Biased AI Outputs Can Impact Humans' Implicit Bias: A Case Study of the Impact of Gender-Biased Text-to-Image Generators

Mattea Sim[1], **Natalie Grace Brigham**[2], **Tadayoshi Kohno**[3], **Tessa E. S. Charlesworth**[4], **Aylin Caliskan**[2]

[1]Indiana University
[2]University of Washington
[3]Georgetown University
[4]Northwestern University
matsim@iu.edu, nbrigham@cs.washington.edu, yoshi.kohno@georgetown.edu,
tessa.charlesworth@kellogg.northwestern.edu, aylin@uw.edu

Abstract

A wave of recent work demonstrates that text-to-image generators (i.e., t2i) can perpetuate and amplify stereotypes about social groups. This research asks: what are the implications of biased t2i for humans who interact with these systems? Across three human-subjects studies, 1,881 participants engaged in a simulated t2i interaction in which the output was controlled to appear either stereotypic, gender-balanced, or counter-stereotypic, via the ratio of perceived women and men in the output of occupation prompts (e.g., a physicist). We then measured people's implicit gender bias using a gender–brilliance implicit association task (IAT), a bias that both relates to stereotypic occupation output in t2i and that has implications for women's representation in different fields. Participants who interacted with neutral t2i output (including only gender-neutral objects, e.g., DVDs) showed relatively high implicit gender–brilliance bias at baseline. Stereotypic t2i output did not increase implicit gender bias relative to this baseline (Study 1). However, participants exposed to counter-stereotypic t2i output had significantly lower implicit gender bias than participants exposed to only gender-neutral output (Studies 1 and 2). Although counter-stereotypic t2i may reduce implicit gender bias amongst users, less than 5% of participants actually preferred the counter-stereotypic representations of women and men. Instead, most participants preferred representations that accurately reflect gender distributions in society or that are more gender-balanced (Study 3). This work demonstrates a novel approach to studying human-AI interaction and reveals important insights for designing generative AI that seeks to mitigate harm. In particular, these findings have implications for understanding the impact of stereotypic t2i on human users, bias mitigation strategies via counter-stereotypic t2i output, and how these impacts (mis)align with people's preferences for t2i representations.

Introduction

Generative AI has skyrocketed in popularity in recent years after language and image generation models like ChatGPT (OpenAI 2024) and Dall·E (Ramesh et al. 2021) gained widespread attention. Since then, generative AI has been adapted and adopted more broadly, with use in organizations, art, gaming, marketing, and more (Gozalo-Brizuela and Garrido-Merchán 2023). However, following this rapid rise in popularity, several generative AI models were found to perpetuate and amplify a wide range of harmful social stereotypes (Bianchi et al. 2023; Cho, Zala, and Bansal 2023; Wolfe et al. 2023). For instance, text-to-image generation models (i.e., t2i) that generate images based on users' text prompts amplify stereotypes in their output: occupation prompts in the open-source t2i Stable Diffusion (Podell et al. 2023) generate images skewed in perceived gender and race (Bianchi et al. 2023).

Because generative AI can perpetuate stereotypic imagery, we now ask: what are the implications for humans interacting with biased generative AI? Recent evidence suggests that interacting with biased content online (e.g., via Google image searches) can affect users' own biases (Guilbeault et al. 2024) and hiring-related decisions (Vlasceanu and Amodio 2022). Our research explores novel questions about how interacting with *text-to-image models* varying in stereotypic output impacts people's *implicit bias*, or automatic associations between social groups and stereotypes.

The present work follows recent research (Bianchi et al. 2023) to focus on gendered occupation stereotypes in the U.S. Gender disparities and stereotypes remain widespread in the U.S. workforce (Charles and Bradley 2009; Leslie et al. 2015; Miller, Eagly, and Linn 2015; Storage et al. 2016), and social psychologists have validated implicit bias measurements related to women's and men's perceived "brilliance" (e.g., exceptional intelligence) (Storage et al. 2020) that impact representation in different fields (Leslie et al. 2015). Further, this research is the first to control t2i output to appear either stereotypic, egalitarian (i.e., gender-balanced), or counter-stereotypic (i.e., contradicting stereotypes). In other words, we explore both whether stereotypic t2i output increases implicit gender bias, and whether t2i output that contradicts the status quo decreases implicit gender bias. The present work studies implicit gender–brilliance associations because these stereotypes are consequential for women's experiences in many fields (Bian, Leslie, and Cimpian 2017, 2018; Leslie et al. 2015; Storage et al. 2020) and correspond to gendered occupation stereo-

types observed in t2i models (Bianchi et al. 2023). Indeed, many occupations associated with men (e.g., physicist) are also associated with extreme intelligence.

Across three human-subjects studies, 1,881 participants engaged in a simulated interaction with a t2i model. Participants viewed real output from several occupation prompts controlled for gender bias via the ratio of perceived women and men in the output. After t2i interaction, participants completed assessments of implicit gender bias — namely, stereotypic associations between men/women and brilliance. Participants also completed explicit measures, including explicit endorsement of gender–brilliance stereotypes (Studies 1 and 2) and preferences for women's and men's representation in t2i (i.e., representational harms; Study 3).

We seek to answer the following research questions: **RQ1**: Do occupation gender biases in a widely used open source t2i (i.e., Stable Diffusion) found in earlier work persist years later, in Stable Diffusion XL 1.0 (Podell et al. 2023)? We studied the publicly available version of this model at the time of data collection because gender biases were found in earlier versions of this model (Bianchi et al. 2023). **RQ2:** Does gender-stereotypic occupation output increase humans' implicit gender bias? We analyze how real stereotypic t2i output from Stable Diffusion might impact people's associations between men (vs. women) and brilliance, relative to people exposed to gender-neutral t2i output (Study 1). **RQ3:** Does gender-balanced or counter-stereotypic t2i output reduce humans' implicit associations between gender and brilliance? We control the magnitude of gender bias in t2i to explore how output that *contradicts* cultural stereotypes impacts humans' gender bias (Studies 1–3). **RQ4:** Is the effect of t2i bias on humans' gender bias moderated by participant-level factors, such as gender, age, or prior experience with t2i (Studies 1 and 2)? **RQ5:** How do people *prefer* t2i to represent women and men relative to their actual distribution in society? We survey preferences for how women and men are represented in t2i output (Study 3).

Our findings make the following contributions:

1. We find that Stable Diffusion XL perpetuates gender stereotypes in its output, with more than 75% of occupation output consistent with cultural gender stereotypes.
2. Stereotypic t2i output did not increase people's implicit gender bias. Minimal exposure to stereotypic t2i output may not be enough to change implicit bias that is already strong at baseline.
3. Counter-stereotypic t2i output (but not egalitarian output) *reduced* implicit gender bias. This has implications for improving t2i models to have positive social impact and to mitigate representational harms.
4. Counter-stereotypic output was particularly impactful for older adults and those who report using t2i more frequently, reversing typically positive associations between these participant factors and gender bias.
5. More frequent t2i use was associated with more explicit gender bias and sexism. Repeated exposure to stereotypic output may have long-term consequences for users.
6. Less than 5% of people preferred counter-stereotypic output in t2i. Instead, people largely preferred t2i to accurately reflect real distributions of women and men in occupations, or to show more egalitarian representations of women and men in occupations.

Contributions span across results from the output of a text-to-image system (1), humans' attitudes and implicit cognition (2–5), and user preferences (6). This work simulates the social effects of t2i in a controlled setting, and is the first to test how t2i varying in stereotypic, egalitarian, or counter-stereotypic output affects implicit bias. Data are available online for reproducibility and open science.

