, including relevance, stance, topics, and frame detection.

the three topics for each post were coded, the overall post's maleness score was computed as the simple average of the three values. The SI appendix B provides the full classification protocol.[9]

Table 1 provides the number of observations by treatment (gender of poster), veracity of information (true or false statement) and topic domain as determined by the ChatGPT-4 procedure.[10] Note that although the topic domain takes on values between -0.5 (fully female-typed) and 0.5 (fully male-typed), for presentation purposes we have broken the variable up into three bins: male (value higher than 0.25); female (value less than 0); neutral (value between 0 and 0.25). In subsequent analysis, we use the more continuous version.

[Table 1 about here]

2.4 Descriptive statistics for main outcomes

Reactions to posts represent our outcome variables: *flagged*, *liked*, and *retweeted*. Because subjects could abstain from any of these reactions, we can also measure the total level of engagement (*engaged*) which takes on the value of 1 if any of the three reactions is activated for a given post. In the SI appendix E1, we show that the results are even stronger when we restrict the sample only to those participants who engaged in some way with a given post.

Across all posts, the engagement rate in our sample was 32%. A large fraction of the subjects did not react to any of the 22 posts they viewed (35%), but subjects who did react in some way did so with high frequency: for example, 20% of men and 16% of women reacted to more than 80% of the tweets they came across on their feed (see the distribution in the SI appendix, Figure C2).

On average, 10% of all tweets were flagged, 13% were liked, and 10% were retweeted. False tweets are 5pp more likely to be flagged than true tweets (t-test p-value <0.001) and are significantly less likely to be liked (p-value = 0.010) or retweeted (p-value = 0.010), although the magnitudes of the differences are less pronounced for these measures (see Figure C3 in the SI appendix).

[9] The reclassification of tweets into domains resulted in some switching from male to female and vice versa, but most simple shifted away from extremes, while maintaining the overall gender skew. More posts that were originally coded as female-typed based on the crude online classification moved toward neutral or even male-leaning than the other way around. This is because the majority of posts relate to politics or policy, even when the topic is education or healthcare. Based on the literature on gender stereotypes, the AI algorithm classified the topic of politics as male-leaning. For this reason, the female bin range in Table 1 is wider than the male bin. The SI appendix F repeats the analysis using the classification of topics into domains sourced from the fact-checking websites, finding that the results are robust to using the original broad topic domains.

[10] The numbers reflect our final sample that excludes subjects who failed attention checks and drops the observations for Tweet ID 47 which was incorrectly coded such that a non-human poster had two times the likelihood of appearing as either male or female poster. Results are robust to including this tweet (see the SI appendix).

3 Experimental Hypotheses and the Empirical Approach

Our hypotheses pertain to gender differences in reacting to the same information presented by posters of different genders. Our pre-analysis plan focuses on the negative action of flagging, formulating the following hypotheses we test with our experiment.[11]

3.1 Pre-registered experimental hypotheses

Hypothesis 1a: On average, tweets posted by men and women are equally likely to be flagged as misinformation.

Multiple factors that may influence one's likelihood of flagging/mistrusting posts by men and women. Previous research has found that women scored higher on implicit trustworthiness than men (Haran 2018) which may lead to lower incidence of flagging. On the other hand, men may be considered more competent than women (Fiske 2018; Coffman et al 2021b). Finally, because some statements are male-stereotyped and some are female-stereotyped, by design, we would expect to see no effect of poster gender, on average.

Hypothesis 1b: Men and women are, on average, equally capable of identifying (flagging) false tweets.

This is formulated as a null because previous findings regarding gender differences in fake news detection are mixed. Almenar et al. (2021) find no significant gender differences, while Arin et al. (2023) find that women detect fake news less frequently than men.

Hypothesis 2: Men are significantly more likely than women to be believed (flagged less) when posting on topics perceived as male-typed, while women are more likely than men to be believed (flagged less) when tweeting about female-typed topics.

