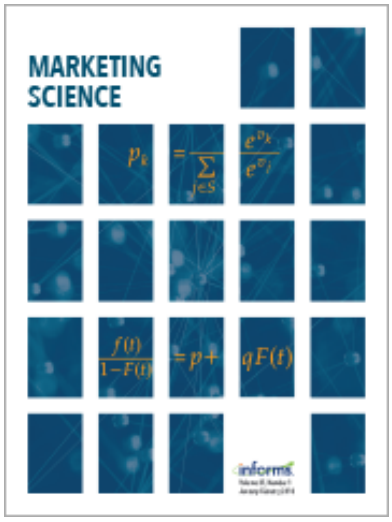


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Frontiers: Discrimination Against Femininity in Headshots: A Field Experiment with AI-Enabled Controllable Stimuli Generation

Lan E. Luo, Olivier Toubia

To cite this article:
Lan E. Luo, Olivier Toubia (2026) Frontiers: Discrimination Against Femininity in Headshots: A Field Experiment with AI-Enabled Controllable Stimuli Generation. Marketing Science

Published online in Articles in Advance 10 Sep 2026

. https://doi.org/10.1287/mksc.2024.1176

https://pubsonline.informs.org/journal/mksc

MARKETING SCIENCE
Articles in Advance, pp. 1–13
ISSN 0732-2399 (print), ISSN 1526-548X (online)

Frontiers: Discrimination Against Femininity in Headshots: A Field Experiment with AI-Enabled Controllable Stimuli Generation

Lan E. Luo,[a,*] **Olivier Toubia**[b]

[a] Yale School of Management, New Haven, Connecticut 06511; [b] Columbia Business School, New York, New York 10027

*Corresponding author

lan.luo@yale.edu (L. E. Luo), ot2107@gsb.columbia.edu (O. Toubia)

https://orcid.org/0000-0001-8189-4162 (L. E. Luo), https://orcid.org/0000-0001-7493-9641 (O. Toubia)

Received: November 1, 2024
Revised: May 22, 2025; April 16, 2026
Accepted: May 20, 2026
Published Online in Articles in Advance: September 10, 2026

https://doi.org/10.1287/mksc.2024.1176

Abstract. We document an understudied form of discrimination based on femininity expressed in headshots. To do so, we use a generative adversarial network to create realistic headshots of people and then manipulate their femininity independently of other attributes such as pose, general facial expression, background, hairstyle, and clothes. Then, we devise an experimental design that allows us to identify the separate and combined causal impact of femininity and gender identity (proxied by gender pronouns) on real-life outcomes. In a field experiment within a naturalistic advertising environment, we find that prospective customers of an education company discriminated against femininity at various stages of the purchase funnel to a largely similar extent for men, women, and nonbinary people. The findings inform managers and policymakers of an important form of discrimination that would otherwise be underestimated by disregarding headshots. Methodologically, we introduce a novel framework for testing specific hypotheses pertaining to causal effects of treatments that are derived from unstructured data. The approach involves using natural text to readily identify hypothesis-relevant features from a generative AI model (in a particularly disentangled and interpretable representation space) and then creating realistic stimuli that vary controllably in those features.

History: Catherine Tucker served as the senior editor.
Funding: This work was supported by the Sanford C. Bernstein & Co. Center for Leadership and Ethics; Columbia Experimental Laboratory in the Social Sciences.
Supplemental Material: The online appendix and data files are available at https://doi.org/10.1287/mksc.2024.1176.

Keywords: gender discrimination • field experiment • interpretable machine learning • representation learning • generative AI • unstructured data • causal inference

1. Introduction

Advances in AI enable us to more cleanly study how differences in the femininity expressed in headshots affect economic outcomes. Headshots are relevant in a wide variety of consequential and marketing-relevant settings such as hiring (e.g., Troncoso and Luo 2023), elections (Olivola and Todorov 2010), sharing economies (Zhang et al. 2025), and advertising (Xiao and Ding 2014); and more broadly in any context involving spokespeople, such as influencer marketing (Feng et al. 2025). Accordingly, headshots are pervasive in the digital economy. For example, headshots are prominently featured on LinkedIn and other social networks, appear in the news when a politician or manager is being discussed, and are often included in corporate websites and online marketplaces. Given their ubiquity, it is important to study whether and how discrimination occurs through headshots. When it comes to discrimination, headshot features that tend to be stable are both particularly pernicious as well as challenging to study, such as femininity. Unlike with hairstyle, clothes, and facial expression, we are unable to bring participants into a laboratory and alter their femininity in a consistent way. Furthermore, many headshot features may correlate with femininity (the latent treatment of interest), potentially influence the outcome, and be unobservable to the researcher. Therefore, it becomes difficult to identify the causal impact of only femininity on outcomes of interest using observational data because of latent omitted variable bias (Fong and Grimmer 2023).