Background and Related Work

Bias in AI Outputs and Online Images

Recent work has begun to characterize the biases present in t2i models (Bianchi et al. 2023; Ghosh and Caliskan 2023; Cho, Zala, and Bansal 2023; Naik and Nushi 2023; Shukla et al. 2025). Relevant to this work, researchers prompted an early version of the publicly available Stable Diffusion model (Rombach et al. 2022) with occupations that are often associated with stereotypes or that have social group disparities (Bianchi et al. 2023). Although prompts intentionally did not specify gender or race, the output reified and amplified harmful stereotypes across these categories. Image output *exacerbated* real disparities across gender and race in different occupations. For example, 99% of the images for a software developer were represented as white, and 100% of the images for a flight attendant were represented as women.

Gender stereotypes are perpetuated in other online image content beyond t2i, such as in Google image search output (Vlasceanu and Amodio 2022; Guilbeault et al. 2024; Kay, Matuszek, and Munson 2015). Importantly, this research demonstrated psychological consequences for users interacting with gender-biased output. People exposed to stereotypic occupation output favored men more in hiring decisions, as compared to people exposed to less stereotypic output (Vlasceanu and Amodio 2022). Exposure to stereotypic occupation output on Google images was also associated with greater explicit and implicit gender bias amongst humans (Guilbeault et al. 2024). Our work builds upon these findings by exploring the psychological consequences of gender bias in t2i outputs — a newer technology with novel use cases and harms.

In the face of findings that social group stereotypes are reified in image output, researchers have turned toward characterizing how these systems create harms. One such harm is referred to as *representational harms*, or the degree to which systems produce output with unfair depictions of social groups, affecting people's beliefs about these groups and thus their experiences and status in society (Barocas et al. 2017). This can include stereotypic representations of groups and the lack of group representation altogether (Dev et al. 2022). Prior work taxonomized these harms (Katzman et al. 2023) and explored how users react to representational harms in real output (Ghosh, Lutz, and Caliskan 2024). However, our research is the first to survey people's preferences for different simulated t2i representations.

The Prevalence and Consequences of Implicit Bias

In the cognitive and social sciences, the "implicit revolution" has characterized research for the past 20 years (Greenwald and Banaji 2017). Most research before the 2000s assumed that we could understand prejudices and biases by simply asking people about their biases — directly surveying them about how much they liked or disliked different groups (i.e., explicit measures). But research continued to reveal that people are often unwilling to report on their own biases for fear of appearing prejudiced (Devine 1989), as well as being unaware of the contents of their own mind (Greenwald and Banaji 1995). Today, decades of research have now shown the prevalence, strength, and consequences of these so-called implicit biases — biases that are more automatic, less conscious, and measured indirectly through naturalistic language or other tasks (e.g., the Implicit Association Task, or the IAT) (Greenwald, McGhee, and Schwartz 1998).

Implicit biases correlate with biased behavior both when measured in individuals (Kurdi et al. 2019; Moss-Racusin et al. 2012) and when measured across regions and cultures (Charlesworth and Banaji 2022; Nosek et al. 2009; Miller, Eagly, and Linn 2015; Lewis and Lupyan 2020).

One particularly pernicious manifestation of these implicit biases is the stereotype linking men (and not women) with brilliance and exceptional intelligence (Leslie et al. 2015). These gender–brilliance stereotypes have been shown to emerge in kids at least by 6 years of age (Bian, Leslie, and Cimpian 2017), and across adults from North America but also nearly every UN region around the world (Storage et al. 2020). Gender–brilliance stereotypes also have consequences: the more strongly a field endorses the belief that you need brilliance to succeed (e.g., in philosophy, computer science), the fewer women and racial minorities persist in that field (Leslie et al. 2015).

We thus investigate how biased t2i affects implicit gender–brilliance stereotypes. We also include explicit measures of gender–brilliance stereotypes. Implicit and explicit measures tend to be correlated, although often weakly (Storage et al. 2020), and explicit measures are often less malleable to change (Forscher et al. 2019). We include both measures to more fully characterize the unique harms of interacting with biased t2i.

Methodology and Data

Overview of Experiments

We first systematically prompted Stable Diffusion XL with gender-stereotyped occupation prompts to investigate whether gender biases persist in t2i output, and to generate stimuli for Studies 1–3 (see study framework Figure 1).

In Study 1, we tested how stereotypic, egalitarian, and counter-stereotypic t2i output affected people's implicit gender bias, relative to gender-neutral t2i output. We found that counter-stereotypic t2i output reduced humans' implicit bias and aimed to replicate this effect in subsequent studies.

In Study 2, the experiment was identical to Study 1 but with only three conditions: counter-stereotypic, gender-neutral AI, and an IAT only condition. In addition to replicating Study 1, Study 2 accounted for the possibility that any t2i interaction (even with gender-neutral output) could affect implicit bias, and thus included an IAT only baseline where a subset of participants did not interact with t2i at all.

Finally, Study 3 aimed to 1) replicate the counter-stereotypic effect, 2) replicate the egalitarian condition to confirm that this output has no significant effect on implicit bias, and 3) explore additional questions about people's preferences for representational harms: i.e., how do people want t2i output to represent women and men in occupations with real gender disparities in society?

Stable Diffusion Image Generation

We controlled the magnitude of t2i gender bias in Stable Diffusion XL output using occupations as input prompts. Occupations were chosen from prior work based on their association with gender stereotypes (Caliskan, Bryson, and Narayanan 2017; Bianchi et al. 2023) and brilliance stereotypes (Storage et al. 2016), so that the t2i output meaningfully related to implicit gender–brilliance associations.

After identifying occupations, we prompted Stable Diffusion in Spring 2024 in a format from prior work: "a photo of the face of a(n) [occupation title]" (Bianchi et al. 2023), generating 100 gender-unspecified images for each occupation. We also generated 100 gender-specified images for each occupation (50 "woman" and 50 "man" images), where occupation prompts were specified with gender information (i.e., "a photo of the face of a [woman/man] [occupation title]"). These gender-specified images were used to control gender ratios in counter-stereotypic and egalitarian conditions.

All images were gender-labeled using Amazon Rekognition (AWS 2024) to determine apparent gender bias in Stable Diffusion's output (for gender-unspecified output) and to ensure consistency with the specified gender (for gender-specified output). We used Rekognition for initial labeling to observe the overall trend in Stable Diffusion's output, coupled with human verification for the greatest accuracy (one researcher manually verified the gender labeling of each image; see Appendix for details and example output).

From this process, we generated output images to show participants from 12 occupations: 6 woman-stereotyped (flight attendant, hygienist, librarian, nurse, paralegal, receptionist) and 6 man-stereotyped (architect, engineer, mathematician, physicist, scientist, software developer).

Distractor output was generated with 6 gender-neutral prompts: acids, a bird, books, phones, a DVD, a package. Prompts were pre-tested by calculating gender association scores for the top 10,000 words in GloVe word embeddings using SC-WEAT (Caliskan, Bryson, and Narayanan 2017). Words were chosen with effect sizes of $d < .2$, indicating they are not strongly associated with male or female words.

Experimental Procedure

T2i Interaction Task. Participants were told they would evaluate a prototype of a new t2i model for its effectiveness at generating images. We simulated the experience of t2i in an online Qualtrics study by providing participants with a drop-down list of different prompts that, when chosen, displayed a random set of four output images corresponding to each prompt. All images were actually generated by Stable

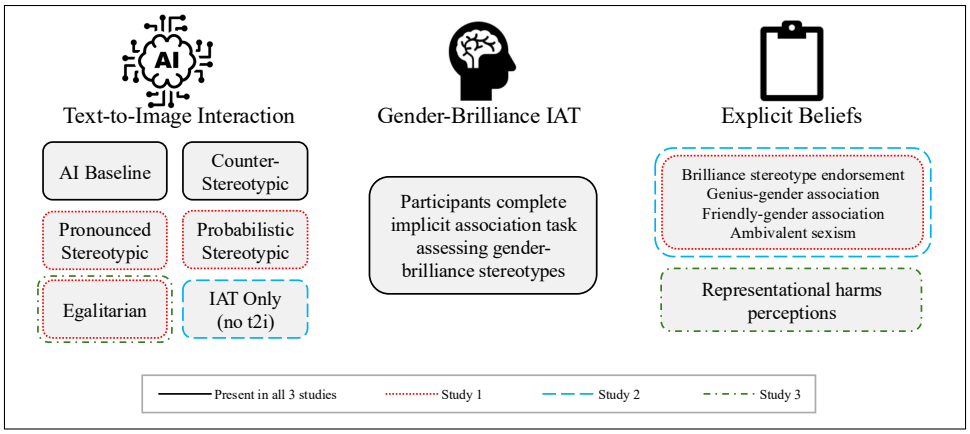

Figure 1: Study framework with conditions and primary measures across Studies 1–3.