Scrolling through social media posts is a low-attention environment (Epstein et al. 2023) which is conducive to the use of impulsive reasoning over analytical thinking, resulting in the use of representativeness or stereotypes to make decisions (Kahneman and Tversky 1972). This in fact has been shown as one of the main causes of the spread of misinformation online, rather than any kind of political or ideological motives (Pennycook et al. 2021). Therefore, we expect that gender stereotypes would be particularly likely to act as heuristics to identifying tweets that seem concerning or fake, leading to gender gaps in flagging probabilities based on the gender-congruency of the topic, as has been demonstrated by previous studies (Bordalo et al 2019; Coffman et al 2021b).

[11] Our predictions focus on flagging, rather than liking and retweeting, for a number of reasons. Access to flagging is less salient on Twitter than either liking or retweeting, and therefore requires a more thoughtful and deliberate action. Furthermore, liking is likely confounded by the attractiveness of the profile image. Assessing the relationship between the poster's physical appearance and the user's reaction to their post is beyond the scope of this study. Flagging, on the other hand, is clearly about the statement in the post, rather than the profile image, since no profile image in our database is in any way offensive or concerning. For retweeting, the reaction is ambiguous; furthermore, it is unclear why one might choose to 'retweet' a post in an abstract setting of an experiment. Figure 3 demonstrates that, while flagging is significantly reduced for true posts, the differences for liking and retweeting true and false posts are much smaller. The SI appendix Table D1 presents the results from the main analysis for these two outcomes.

Finally, we pre-registered testing for heterogeneous effects by user gender (1=female; 0=male) and falseness of tweet information (1=fake; 0=true), being *ex ante* agnostic as to the direction of these heterogeneities.[12]

3.2 Empirical approach

The main empirical strategy is to compare our primary outcome of *flagging* of the same tweet posted by a male- and a female-presenting poster by men and women, conditional of the veracity of the tweet, controlling for a preregistered set of demographic characteristics. Specifically, we estimate the linear probability model for the sample as a whole and breaking the sample up by gender of the subject:

$$\Pr(Y_{it}) = \beta_1 FemPost_{it} + \beta_2 Maleness_t + \beta_{12} FemPost \times Maleness_{it} + \\ + \gamma_1 FemUser_i + \gamma_2 False_t + \gamma_{12} FemUser \times False_{it} + X_i'\theta + \varepsilon_{it} \quad (1)$$

where Y_{it} represents flagging of tweet t by subject i; $FemPost_{it}$ takes on a value of 1 if tweet t in front of subject i shows a female-presenting poster profile image; $Maleness_{it}$ is the average stereotype of the topic of tweet t (ranging from -0.5 = fully female to 0.5 = fully male); $FemPost \times Maleness_{it}$ is the interaction of the two treatment variables; $FemUser_i$ is 1 if the subject is female and 0 if male (note that users reporting other gender are excluded); $False_t$ takes on a value of 1 if tweet t is false, and X_i is a set of preregistered controls for race, Hispanic ethnicity, age, education, and employment status. For simplicity of interpretation, we exclude the 14 subjects who indicated other gender and keep only the observations with female- or male-presenting profile images (see SI appendix E1 for robustness checks when we keep these observations, omit demographic controls, and estimate the probit model for binary choice variables). As a result, the sample consists of 11,666 observations at the tweet-subject level. Standard errors are clustered at the tweet level.

4 Results

4.1 Posts by female posters are less likely to be flagged overall

We begin by estimating Eq (1) regardless of topic domain, in order to test Hypothesis 1a and Hypothesis 1b. Column 1 of Table 2 presents the findings. First, we observe that posts randomly accompanied by a female-presenting poster are actually about 1pp less likely to be flagged than posts accompanied by a male-presenting profile image. This relatively small but significant average effect persists after controlling for whether the tweet is false (Column 2) and after we include controls for the gender stereotype associated with the topic (Column 3). Thus, on average, we find evidence against Hypothesis 1a which posits that posts by men and women would be equally likely to be flagged.

[12] SI appendix Table E8 estimates the long model favored by Muralidharan et al. (2023) with triple interactions between gender of poster, gender of user, and the maleness index. The results from that analysis are consistent with our preferred specifications.

[Table 2 about here]

Consistent with our prediction (Hypothesis 1b), Column 2 of Table 2 further reveals no gender differences in flagging of false tweets: men and women are both 4pp more likely to flag false posts than true posts.