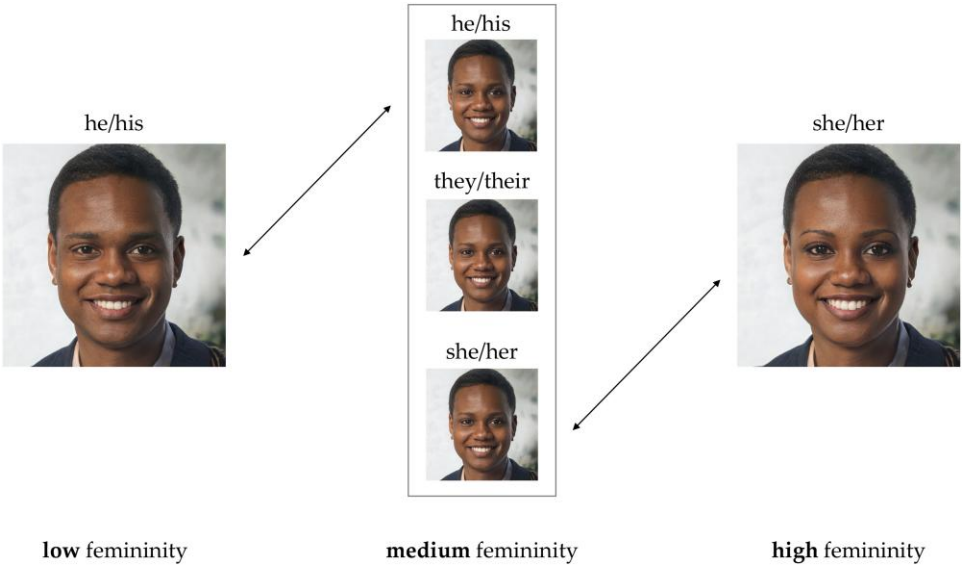
To resolve the challenge of studying the causal impact of femininity on market-relevant outcomes, we use particularly disentangled representations from deep generative models to controllably produce a

realistic set of stimuli that vary only on femininity, seeking to keep all other attributes of the headshot such as pose, general facial expression, background, hairstyle, and clothes constant. Our approach leverages StyleGAN2 (Karras et al. 2020), a generative adversarial network (GAN) (see Online Appendix A), which is capable of generating realistic images. StyleGAN2 has a particular latent space with disentangled properties, meaning that each latent dimension ideally controls just a single visual attribute (Wu et al. 2021). We adapted an approach to identify the relevant latent dimensions in StyleGAN2 corresponding to femininity using natural text (e.g., a photo of a female face; Patashnik et al. 2021). Because varying femininity may also affect the inferences people make about a person's gender identity (see Online Appendix D.2.2), we address this by devising a novel experimental design. The design involves explicitly yet subtly attaching gender pronouns (e.g., he/his, she/her, they/their) to headshots, and this allows us to separately identify the impact of femininity and gender identity (proxied by gender pronouns) on outcomes of interest (see Figure 1).

We apply our experimental design in a field experiment (within a naturalistic advertising environment) involving an online education service to assess impact on customer behavior. Specifically, we advertise an online tutoring company on Meta and experimentally vary the femininity and gender pronouns of the tutor featured on the company's landing page. We address the issue of algorithmic selection of ads in Meta A/B tests (Braun et al. 2024) by applying the treatment to the landing page rather than the ad itself, ensuring unconfoundedness. We find that prospective customers discriminated against femininity to a largely similar extent for men, women, and nonbinary people. Femininity main treatment effects are somewhat small but meaningful in magnitude: relative to low femininity, high femininity significantly decreases the probability of various customer actions by 2.3%–4% in spite of increasing the time spent on the website by 20.1 seconds.