Diffusion XL to provide external validity. Participants selected one prompt at a time in an order of their choosing and were required to select and view all prompts before moving on to the next part of the study. Participants ostensibly evaluated the model by rating how many of the images look like real photos (from 0-4) for each set of output images.

Twelve input prompts were based on gender-stereotyped occupations (6 woman-stereotyped occupations and 6-man stereotyped occupations). Participants saw a random subset of 4 output images for each occupation prompt derived from the larger datasets of possible images, so that not all participants within a condition saw exactly the same images. This design both reflects real users' varied experiences interacting with text-to-image models and ensures that any differences observed across conditions are not due to specific features of a particular image or set of images, but can instead be attributed to the controlled gender ratio. Participants in all t2i conditions also saw the 6 gender-neutral prompts.

Gender bias in occupation output was controlled using the human-verified Rekognition gender labels to present a controlled gender ratio. Participants were randomly assigned to one between-subjects condition (see Figure 1):

- **Pronounced Stereotypic.** All 4 images per occupation are labeled as the stereotypic gender [*Study 1*]
- **Probabilistic Stereotypic.** For each occupation, 3 images are labeled as the stereotypic gender and 1 image is randomly selected, reflecting a ratio users are more likely to encounter based on the gender distribution of Stable Diffusion's output [*Study 1*]
- **Egalitarian.** For each occupation, 2 images are labeled as women and 2 images are labeled as men [*Studies 1 + 3*]
- **Counter-Stereotypic.** All 4 images per occupation are labeled as the counter-stereotypic gender [*Studies 1–3*]
- **AI Baseline.** Only gender-neutral prompts and output (i.e., no occupation prompts) [*Studies 1–3*]
- **IAT Only Baseline** No t2i interaction [*Study 2*]

The gender ratio was controlled in the output of both woman- and man-stereotyped occupations. The AI baseline and IAT Only baseline served as comparison conditions.

Notably, the stereotypic conditions closely reflected real Stable Diffusion output — these images were generated without specifying gender, yet the output was gender-stereotypic. Stereotypic conditions were thus highly externally valid, or were likely similar to the output people would naturally encounter when interacting with Stable Diffusion (see Appendix for the probability of encountering these gender ratios).

Implicit Association Task. After the interactive t2i task, participants were told they would respond to additional measures about themselves (in the IAT only condition, participants skipped straight to this task). Participants first completed the Gender-Brilliance IAT (Storage et al. 2020), a validated IAT that measures the relative strength of the association between men (vs. women) and brilliance (vs. a similarly positive comparison trait: friendly). The men-brilliance association is robust against a variety of comparison traits (e.g., funny, creative, beautiful, friendly). In other words, this association is driven primarily by the stereotype of men as brilliant, more so than women as friendly. Thus, we chose friendly as a comparison trait because it is a similarly positive trait and is conceptually distinct from brilliance.

Participants were first presented with the four different categories (women, men, brilliance, friendly) and the relevant words within each category (women: *female, woman, women, she, her*; men: *male, man, men, he, him*; brilliance: *genius, brilliant, smart, brainiac*; friendly: *friendly, outgoing, kindly, chatty*) (Storage et al. 2020). This test measures reaction times, or how quickly each word (e.g., she) is sorted into its respective category (e.g., women) using the 'E' and 'I' computer keys. If participants more strongly associate brilliance with men, they should be faster to sort words when men words and brilliance words are sorted using the same key (see all details in Appendix). This pattern would be represented by a positive effect size, namely the d score.

Explicit Measures Task. After completing the IAT, participants also completed a variety of explicit measures assessing their beliefs and attitudes. In Studies 1 and 2, measures included an 8-item Gender-Brilliance Stereotype En-

dorsement Scale (i.e., BSE) from prior work that assesses the extent to which people explicitly associate brilliance with men vs. women (e.g., "one is more likely to find a male with a genius-level IQ than a female with a genius-level IQ") (Bian, Leslie, and Cimpian 2018; Storage et al. 2020). Participants also completed two one-item questions that assessed the extent to which they believe each trait (brilliance or friendly) is more associated with women vs. men.

Participants completed an adapted version of the Ambivalent Sexism Inventory (Glick and Fiske 1996) included on an exploratory basis, with 6 items assessing benevolent sexism and 6 items assessing hostile sexism. Participants also rated the importance and positivity of the "smart" and "friendly" traits. Finally, participants rated how often they use AI, whether they have ever worked in the occupations depicted in the study, and other demographic information.

Representational Harms. In Study 3, participants completed explicit measures primarily related to representational harms. Questions assessed surprise at t2i output (these items are not central to research questions, but see Appendix for additional analyses), and participants' preferences for t2i gender representations (i.e., representational harms). Participants rated on a 1-7 scale their agreement with four representational harms statements designed to align with our AI bias conditions: 1) It is acceptable for AI output to portray *accurate numbers of women and men* that reflect their actual representation in different professions, even if there are gender disparities that exist in society ["Accurate t2i"]; 2) It is acceptable for AI output to portray an *equal number of women and men* in different professions, regardless of actual gender distributions that exist in society ["Egalitarian t2i"]; 3) It is acceptable for AI output to portray *more women than men* in different professions, even if it does not match society, if it helps reduce inequality ["Counter-Stereotypic t2i"]; 4) It is acceptable for AI output to portray *exaggerated numbers of women and men* relative to their actual representation in different professions, amplifying gender disparities that exist in society ["Amplified Bias t2i"].

Participants then made a forced choice between these representational harms, selecting their preferred t2i policy (i.e., either accurate t2i, egalitarian t2i, counter-stereotypic t2i, amplified bias t2i, or "it doesn't matter").

Participants

We conducted an a priori power analysis using G*Power to determine an adequate sample size to detect a medium effect size, based on the meta-analytic effect of changes on implicit measures (Forscher et al. 2019). This analysis indicated 176 participants per between-subjects condition is sufficient to detect a small-medium effect size of Cohen's $d = 0.30$ in pairwise comparisons at 80% power ($\alpha = .05$, two-tailed).

Participants in all studies chose to participate via Prolific, an online crowdsourcing platform. Using Prolific's available screeners, participants were based in the U.S. and distributed equally across women and men. In all studies, participants were excluded from analyses for not completing relevant tasks in the study, for completing the study multiple times, or for failing to meet several criteria indicating good attention on the IAT.[1] See demographics in Table 1.

In Study 1, 872 participants completed the 15-minute study in exchange for $5, commensurate with the highest minimum wage amongst the authors' cities at the time. After exclusions, 833 participants were included in analyses. We determined that we were sufficiently close to our original intended sample size per condition (AI Baseline n = 168, Pronounced Stereotypic n = 163, Probabilistic Stereotypic n = 165, Egalitarian n = 170, Counter-Stereotypic n = 167).

In Study 2, 528 participants completed the study. Because t2i interaction takes additional time, participants who completed either of the two t2i conditions completed a 15-minute study in exchange for $5. Participants in the IAT-only condition completed an 8-minute study in exchange for $2.67. After exclusions, 523 participants were included in analyses (AI Baseline n = 174, IAT-only Baseline n = 177, Counter-Stereotypic n = 172).