4.2 Men are more likely to flag posts by women on male-stereotyped topics

So far, we have established that, on average, female posters enjoy an advantage over male posters. Our main question is whether the effect of poster gender varies by the stereotype associated with the post's topic. Hypothesis 2 predicts that women would be more likely to be flagged in male-domains, and that the gender of the user reading the tweet would play a role. Columns 4 and 5 of Table 3 estimate Eq (1) by gender of user, providing support for this hypothesis.

While female users are equally likely to flag posts by men and women, regardless of topic domain (Column 4), the probability of being flagged by a male user increases significantly as the topic of the post becomes more male-stereotyped (Column 5). In fact, as the maleness index increases from -0.5 (fully female-typed) to 0.5 (fully male-typed), a female poster's probability of being flagged by a male user rises by 6pp. This is a large effect in magnitude relative the 11% flagging probability for the omitted category. The interaction has the opposite effect among female users, although it is not statistically significant.[13]

We conclude that the overall advantage for female posters (Columns 1-3) derives from both men and women flagging female posters relatively less than male posters when the topic of the post is stereotypically female-oriented. On the other hand, men are likely to dislike or mistrust the posts authored by women on male-stereotyped topics (Column 5).

4.3 Potential mechanisms

Following our pre-analysis plan, we test a number of potential explanations for the observed pattern of flagging behavior. First, we consider the possibility that men may get more outraged upon encountering false information from women than from men, if the topic is perceived as male-typed. Panel A of Table 3 estimates our main specifications from Columns 4 and 5 of Table 2, breaking up the sample of posts into true and false statements. We find that women are more likely than men to have their tweets flagged by men as the male stereotype of the topic increases, regardless of whether their statements are true or false. The effect is actually more pronounced for true tweets, although difference between true and false tweets is not statistically significant.

[Table 3 about here]

Second, individuals with conservative views might be more prone to adhering to and enforcing traditional gender stereotypes. Panel B of Table 2 estimates the effect of poster gender interacted with topic stereotype separately for users who self-identify as conservative or very conservative

[13] In accordance with our pre-analysis plan, we consider the effects of potential moderator covariates: ideology (liberal v. conservative), experience with Twitter, religion, and past experience with misinformation. SI appendix table D1 shows that the interaction term decreases slightly in magnitude but remains large and significant.

and those who identify as liberal or very liberal. We find that flagging behavior does not significantly differ across the political spectrum. Both conservative and liberal men demonstrate a higher propensity to flag posts by women than posts by men as our maleness index rises.

Third, it is possible that the effects are concentrated among users less accustomed to social media discourse. Panel C of Table 3 delves into the impact of the poster's gender and topic domain on flagging behavior of users who use Twitter (either have an account or at least browse) and users who never use Twitter. The results reveal that familiarity with the Twitter platform has no heterogenous impact on the results.

Having not found support for these alternative mechanisms, we are left with the conclusion that the disproportionate mistrust displayed by men toward women is rooted in the unconscious bias against women voicing opinions on male-stereotyped topics.

5 Field Data Parallels

The experimental results suggest that women are mistrusted by men when making statements on male-stereotypical topics. This section explores whether this pattern is consistent with what we observe in field data. To that end, we leverage an existing dataset collected by Allen et al (2022) from "Community Notes" (formerly called "Birdwatch"), Twitter's crowdsourced fact-checking product. The dataset consists of all notes and ratings of tweets over the period between 1/28/21 and 6/29/21, as well as additional information on the characteristics of raters ("Birdwatchers"), including gender. The original dataset comprises 28,700 tweet x Birdwatch note observations, across a total of 2,910 unique tweets.

Community Notes are used to provide fact-check information for tweets that the Birdwatchers believe to be potentially misleading (89.6% of notes classified the tweet as potentially misleading). Within the sample of misleading tweets, we inquire whether the Birdwatchers flagged them as considerably harmful. In the data, the two possible classifications of harm are "considerable" and "little." We further restrict the sample to individual, rather than institutional, Birdwatchers and to tweets that are posted by individuals rather than institutions or media outlets, such as CNN, in order to be able to identify their gender. We also exclude tweets that have no text, only contain links, only contain images, or are not in English (so we can identify tweet topic). Finally, we exclude prominent tweeters that appear very frequently in the dataset, as Birdwatchers may have strong associations with these posters beyond their gender.[14] The resulting final sample contains 1,944 unique tweets, 1,325 posted by men and 619 by women.