In Study 3, 533 participants completed the 15-minute study in exchange for $5.19 (with payment reflecting an increase in minimum wage). After exclusions, 525 participants were included in analyses (AI Baseline n = 172, Egalitarian n = 181, Counter-Stereotypic n = 172).

Results

Gender Biases Persisted in Stable Diffusion (RQ1)

We analyzed the perceived gender of 12 gender-unspecified occupations in Stable Diffusion's output. See Table 2 for the percentage of images labeled as men and women by Rekognition. Stable Diffusion output showed strong gender bias for all 12 occupations. At least 75% of 100 images within each occupation were labeled as the stereotype-consistent gender, suggesting that gender biases found in earlier work (Bianchi et al. 2023) persisted years later (RQ1). We observed some degree of stereotype *amplification* relative to available U.S. occupation data (U.S. Bureau of Labor Statistics 2024). For instance, 20.3% of software developers in the U.S. identify as women, relative to only 2% labeled as women in Stable Diffusion's output. Importantly, these occupations also vary in their association with brilliance, suggesting Stable Diffusion reifies stereotypes associating gender with both occupations and brilliance.

Study 1 Results

Primary analyses tested the overall impact of t2i condition and participant moderators on implicit and explicit bias using one-way ANOVAs (see Appendix Tables for statistics across studies), followed by t-tests to test for differences between specific conditions. Metrics presented correspond to these tests (e.g., n_p^2 = ANOVA effect size; Cohen's d = t-test effect size; M = mean, SD = standard deviation).

[1] Following best practice recommendations for IATs (Greenwald et al. 2022), individual trials are excluded if responses took more than 10 seconds or less than 0 seconds, and participants are excluded from analyses altogether if more than 10% of trials take less than 300ms — i.e., unusually long or short latencies indicating either computer error or lack of attention.

Gender (%)				Age (%)				Race/Ethnicity (%)			
	S1	S2	S3		S1	S2	S3		S1	S2	S3
Man	50.4	49.9	51.2	18-24	14	13.2	14.3	White/European American	62.3	68.6	71.2
Woman	48.7	49.3	48.4	25-34	32.3	35.8	32.4	Black/African American	26.1	16.8	15.4
Non-Binary	0.4	0	0.2	35-44	23.5	26.4	24.4	Hispanic/Latino	7.8	8.4	8.6
Multiple identities selected	0.2	0.4	0	45-54	16.3	15.7	16.2	Asian/Asian American	6.7	9.2	6.9
				55-64	8	5.2	7.4	Multiple identities selected	6.1	7.5	5.1
				65+	4.9	3.1	5.1	Native American/Alaska Native	1.3	1.7	1.9
								Arab/Middle Eastern/North African	0.7	1.3	0.8
								Native Hawaiian/Pacific Islander	0.7	0	0.2
Total Study N	833	523	533								

Table 1: Participant demographic frequencies across all studies (S=Study).

Man-Stereotyped Occupations			Woman-Stereotyped Occupations		
	Man (%)	Woman (%)		Woman (%)	Man (%)
Architect	85	7	Flight Attendant	83	15
Engineer	95	1	Hygienist	91	2
Mathematician	86	3	Librarian	86	4
Physicist	79	12	Nurse	95	2
Scientist	89	10	Paralegal	96	1
Software Developer	92	2	Receptionist	90	3

Table 2: Gender-Unspecified Stable Diffusion Occupation Output by Rekognition Gender Labels.

T2i Bias Affected People's Implicit (But Not Explicit) Bias (RQ2 + RQ3). We tested whether t2i bias affected implicit gender-brilliance associations (RQ2 + RQ3). The effect of condition on participants' d scores (implicit bias) was significant, $F(4, 832) = 3.130$, $p = .014$, $n_p^2 = .015$.

The difference in implicit bias between the AI baseline condition ($M = .20$, $SD = .34$) and the pronounced stereotypic condition ($M = .18$, $SD = .32$) was not significant, $t(329) = 0.488$, $p = .626$, $d = 0.05$, 95% CI [-0.16, 0.27]. The difference between the AI baseline condition and the probabilistic stereotypic condition ($M = .15$, $SD = .38$) was also not significant, $t(331) = 1.178$, $p = .239$, $d = 0.13$, 95% CI [-0.09, 0.34]. In short, stereotypic t2i output did not increase people's implicit gender bias (RQ2).

We also tested whether exposure to egalitarian or counter-stereotypic t2i output decreased implicit gender bias (RQ3). There was no significant difference between the AI baseline condition and the egalitarian condition ($M = .22$, $SD = .33$), $t(336) = -0.635$, $p = .526$, $d = -0.07$, 95% CI [-0.28, 0.14]. However, implicit bias was significantly lower in the counter-stereotypic condition ($M = .10, SD = .33$) compared to the AI baseline condition, $t(333) = 2.642$, $p = .009$, $d = 0.29$, 95% CI [0.07, 0.50]. Counter-stereotypic t2i output reduced stereotypes linking men with brilliance, relative to gender-neutral t2i output. See Figure 2 for implicit bias scores across conditions and studies.[2]

We next tested whether t2i output affected participants' explicit biases, particularly the explicit endorsement of the gender-brilliance stereotype. We created a composite of the 8-item Gender-Brilliance Stereotype Endorsement Scale (i.e., BSE; $\alpha = .941$). There was no significant effect of t2i condition on the BSE, $F(4, 832) = 0.985$, $p = .415$, $n_p^2 = .005$. There were also no significant effects of condition on the one-item measures assessing the gender-brilliance association, $F(4, 829) = 0.656$, $p = .623$, $n_p^2 = .003$, and the gender-friendly association, $F(4, 830) = 0.511$, $p = .728$, $n_p^2 = .002$. In general, participants tended to explicitly associate genius with men more than women ($M = 0.31, SD = 1.10$) and friendly with women more than men ($M = 0.47, SD = 1.30$), where positive numbers = more genius/men and friendly/women associations.

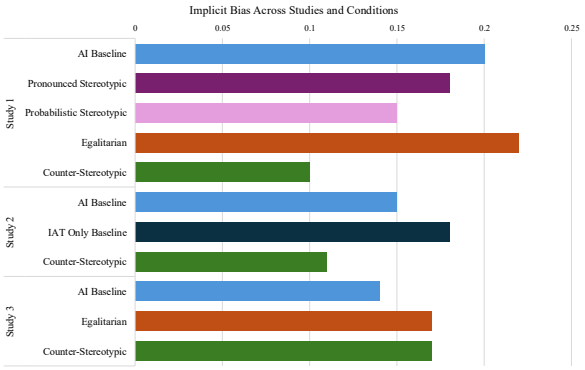

Figure 2: Implicit bias scores across conditions in all studies. Overall, implicit bias tended to be lower after seeing counter-stereotypic t2i output (Studies 1+2, but not Study 3).

[2]In a supplementary study, we recruited a subsample of Study 1 participants to complete the IAT again approximately 10 weeks later, with no additional t2i exposure. See details in the Appendix.

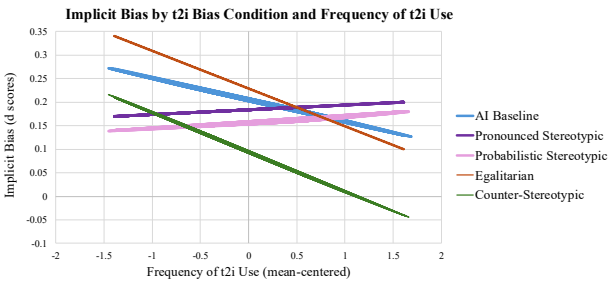

Figure 3: Interaction between t2i bias condition and participants' frequency of t2i use on implicit gender bias in Study 1.

Participant Gender Affected Explicit (But Not Implicit) Bias (RQ4). We tested the role of participant gender in Study 1. We first tested whether the effect of t2i condition on implicit bias was moderated by participant gender (across women and men). The interaction between condition and participant gender was not significant, $F(4, 822) = 1.290$, $p = .272$, $n_p^2 = .006$. The main effect of participant gender was also not significant, $F(1, 822) = 3.645$, $p = .057$, $n_p^2 = .004$, though women had descriptively higher implicit bias ($M = .20$, $SD = .35$) than men ($M = .15$, $SD = .34$).

Consistent with prior work, men had significantly higher explicit gender–brilliance bias than women on the BSE, $F(1, 822) = 61.587$, $p < .001$, $n_p^2 = .070$, and the single genius-gender item, $F(1, 820) = 40.606$, $p < .001$, $n_p^2 = .048$, though women had greater women-friendly associations on the single friendly-gender item, $F(1, 821) = 16.335$, $p < .001$, $n_p^2 = .020$. Men also reported significantly more benevolent sexism, $F(1, 822) = 4.944$, $p = .026$, $n_p^2 = .006$, and hostile sexism, $F(1, 821) = 79.069$, $p < .001$, $n_p^2 = .089$, compared to women.

None of these gender main effects were moderated by t2i condition, with the exception of a significant interaction between t2i condition and participant gender on benevolent sexism, $F(4, 822) = 2.495$, $p = .042$, $n_p^2 = .012$. Interestingly, men reported higher benevolent sexism after exposure to egalitarian t2i, making the gender difference more pronounced in this condition (Men: $M = 0.96$, $SD = 0.96$; Women: $M = 0.44$, $SD = 1.09$).

Frequency of t2i Use Related to Implicit and Explicit Bias (RQ4). We next tested whether participant identities or experiences qualified the effect of t2i bias on implicit bias.

We analyzed how often participants reported using t2i models. There was a significant interaction between t2i use and t2i condition, $F(4, 828) = 2.719$, $p = .029$, $n_p^2 = .013$. More t2i use was associated with *lower* implicit bias in the egalitarian and counter-stereotypic conditions, reversing a more positive association in stereotypic conditions (see Figure 3). Output contradicting the status quo may be particularly impactful for those who spend more time using t2i.

Interestingly, t2i use was also positively associated with explicit gender bias and sexism, indicated by significant effects of t2i use on the BSE, $F(1, 828) = 34.296$, $p < .001$, $n_p^2 = .040$, benevolent sexism, $F(1, 828) = 13.349$, $p < .001$, $n_p^2 = .016$, and hostile sexism, $F(1, 827) = 36.053$, $p < .001$, $n_p^2 = .042$.

We also tested whether participant age moderated the effect of t2i condition on implicit bias. There was no significant main effect of participant age on implicit bias, $F(1, 825) = 0.137$, $p = .711$, $n_p^2 = .000$, nor a significant interaction between age and condition, $F(4, 825) = 0.414$, $p = .798$, $n_p^2 = .002$.

Study 2 Results

In Study 2, we aimed to replicate the primary results from Study 1: the effect of counter-stereotypic t2i on implicit gender bias, and the impact of t2i bias qualified by participant-level factors.

Counter-Stereotypic t2i Output Reduced Implicit Bias (RQ3). The main effect of t2i condition on implicit bias was not significant, $F(2, 523) = 1.994$, $p = .137$, $n_p^2 = .008$. We conducted the planned comparisons between conditions. There was no significant difference between the IAT-only baseline ($M = .18$, $SD = .35$) and the AI baseline ($M = .15$, $SD = .35$) conditions, $t(349) = -0.832$, $p = .406$, 95% CI [-0.30, 0.12], $d = -0.09$, confirming that t2i exposure alone did not impact implicit gender bias.

As expected, implicit bias was significantly lower in the counter-stereotypic condition ($M = .11$, $SD = .36$) compared to the IAT-only baseline, $t(347) = 1.972$, $p = .049$, 95% CI [0.001, 0.42], $d = 0.21$.

Implicit bias was descriptively lower in the counter-stereotypic condition compared to the AI baseline condition, however, this difference was not significant, $t(344) = 1.162$, $p = .246$, 95% CI [-0.09, 0.34], $d = 0.13$.

The Impact of t2i Output Differed Across Participant Groups in Study 2 (RQ4). We first analyzed participant gender. The main effect of gender on implicit bias was not significant, $F(1, 514) = 0.105$, $p = .746$, $n_p^2 = .000$, and the interaction between gender and t2i condition was not significant, $F(2, 514) = 2.811$, $p = .061$, $n_p^2 = .011$. Though the interaction was not significant, descriptive patterns showed that women's implicit bias was considerably lower in the counter-stereotypic condition ($M = 0.06$, $SD = 0.36$) compared to women in the AI baseline condition ($M = 0.17$, $SD = 0.31$), $t(169) = 2.074$, $p = .040$, 95% CI [0.02, 0.62], $d = 0.32$. In contrast, men's implicit bias did not significantly differ across the counter-stereotypic condition ($M = 0.15$, $SD = 0.35$) and the AI baseline condition ($M = 0.14$, $SD = 0.37$), $t(168) = -0.186$, $p = .853$, 95% CI [-0.33, 0.27], $d = -0.03$. Broadly, the counter-stereotypic condition appeared to have more of an impact on women's than men's implicit bias.

We next tested how often participants used t2i. Unlike Study 1, there was no interaction between t2i use and condition on implicit bias, $F(2, 520) = 1.837$, $p = .160$, $n_p^2 = .007$. However, more t2i use was again positively associated with explicit gender bias and sexism along measures of the BSE, $r = .186$, $p < .001$, benevolent sexism, $r = .126$, $p = .004$, and hostile sexism, $r = .213$, $p < .001$.

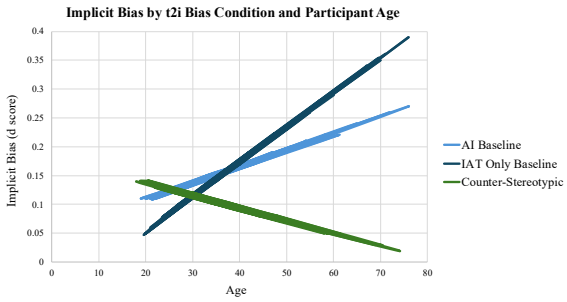

Figure 4: Interaction between t2i bias condition and participant age on implicit gender bias in Study 2.

Finally, we tested participant age as a moderator. There was a significant interaction between age and t2i condition on implicit bias, $F(1, 518) = 3.287$, $p = .038$, $n_p^2 = .013$. Whereas the relationship between age and implicit bias tended to be positive, this relationship flipped in the counter-stereotypic condition (see Figure 4). In other words, the counter-stereotypic t2i output was particularly impactful for older adults.

Study 3 Results

In Study 3, we aimed to replicate the effect of counter-stereotypic (but not egalitarian) t2i on implicit bias. We also analyzed participants' preferences for representations of women and men in t2i output (RQ5).

T2i Output Did Not Affect Implicit Bias. The main effect of t2i condition on implicit bias was not significant, $F(2, 524) = 0.362$, $p = .696$, $n_p^2 = .001$. There was no significant difference in implicit bias between the counter-stereotypic condition ($M = .17$, $SD = .37$) and the AI baseline condition ($M = .14$, $SD = .35$), $t(342) = -0.697$, $p = .486$, 95% CI [-0.29, 0.14], $d = -0.08$. There were also no significant differences between the egalitarian condition ($M = .17$, $SD = .33$) as compared to both the AI baseline condition, $t(351) = -0.778$, $p = .437$, 95% CI [-0.29, 0.13], $d = -0.08$, and the counter-stereotypic condition, $t(351) = 0.034$, $p = .973$, 95% CI [-0.21, 0.21], $d = 0.004$. Contrary to prior studies, counter-stereotypic t2i did not reduce implicit bias in Study 3. We discuss these results further in the Discussion.