We examine whether the perceived level of harm varies by poster gender. Overall, tweets by women are 6 pp more likely to be flagged as harmful than tweets by men (p-value 0.002). Heterogeneity analysis by gender of the Birdwatcher reveals that female Birdwatchers are 13pp more likely to flag posts by female tweeters than by male tweeters, while male Birdwatchers are only 4pp more likely to do so (see SI appendix Figure E1). These results are robust to the inclusion

[14] In particular, we drop tweets by Lauren Boebert, Candace Owens, POTUS, Donald Trump Jr., Jim Jordan, AOC, Alex Berenson, Ted Cruz, Marjorie Taylor Greene, Aaron Rupar, Jack Posobiec. Including these tweets and corresponding notes for these sources does not substantively change the results.

of controls for indicators for age, follower counts, and partisanship scores for tweeters and Birdwatchers, as well as an indicator for Birdwatcher notes being rated as helpful by third-party raters (see Tables E9 and E10 in the SI appendix). [15]

Of course, differences by Birdwatcher gender might be driven by selection. First, the sample of Birdwatchers is considerably male-skewed (only 25 percent of the sample of Birdwatchers is identified as female). Second, female Birdwatchers may be more likely to notice and follow female tweeters, and vice versa.

Finally, we address the question of whether Birdwatchers' flagging behavior varies by gender stereotype of topic domain. This examination draws a parallel between our experimental findings and field data. To achieve this, we follow the topic classification algorithm similar to the one used in our main study, albeit simplified due to the larger number of tweet observations.[16] The algorithm codes an average topic score (maleness, as in the experimental data) which assigns a value of 0-10 to each tweet's domain based on gender stereotypes with 10 being the most masculine.

Table 4 estimates the effect of poster gender on the probability of tweet being flagged as harmful, as the topic domain varies from least to most masculine. Our preferred specification that mirrors the analysis in Table 3 includes controls for Birdwatcher age, gender, partisanship status and clusters standard errors at the tweeter and Birdwatcher level. On average, we find that regardless of Birdwatcher gender, women's tweets are significantly more likely to be flagged as the maleness of the topic increases. This result is broadly consistent with our experimental findings and is robust to alternative specifications (see online SI appendix E2).

[Table 4 about here]

6 Discussion

Overall, our results indicate that women are judged more harshly than men when sharing information on topics that are perceived to be male-typed on social media. We find this pattern both in the field data and in our experimental study. This is consistent with the literature on the effect of stereotypes on gender gaps in other contexts (Bordalo et al 2019).

Our experimental design allows us to rule out a number of alternative explanations for the gender difference. For example, in actual social media interactions, women may share information on different topics and may phrase their posts differently than men, which may contribute to differential reactions to their posts. But these channels are shut down in our study: even though the posts are worded identically, women are flagged more than men for statements on male-typed

[15] For details of variable construction, refer to Allen et al. (2022). Gender and age are predicted using the M3Model described in Wang et al. (2019). Categories for age are <=18, 19-29, 30-39, >=40; Gender is coded as female and not female. Partisanship is inferred using accounts the user follows, using the method from Barberá et al. (2015) and coded on a [-2.5,2.5] scale, with more positive values indicative of greater affinity for the Republican party and negative values for the Democratic party.

[16] Note that we also used ChatGPT to code the "tone" for each tweet. However, the vast majority of tweets received a neutral classification and there was not enough variation in this variable to use it for analysis. Refer to the online SI appendix H for details.

topics (by male users) and overall less than men on female-typed topics. Differences in user ideology, familiarity with Twitter, and religiosity do not explain these effects. We also find no evidence that the gender gap vanishes when the post is truthful.

The finding that women are mistrusted particularly when making statements on male-typed topics carries implications beyond the realm of social media. Being perceived as credible is crucial for success across various fields, including the labor market, politics, and everyday social interactions. It is important for performance: lawyers must convince juries; start-up CEOs must convince venture capital investors; academic researchers must convince reviewers and journal editors. Conditional on performance being imperfectly observable, workers must also convince their managers that their efforts deserve recognition when applying for pay raises and promotions. Therefore, doubts about credibility of women's statements could directly contribute to the large gender gaps in law, politics, STEM, and other male-dominated fields. Because the inherent truthfulness (or falsehood) of statements does not seem to affect the level of mistrust, our findings point to the need for a more fundamental change in gender stereotypes about topic domains.