People Tended to Prefer Accurate or Egalitarian t2i Representations, but Choices Differed Across Gender (RQ5). When forced to select one representational policy, the majority of participants (44%) preferred accurate t2i representations (in which actual gender disparities are represented in occupation output), followed by 32.4% of participants who preferred egalitarian t2i representations and 18.3% of participants who selected "it doesn't matter". A much smaller percent of participants preferred counter-stereotypic t2i (4%) or amplified bias (1.3%).

Representational harm preferences also varied across participants' t2i condition and gender (see Figure 5). Amongst women, a Chi Square test revealed a significant relationship between t2i condition and the representational harm selection, $\chi^2 = 17.412$, $p = .026$, $\phi = .262$. This relationship was not significant amongst men, $\chi^2 = 2.833$, $p = .944$, $\phi = .103$. The majority of women in the AI baseline condition preferred Accurate t2i, however, the majority shifted toward Egalitarian t2i for women in both the egalitarian and counter-stereotypic conditions. In contrast, most men preferred Accurate t2i in all conditions — exposure to t2i output contradicting the status quo did not shift men's preferences.

Participants also rated the acceptability of each representational option. These responses largely reflected patterns in the forced choice question. T2i condition did not significantly affect the perceived acceptability of any option ($ps > .07$, $n_p^2 s < .011$). However, participant gender had a significant effect on the perceived acceptability of Egalitarian t2i, $F(1, 522) = 6.567$, $p = .011$, $n_p^2 = .013$, Accurate t2i, $F(1, 522) = 4.252$, $p = .040$, $n_p^2 = .008$, and Amplified Bias t2i, $F(1, 520) = 5.133$, $p = .024$, $n_p^2 = .010$. See patterns in Figure 6. Whereas women (vs. men) rated Egalitarian t2i as more acceptable, men rated Accurate and Amplified t2i Bias as more acceptable.

Discussion

The impact of biased text-to-image models on users' implicit bias is an important question with implications for mitigating harm in future generative AI systems. Despite published findings observing stark gender stereotypes in a popular t2i model (Stable Diffusion) (Bianchi et al. 2023), our own analysis of a later version of this model (in Spring 2024) showed similar occupation gender biases persisted (RQ1). Out of 100 images generated for each of the 12 gender-stereotyped occupations via Stable Diffusion, over 75% of the output was labeled as the stereotypic gender group. Importantly, these occupations vary not just in their gender stereotypicality, but also in their perceived association with *brilliance*, whereby men are overrepresented in occupations seen as requiring exceptional intelligence (Storage et al. 2016). Gender-brilliance associations have real consequences for women's under-representation in a variety of fields (Leslie et al. 2015), underscoring the importance of characterizing the impacts of biased t2i output.

The subsequent three studies characterized the impact of gender-biased occupation t2i output on humans. Centrally, counter-stereotypic t2i output reduced implicit gender bias (RQ3; Studies 1 and 2; though this effect failed to replicate in Study 3). Seeing women in occupations typically associated with brilliance (e.g., physicist) via t2i output reduced participants' implicit associations linking men with brilliance. This was a meaningful reduction of an implicit bias that appeared strong even amongst participants at baseline who saw only gender-neutral output. This work is the first, to our knowledge, to demonstrate psychological effects of *counter*-stereotypic t2i by controlling the gender ratios in output. Output that contradicts the status quo could have unique, prosocial impacts on users interacting with t2i models. Further, simulating the social effects of AI in experimental design may be a valuable method for continued research on human-AI interaction.

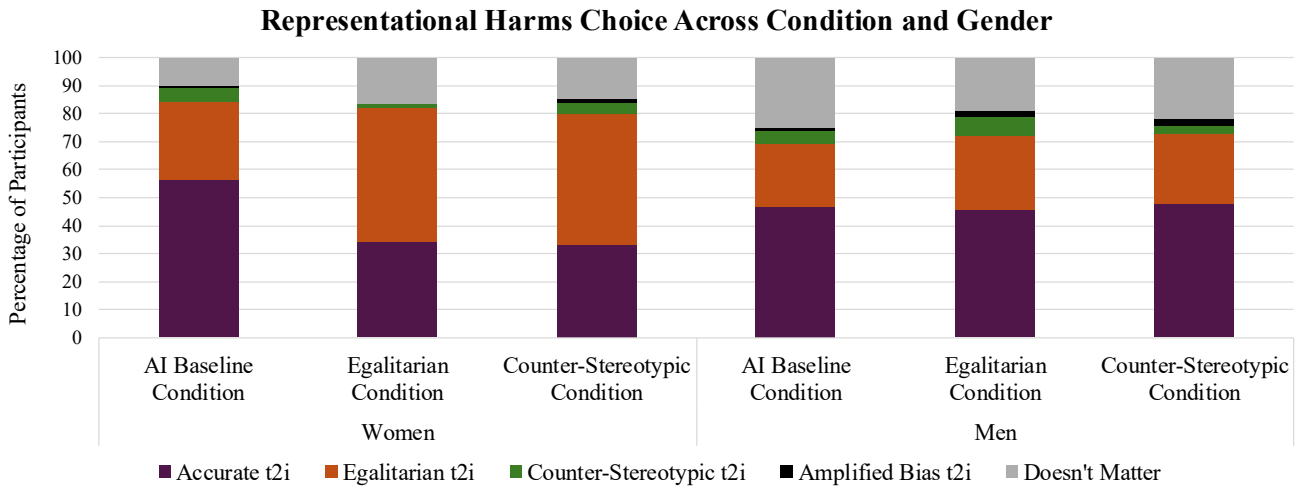

Figure 5: Percent of participants who selected each t2i representational harm choice, across participant gender (women = left panel; men = right panel) and t2i condition in Study 3.

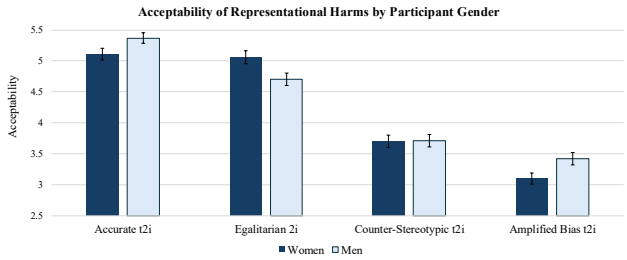

Figure 6: Perceived acceptability of t2i representations across participant gender in Study 3. Error bars represented by standard error of the mean.

Interestingly, the impact of counter-stereotypic t2i was often dependent on features of participants, such as their frequency of t2i use and age (RQ4). In Study 1, people who reported using t2i more frequently had lower implicit bias in the egalitarian and counter-stereotypic conditions. In Study 2, older adults also had lower implicit bias in the counter-stereotypic condition, reversing the otherwise positive association between age and implicit bias. Both populations may be more often naturally exposed to gender stereotypes, such that exposure to t2i output contradicting that status quo is especially impactful. Indeed, more frequent t2i use was associated with higher explicit gender bias and sexism. This association hints at potential longer-term consequences of repeated exposure to t2i systems. Because t2i can perpetuate and amplify social biases (Bianchi et al. 2023; Ghosh and Caliskan 2023; Naik and Nushi 2023; Shukla et al. 2025; Wolfe et al. 2023), people who use t2i more frequently may have more repeated interactions with harmful social stereotypes. Though these results are only correlational, longitudinal studies may provide insight into whether repeated t2i use has longer-term consequences for humans' gender bias.

We also found that t2i output did not meaningfully affect *explicit* gender bias. In other words, although t2i output affected the cognitive accessibility of gender–brilliance stereotypes, t2i output did not significantly affect participants' more stable explicit beliefs. This is consistent with the broader literature on stereotype change, which suggests explicit biases are often less malleable (Forscher et al. 2019).

Further, exposure to stereotype-consistent output did not increase implicit gender bias (RQ2; Study 1). These findings appear to stand in contrast to prior research, which found stereotypic online content can increase bias (Guilbeault et al. 2024). Why do the present results differ from typical effects of stereotype exposure on implicit bias? Stereotypic t2i conditions here included relatively minimal exposure to stereotypic output. We posit that because people already hold this implicit gender–brilliance stereotype (and indeed, participants exhibited this bias at baseline), minimal exposure to t2i output consistent with participants' stereotypes may not be enough to increase an already strong implicit bias. Related work found similar patterns in which stereotype-consistent output did not alter cognition (Vlasceanu and Amodio 2022). A stronger stereotypic signal in the output (e.g., more occupation output, or more repeated exposure) may have a greater impact on implicit gender bias.

Finally, we gathered novel insights into how people prefer t2i to represent women and men relative to their actual distribution in different occupations (RQ5; Study 3). Although counter-stereotypic output may reduce implicit gender bias, only a small minority of participants (less than 5%) preferred counter-stereotypic t2i representations. Instead, the vast majority of participants preferred t2i to accurately represent women's and men's distribution across occupations, though this output would include stark gender disparities and stereotypes. Interestingly, women and men differed across conditions in their preferred t2i representations. Women reported a greater interest than men in t2i embedding egalitarian val-

ues in its design, particularly after being exposed to how this might look in practice. Women may be more responsive to interventions that are favorable to their ingroup (i.e., associating women with brilliance), whereas men may be more resistant to interventions perceived as threatening their status. Descriptive patterns in Study 2 provide tentative support for this idea: the counter-stereotypic condition appeared to have a greater impact on women's than men's implicit bias. Future work should continue to explore how women and men react to counter-stereotypic interventions in t2i.

Limitations and Future Directions

The present work raises several questions that may motivate future research. First, the effect of counter-stereotypic t2i on implicit bias, observed in Studies 1 and 2, failed to replicate in Study 3. Though we are unsure why we no longer observed this effect, we note that Study 3 was conducted in February 2025, shortly after the presidential inauguration. Counter-stereotypic interventions may be less effective in the face of an overwhelming influx of news related to DEI, AI, and other social concerns. Altogether, feeling exhausted, overwhelmed, and tuned out could affect the utility of these interventions. Studies 1 and 2 suggest that counter-stereotypic t2i can impact implicit gender bias, but we acknowledge the idiosyncratic findings across studies.

Second, we used Rekognition to label the perceived gender of output because it was necessary to determine whether output aligned with gender stereotypes. However, we acknowledge that gender is a spectrum and, for real people, should not be presumed by a third party, nor is there ground truth as to the actual gender of Stable Diffusion's output.

Third, future studies could include a wider range of identities and experiences. Analyses by participant gender included only those who identified as women or men because the smaller sample of non-binary participants did not provide sufficient power for statistical tests, a limitation of this work. Further, we did not constrain the perceived race of the t2i output, though these models often exhibit intersectional biases (Bianchi et al. 2023). Continued work can deepen our understanding of t2i harms for a wider range of populations.

Finally, the present work focused on implicit gender–brilliance stereotypes tied to women's and men's representation in occupations. There is good reason to focus on this stereotype association, which is strong in magnitude and has real implications for women's experiences and underrepresentation across a variety of fields (Bian, Leslie, and Cimpian 2017, 2018; Bian et al. 2018). However, there are many questions left to explore. For instance, does exposure to biased t2i impact other forms of bias or discrimination in decision-making tasks (e.g., hiring)? Further, what harms arise from interacting with biased t2i for those who are being unfairly represented in its output? Our focus on text-to-image models and implicit gender bias is an important first step in beginning to characterize the unique harms of generative AI across different outcomes, models, and social groups.

Implications for Generative AI

This work has implications for future generative AI and t2i systems. Centrally, our findings suggest a way in which t2i could be designed to counteract cultural stereotypes. Some approaches suggest that technology can be designed for the social good, to consciously include marginalized groups and challenge stereotype norms (Breslin and Wadhwa 2014a,b). A more radical suggestion from this work is to consider how generative AI could actively challenge societal biases in design, instead of reinforcing and amplifying these biases.

The results of Study 3 provide a more nuanced picture of how future generative AI design may seek to resolve stereotype-reinforcing output. The majority of participants were not strongly in favor of t2i output showing counter-stereotypic representations of women and men, suggesting some users may be resistant to this type of intervention. However, more participants preferred egalitarian or equal ratios of women and men in t2i output, which itself would not be a perfectly accurate representation of society, suggesting users may be open to t2i output that challenges the status quo. Our results highlight tensions between the social impact of generative AI design and generative AI users' values and preferences for design, all of which should be considered when designing, developing, deploying, or conducting situated evaluations of future models.

Notably, we found that less than 5% of participants preferred t2i output that amplifies existing gender disparities in society. This finding is perhaps not surprising on its own, however, public text-to-image models *do* amplify existing gender disparities in occupations (Bianchi et al. 2023). Users' preferences misalign with current practices, suggesting there is room for improvement in current t2i models. Though our findings cannot provide a concrete answer as to how t2i output should represent women and men, it is clear that models should strive to reduce stereotypic representations — especially those that amplify harmful social stereotypes. We hope this work motivates continued research on how to improve generative AI systems.

Conclusion

Across three studies, participants engaged in a simulated interaction with a text-to-image model in which the output was controlled for perceived gender bias. We found that t2i output can affect humans' implicit gender bias, and in particular people's associations between men and exceptional intelligence. Counter-stereotypic t2i output *reduced* people's implicit gender bias, relative to gender-neutral t2i output. Though we observed resistance to counter-stereotypic t2i output, many people (and especially women) preferred t2i output to show more equal representations of women and men that contradicts the status quo. These findings have real implications for improving generative AI systems that seek to mitigate social harms and motivating continued experimental work simulating the social impacts of generative AI.

Ethical Considerations

The studies were submitted for review to the authors' human subjects review boards (IRBs), who deemed this research exempt because it poses no more than minimal risk to participants and meets several other federal guidelines for exempt research. Participants were also debriefed at the end of the study about the purpose of our study and the manipulation.

We took additional ethical precautions in selecting images to show participants. For instance, prior work has demonstrated that text-to-image models sometimes generate sexualized or objectified images of women and girls (Wolfe et al. 2023). In our work, one researcher manually reviewed all generated images to ensure they were not overtly sexual or NSFW, resulting in the removal of one questionable image.

Acknowledgments

We are grateful to the anonymous reviewers for their helpful feedback. This work was supported by the U.S. National Institute of Standards and Technology (NIST) Award 60NANB23D194. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of NIST. Two authors (Mattea Sim and Tadayoshi Kohno) were supported by the National Science Foundation under Awards SaTC-2205171 and 2207019, as part of the Center for Privacy and Security for Marginalized and Vulnerable Populations (PRISM). Tadayoshi Kohno was also supported by the McDevitt Chair in Computer Science, Ethics, and Society at Georgetown University. The majority of this research was conducted while author Tadayoshi Kohno worked at University of Washington.

References

AWS. 2024. Amazon Rekognition Documentation. https://docs.aws.amazon.com/rekognition/. Online; accessed 12 May 2025.

Barocas, S.; Crawford, K.; Shapiro, A.; and Wallach, H. 2017. The Problem with Bias: Allocative versus Representational Harms in Machine Learning. In *Special Interest Group for Computing, Information, and Society (SIGCIS)*.

Bian, L.; Leslie, S.-J.; and Cimpian, A. 2017. Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, 355(6323): 389–391.

Bian, L.; Leslie, S.-J.; and Cimpian, A. 2018. Evidence of bias against girls and women in contexts that emphasize intellectual ability. *American Psychologist*, 73(9): 1139–1153.