REFERENCES

Abeler, Johannes, Daniele Nosenzo, and Collin Raymond. 2019. "Preferences for Truth-Telling." *Econometrica* 87 (4): 1115–53.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31 (2): 211–36.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The Welfare Effects of Social Media." *American Economic Review* 110 (3): 629–76.

Allen, Jennifer, Cameron Martel, and David G. Rand. 2022. "Birds of a Feather Don't Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter's Birdwatch Crowdsourced Fact-Checking Program." Paper presented at the (hybrid) CHI Conference on Human Factors in Computing Systems, New Orleans, LA.

Almenar, Ester, Sue Aran-Ramspott, Jaume Suau, and Pere Masip. 2021. "Gender Differences in Tackling Fake News: Different Degrees of Concern, but Same Problems." *Media and Communication* 9 (1): 229-38.

Arin, K. Peren, Deni Mazrekaj, and Marcel Thum. 2023. "Ability of Detecting and Willingness to Share Fake News." *Scientific Reports* 13 (1): 7298.

Barberá, Pablo, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. "Tweeting from Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science* 26 (10): 1531–42.

Beltran, Javier, Aina Gallego, Alba Huidobro, Enrique Romero, and Lluís Padró. 2021. "Male and Female Politicians on Twitter: A Machine Learning Approach." *European Journal of Political Research* 60 (1): 239-51.

Bohren, J. Aislinn, Alex Imas, and Michael Rosenberg. 2019. "The Dynamics of Discrimination: Theory and Evidence." *American Economic Review* 109(10): 3395–436.

Bordalo, Pedro, Katherine B. Coffman, Nicola Gennaioli, and Andrei Shleifer. 2019. "Beliefs about Gender." *American Economic Review* 109 (3): 739–73.

Buchanan, Tom. 2020. "Why Do People Spread False Information Online? The Effects of Message and Viewer Characteristics on Self-Reported Likelihood of Sharing Social Media Disinformation." *PLoS One* 15 (10): e0239666.

Coffman, Katherine B., Christine L. Exley, and Muriel Niederle. 2021a. "The Role of Beliefs in Driving Gender Discrimination." *Management Science* 67 (6): 3551-69.

Coffman, Katherine B., Flikkema, Clio Bryant, and Olga Shurchkov. 2021b. "Gender Stereotypes in Deliberation and Team Decisions." *Games and Economic Behavior* 129 (1): 329-49.

Cohn, Alain, Maréchal, Michel André, David Tannenbaum, and Christian Lukas Zünd. 2019. "Civic Honesty Around the Globe." *Science* 365 (6448): 70–73.

Coppock, Alexander and Oliver A. McClellan. 2019. "Validating the Demographic, Political, Psychological, and Experimental Results Obtained from a New Source of Online Survey Respondents." *Research & Politics* 6 (1): 1-14.

Crawford, Kate, and Tarleton Gillespie. 2016. "What is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint." *New Media & Society* 18 (3): 410–28.

Dabbous, Amal, Karine Aoun Barakat, and Beatriz de Quero Navarro. 2022. "Fake News Detection and Social Media Trust: A Cross-Cultural Perspective." *Behaviour & Information Technology* 41 (14): 2953−72.

Elder, Elizabeth Mitchell and Matthew Hayes. 2023. "Signaling Race, Ethnicity, and Gender with Names: Challenges and Recommendations." *The Journal of Politics*, 85 (2): 764-70. doi.org/10.1086/723820

Epstein, Ziv, Nathaniel Sirlin, Antonio Arechar, Gordon Pennycook, and David Rand. 2023. "The Social Media Context Interferes with Truth Discernment." *Science Advances* 9 (9).

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. "ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences* 120 (30).

Goldin, Claudia, Sari Pekkala Kerr, Claudia Olivetti, and Erling Barth. 2017. "The Expanding Gender Earnings Gap: Evidence from the LEHD-2000 Census." *American Economic Review* 107 (5): 110–14.

Hu, Lingshu, and Michael Wayne Kearney. 2021. "Gendered Tweets: Computational Text Analysis of Gender Differences in Political Discussion on Twitter." *Journal of Language and Social Psychology* 40 (4): 482-503.

Huddy, Leonie, and Nayda Terkildsen. 1993. "Gender Stereotypes and the Perception of Male and Female Candidates." *American Journal of Political Science* 37 (1): 119–47.

Lundberg, Shelly, and Jenna Stearns. 2019. "Women in Economics: Stalled Progress." *Journal of Economic Perspectives* 33 (1): 3–22.

Meagher, Benjamin R. 2017. "Judging the Gender of the Inanimate: Benevolent Sexism and Gender Stereotypes Guide Impressions of Physical Objects." *British Journal of Social Psychology* 56 (3): 537–60.

Michelmore, Kathering and Sharon Sassler. 2016. "Explaining the Gender Wage Gap in STEM: Does Field Sex Composition Matter?" *RSF: The Russell Sage Foundation Journal of the Social Sciences* 2 (4): 194–215.

Muralidharan, Karthik, Mauricio Romero, and Kaspar Wüthrich. 2023. "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments." *The Review of Economics and Statistics*, 1–44.

Pennycook, Gordon, and David G. Rand. 2020. “Who Falls for Fake News? The Roles of Bullshit Receptivity, Overclaiming, Familiarity, and Analytic Thinking.” *Journal of Personality* 88 (2): 185–200.

Pennycook, Gordon, and David G. Rand. 2022. “Accuracy Prompts are a Replicable and Generalizable Approach for Reducing the Spread of Misinformation.” *Nature Communications* 13: 2333.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. “Shifting Attention to Accuracy Can Reduce Misinformation Online.” *Nature* 592: 590–95.

Shurchkov, Olga. 2012. “Under Pressure: Gender Differences in Output Quality and Quantity under Competition and Time Constraints.” *Journal of the European Economic Association* 10 (5): 1189–1213.

Shurchkov, Olga, and Alexandra V.M. van Geen. 2019. “Why Female Decision-Makers Shy Away from Promoting Competition.” *Kyklos* 72 (2): 297–331.

United States Census Bureau. 2022. “U.S. Census Bureau QuickFacts: United States” https://www.census.gov/quickfacts/ (accessed August 1, 2023)

United States Census Bureau. 2023. “America is Getting Older." *United States Census Bureau,* Press Release Number CB23–106.

van der Does, Tamara, Mirta Galesic, Zackary Okun Dunivin, and Paul E. Smaldino. 2022. “Strategic Identity Signaling in Heterogeneous Networks.” *Proceedings of the National Academy of Sciences, USA* 119 (10): e2117898119.

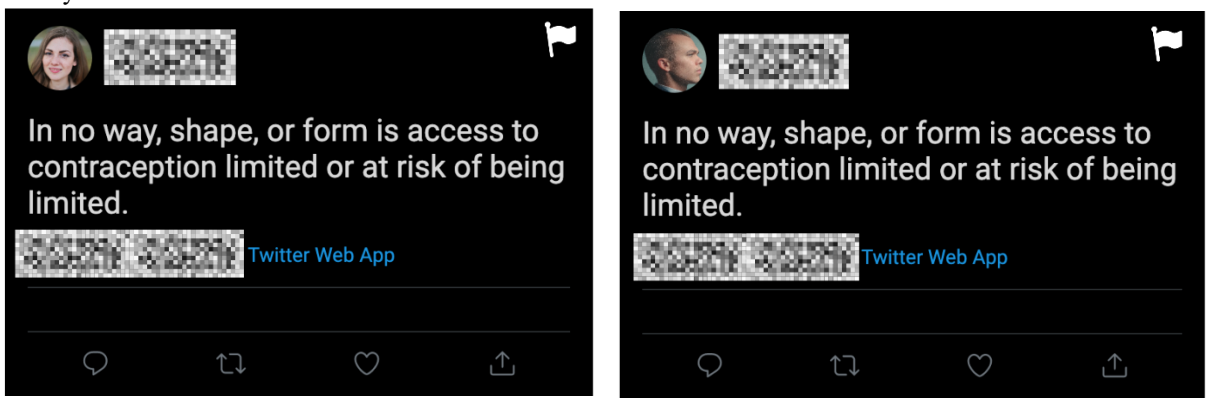
Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. “The Spread of True and False News Online.” *Science* 359 (6380): 1146–51.