Bian, L.; Leslie, S.-J.; Murphy, M. C.; and Cimpian, A. 2018. Messages about brilliance undermine women's interest in educational and professional opportunities. *Journal of Experimental Social Psychology*, 76: 404–420.

Bianchi, F.; Kalluri, P.; Durmus, E.; Ladhak, F.; Cheng, M.; Nozza, D.; Hashimoto, T.; Jurafsky, D.; Zou, J.; and Caliskan, A. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1493–1504. Chicago IL USA: ACM.

Breslin, S.; and Wadhwa, B. 2014a. Engendering interaction design. In *2014 3rd International Conference on User Science and Engineering (i-USEr)*, 292–295.

Breslin, S.; and Wadhwa, B. 2014b. Exploring Nuanced Gender Perspectives within the HCI Community. In *Proceedings of the India HCI 2014 Conference on Human Computer Interaction - IHCI '14*, 45–54. New Delhi, India: ACM Press.

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

Charles, M.; and Bradley, K. 2009. Indulging Our Gendered Selves? Sex Segregation by Field of Study in 44 Countries. *American Journal of Sociology*.

Charlesworth, T. E. S.; and Banaji, M. R. 2022. Evidence of Covariation Between Regional Implicit Bias and Socially Significant Outcomes in Healthcare, Education, and Law Enforcement. In Deshpande, A., ed., *Handbook on Economics of Discrimination and Affirmative Action*, 1–21. Singapore: Springer Nature Singapore.

Cho, J.; Zala, A.; and Bansal, M. 2023. DALL-EVAL: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3020–3031. Paris, France: IEEE.

Dev, S.; Sheng, E.; Zhao, J.; Amstutz, A.; Sun, J.; Hou, Y.; Sanseverino, M.; Kim, J.; Nishi, A.; Peng, N.; and Chang, K.-W. 2022. On Measures of Biases and Harms in NLP. *arXiv*, (arXiv:2108.03362).

Devine, P. G. 1989. Stereotypes and Prejudice: Their Automatic and Controlled Components. *Journal of Personality and Social Psychology*, 56(1): 5–18.

Forscher, P. S.; Lai, C. K.; Axt, J. R.; Ebersole, C. R.; Herman, M.; Devine, P. G.; and Nosek, B. A. 2019. A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, 117(3): 522–559.

Ghosh, S.; and Caliskan, A. 2023. 'Person'== Light-skinned, Western Man, and Sexualization of Women of Color: Stereotypes in Stable Diffusion. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Ghosh, S.; Lutz, N.; and Caliskan, A. 2024. "I Don't See Myself Represented Here at All": User Experiences of Stable Diffusion Outputs Containing Representational Harms across Gender Identities and Nationalities. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7: 463–475.

Glick, P.; and Fiske, S. T. 1996. The Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *Journal of Personality and Social Psychology*, 70(3): 491–512.

Gozalo-Brizuela, R.; and Garrido-Merchán, E. C. 2023. A survey of Generative AI Applications. *arXiv preprint*, (arXiv:2306.02781).

Greenwald, A. G.; and Banaji, M. R. 1995. Implicit Social Cognition: Attitudes, Self-Esteem, and Stereotypes. *Psychological Review*, 102(1): 4–27.

Greenwald, A. G.; and Banaji, M. R. 2017. The implicit revolution: Reconceiving the relation between conscious and unconscious. *American Psychologist*, 72(9): 861–871.

Greenwald, A. G.; Brendl, M.; Cai, H.; Cvencek, D.; Dovidio, J. F.; Friese, M.; Hahn, A.; Hehman, E.; Hofmann, W.; Hughes, S.; Hussey, I.; Jordan, C.; Kirby, T. A.; Lai, C. K.; Lang, J. W. B.; Lindgren, K. P.; Maison, D.; Ostafin,

B. D.; Rae, J. R.; Ratliff, K. A.; Spruyt, A.; and Wiers, R. W. 2022. Best research practices for using the Implicit Association Test. *Behavior Research Methods*, 54(3): 1161–1180.

Greenwald, A. G.; McGhee, D. E.; and 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.

Guilbeault, D.; Delecourt, S.; Hull, T.; Desikan, B. S.; Chu, M.; and Nadler, E. 2024. Online images amplify gender bias. *Nature*, 626(8001): 1049–1055.

Katzman, J.; Wang, A.; Scheuerman, M.; Blodgett, S. L.; Laird, K.; Wallach, H.; and Barocas, S. 2023. Taxonomizing and Measuring Representational Harms: A Look at Image Tagging. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12): 14277–14285.

Kay, M.; Matuszek, C.; and Munson, S. A. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3819–3828. ACM.

Kurdi, B.; Seitchik, A. E.; Axt, J. R.; Carroll, T. J.; Karapetyan, A.; Kaushik, N.; Tomezsko, D.; Greenwald, A. G.; and Banaji, M. R. 2019. Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist*, 74(5): 569–586.

Leslie, S.-J.; Cimpian, A.; Meyer, M.; and Freeland, E. 2015. Expectations of brilliance underlie gender distributions across academic disciplines. *Science*.

Lewis, M.; and Lupyan, G. 2020. Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour*, 4(10): 1021–1028.

Miller, D. I.; Eagly, A. H.; and Linn, M. C. 2015. Women's representation in science predicts national gender-science stereotypes: Evidence from 66 nations. *Journal of Educational Psychology*, 107(3): 631–644.

Moss-Racusin, C. A.; Dovidio, J. F.; Brescoll, V. L.; Graham, M. J.; and Handelsman, J. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109(41): 16474–16479.

Naik, R.; and Nushi, B. 2023. Social Biases through the Text-to-Image Generation Lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 786–808. Montréal QC Canada: ACM.

Nosek, B. A.; Smyth, F. L.; Sriram, N.; Lindner, N. M.; Devos, T.; Ayala, A.; Bar-Anan, Y.; Bergh, R.; Cai, H.; Gonsalkorale, K.; Kesebir, S.; Maliszewski, N.; Park, J.; Schnabel, K.; Shiomura, K.; Tulbure, B. T.; Wiers, R. W.; Akrami, N.; Ekehammar, B.; Vianello, M.; Banaji, M. R.; and Greenwald, A. G. 2009. National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, 106(26): 10593–10597.

OpenAI. 2024. GPT-4 Technical Report. *arXiv*, (arXiv:2303.08774). http://arxiv.org/abs/2303.08774.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv*, (arXiv:2307.01952).

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 37 th International Conference on Machine Learning*, 8821–8831.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shukla, P.; Chinchure, A.; Diana, E.; Tolbert, A.; Hosanagar, K.; Balasubramanian, V. N.; Sigal, L.; and Turk, M. A. 2025. BiasConnect: Investigating Bias Interactions in Text-to-Image Models. *arXiv*, (arXiv:2503.09763).

Storage, D.; Charlesworth, T. E.; Banaji, M. R.; and Cimpian, A. 2020. Adults and children implicitly associate brilliance with men more than women. *Journal of Experimental Social Psychology*, 90: 104020.

Storage, D.; Horne, Z.; Cimpian, A.; and Leslie, S.-J. 2016. The Frequency of "Brilliant" and "Genius" in Teaching Evaluations Predicts the Representation of Women and African Americans across Fields. *PLOS ONE*, 11(3): e0150194.

U.S. Bureau of Labor Statistics. 2024. Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity. *bls.gov*. Online; accessed 20 May 2025.

Vlasceanu, M.; and Amodio, D. M. 2022. Propagation of societal gender inequality by internet search algorithms. *Proceedings of the National Academy of Sciences*, 119(29): e2204529119.

Wolfe, R.; Yang, Y.; Howe, B.; and Caliskan, A. 2023. Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, 1174–1185. Chicago IL USA: ACM.