Wang, Zihian, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. 2019. “Demographic Inference and Representative Population Estimates from Multilingual Social Media Data.” *WWW ’19: The World Wide Web Conference*, 2056–2067.

Zhu, Jane M., Arthur P. Pelullo, Sayed Hassan, Lillian Siderowf, Raina M. Merchant, and Rachel M. Werner. 2019. “Gender Differences in Twitter Use and Influence Among Health Policy and Health Services Researchers.” *JAMA Internal Medicine* 179 (12): 1726–29.

Figures

PANEL 1:
Topics: Health/Healthcare; Women's Rights; Politics (Maleness score = -0.17; female-stereotyped)
Veracity: False



PANEL 2:
Topics: Politics; International Relations; Economics/Finance (Maleness score = 0.5; male-stereotyped)
Veracity: True

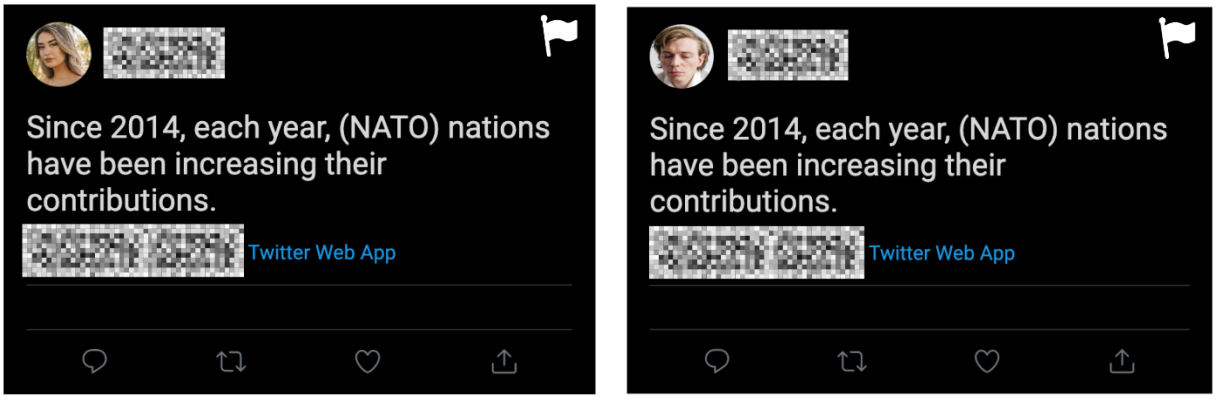


Figure 1: Example tweets with female-presenting (left) and male-presenting (right) profile photos

Tables

Table 1: Number of posts by gender of poster, topic domain, and truthfulness

Topic Domain	Female Poster		Male Poster		Non-Human Poster	
	True	False	True	False	True	False
Male Domain (Maleness >=0.25)	1,031	979	1,002	1,045	974	953
Female Domain (Maleness <0)	884	929	915	948	843	941
Neutral Domain (0<=Maleness<0.25)	1,066	977	1,114	986	1,155	981

Notes: Each cell records the number of tweet x subject observations by treatment arm. Tweet ID 47 was dropped from the analysis as a result of a coding error.

Table 2: Average effect of poster gender, post truthfulness, and topic domain on the probability of flagging

Dep. Var: Pr(Flagged Tweet)	(1)	(2)	(3)	(4)	(5)
User Gender	All	All	All	Female	Male
Female poster	-0.011**	-0.010**	-0.012**	-0.012	-0.011
	(0.005)	(0.005)	(0.006)	(0.007)	(0.009)
Female user		-0.023***	-0.023***		
		(0.008)	(0.008)		
False tweet		0.037***	0.037***	0.051***	0.036***
		(0.010)	(0.010)	(0.011)	(0.010)
False tweet x Female user		0.013	0.013		
		(0.013)	(0.013)		
Maleness index			-0.037*	-0.012	-0.070***
			(0.020)	(0.022)	(0.024)
Maleness x Female poster			0.015	-0.022	0.063**
			(0.020)	(0.023)	(0.028)
Dep. var. mean	0.103	0.103	0.103	0.093	0.113
R-squared	0.053	0.059	0.060	0.050	0.100
No. observations	11,666	11,666	11,666	6,121	5,545

Notes: The sample is restricted to tweets with female or male-presenting profile images and participants who passed both attention checks and participants who did not choose other gender. Tweet id 47 was dropped from the analysis due to coding error. Maleness index ranges from -0.5 (fully female) to 0.5 (fully male). All specifications include demographic controls include age, indicators for race, ethnicity (Hispanic), education, and level of employment. Robust standard errors clustered at the tweet level. Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Heterogeneous effects of poster gender and topic domain on the probability of flagging, by truthfulness of post, political ideology, and Twitter/X use

Dep. Var: Pr(Flagged Tweet)	(1)	(2)	(3)	(4)
User Gender	Female	Male	Female	Male
Panel A:	False Posts		True Posts	
Maleness x Female poster	-0.008	0.031	-0.031	0.088**
	(0.028)	(0.042)	(0.035)	(0.038)
Dep. var. mean	0.120	0.134	0.067	0.093
R-squared	0.074	0.118	0.023	0.091
No. observations	3,044	2,719	3,077	2,826
Panel B:	Conservatives		Liberals	
Maleness x Female poster	-0.065	0.078*	0.036	0.079
	(0.048)	(0.040)	(0.063)	(0.060)
Dep. var. mean	0.078	0.095	0.134	0.112
R-squared	0.082	0.078	0.096	0.147
No. observations	1,596	1,995	1,385	1,931
Panel C:	Twitter User		Twitter Non-User	
Maleness x Female poster	-0.019	0.059*	-0.027	0.070
	(0.029)	(0.031)	(0.035)	(0.074)
Dep. var. mean	0.094	0.078	0.091	0.201
R-squared	0.063	0.037	0.071	0.151
No. observations	3,525	3,975	2,596	1,570

Notes: The sample is restricted to tweets with female or male-presenting profile images and participants who passed both attention checks and participants who did not choose other gender. Tweet id 47 was dropped from the analysis due to coding error. Maleness index ranges from -0.5 (fully female) to 0.5 (fully male). All specifications include the level effects of poster gender, maleness, post being false, and the interaction of false and female poster. Columns 1 and 2 of Panel A estimate the effects for false posts; Columns 3 and 4 estimate the effects for true posts. Columns 1 and 2 of Panel B estimate the effects for subjects who self-identified as conservative or very conservative; Columns 3 and 4 estimate the effects for subjects who self-identified as liberal or very liberal. Columns 1 and 2 of Panel C estimate the effects for subjects who self-identified as having a Twitter/X account and/or browsing Twitter; Columns 3 and 4 estimate the effects for those who do not use or browse it. All specifications include demographic controls include age, indicators for race, ethnicity (Hispanic), education, and level of employment. Robust standard errors clustered at the tweet level. Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Average effect of poster gender and topic domain on the probability of flagging (field data)

Dep. Var: Pr(Flagged Tweet)	(1)	(2)	(3)
Birdwatcher Gender	All	Male	Female
Female poster	-0.232	-0.216	-0.202
	(0.154)	(0.172)	(0.271)
Maleness index	-0.010	-0.015	0.018
	(0.015)	(0.018)	(0.024)
Maleness x Female poster	0.051**	0.047	0.056
	(0.026)	(0.029)	(0.047)
Controls	Yes	Yes	Yes
Dep. var. mean	0.723	0.728	0.714
R-squared	0.370	0.383	0.393
No. observations	13,917	11,179	2,738

Notes: The sample is restricted to Birdwatchers whose gender could be identified which excludes organizations, tweets posted by individuals excluding particularly famous and prolific authors and that contain legible text in English. Controls include Birdwatcher gender, indicators for age, follower counts, status counts, and partisanship scores for tweeters and Birdwatchers. Controls also include an indicator for Birdwatcher notes rated as helpful by third-party raters. Robust standard errors clustered at the tweeter and Birdwatcher level in parentheses. Significance: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.