Our study makes both substantive and methodological contributions that we hope will benefit scholars, managers, and policymakers across a wide range of fields. On the substantive side, we contribute to our understanding of discrimination, stigmatization, and inequities in marketplaces (e.g., Busse et al. 2017, Lambrecht and Tucker 2019, Proserpio et al. 2021, Ozturk et al. 2024). We hope our research can spur managers and policymakers to consider effective interventions. To note, in many cases, gender-blind and headshot-blind policies automatically address femininity bias because femininity is an attribute derived from headshots. Still, to the extent they are generalizable to other types of interactions with faces, we believe our findings have practical implications for the numerous face-to-face contexts in which blind policies are impossible or uncommon (e.g., face-to-face job interviews, real estate negotiations, venture capital pitches). In such contexts, the policy prescriptions to mitigate such discrimination are then a matter of framing (e.g., emphasizing credentials; Cui et al. 2020) or education (Paluck et al. 2021). In particular, gender-related antibias education could

Figure 1. Exemplar Stimuli Design



Notes. Comparing along diagonals (indicated by arrows) identifies the impact of femininity, keeping the gender identity of the person constant. Comparing along the middle column (indicated by the gray rectangle) identifies the impact of gender identity, keeping the femininity of the person constant.

emphasize how femininity is an often neglected channel (relative to gender identity) through which people may (un)wittingly discriminate. This is valuable because, for example, even if employers are aware of and address discrimination on the basis of gender identity, they may still differentially treat candidates unfairly of the same gender identity based on their femininity. Disregarding femininity in such contexts, thus, underestimates the extent of gender-based discrimination.

Methodologically, we develop a novel framework for testing specific hypotheses related to causal effects of latent treatments derived from unstructured data. The approach involves using natural text to readily identify features from a generative AI model (in a particularly disentangled representation space) that are relevant to some interpretable hypothesis of interest and then creating realistic stimuli that vary controllably in those features. More broadly, we believe this conceptual approach presents a rich set of opportunities for computational social science. Namely, researchers can use similar methodological approaches to experimentally identify the impact of a variety of latent treatments derived from text (e.g., reviews, political appeals, news/social media), images (e.g., product photos, political ads, urban/rural imagery), audio (e.g., service calls, political speeches, executive statements), and video.

2. Literature Review

We contribute to three streams of literature. First, we contribute to the literature on gender-based discrimination. Using correspondence studies, gender discrimination (based solely on names or pronouns) is documented in lending (Brock and De Haas 2023), academia (e.g., Milkman et al. 2015), sales (Kelley et al. 2026), and labor markets (e.g., Chan and Wang 2018, Kline et al. 2022). This literature, by omitting headshots, may underestimate the extent of gender-based discrimination. A separate literature, predominately from psychology, studies solely femininity (e.g., Perrett et al. 1998, Zietsch et al. 2015, Athey et al. 2025). The extant literature has been unable to reach clear conclusions as to whether and how people discriminate based on femininity. Normative beliefs about appearance may lead people to negatively evaluate counter-stereotypical faces (Sutherland et al. 2015, Oh et al. 2020). Alternatively, discriminatory behavior may be influenced by positive or negative gender stereotypes (Heilman et al. 2024), which may be exacerbated for feminine women. Regardless of gender identity, people may also discriminate against femininity because of a tendency to perceive feminine faces as less competent (Oh et al. 2019) or favor femininity because of a tendency to perceive feminine faces as more attractive, warm, honest, and cooperative (Perrett et al. 1998, Said and Todorov 2011). Because of these competing predictions, this research serves to explore how femininity in headshots impacts discriminatory behavior rather than a study testing any particular direction of effects (i.e., we test the null hypothesis that there is no discrimination based on femininity expressed in headshots without making an explicit prediction as to whether the null hypothesis is correct). Additionally, in these literatures, gender identity is not considered separately from femininity, which we believe to be a significant gap; regardless of how feminine a person looks, the person may identify as a man, woman, or nonbinary person. Thus, to the best of our knowledge, we are the first to identify (i) the partial effect of femininity accounting for gender pronouns and (ii) their interaction effect. Furthermore, extant literature on femininity generally considers more psychological constructs or stated preferences as outcomes (e.g., Perrett et al. 1998, Scott et al. 2014, Zietsch et al. 2015). Athey et al. (2025) consider the impact of femininity (jointly with gender identity) on real-life outcomes (i.e., preferences for loan campaigns) with observational data and laboratory studies without running a field experiment. Thus, to the best of our knowledge, we are the first in the literature to conduct a field experiment that studies the impact of femininity on real-life outcomes.

Second, we contribute to the literature on using computational methods to controllably generate stimuli that vary on interpretable attributes (i.e., controllable stimuli generation). The psychology literature studying femininity often does so using face morphing technologies such as FaceGen Modeller (e.g., Said and Todorov 2011, Nakamura and Watanabe 2020) or Psychomorph (e.g., Scott et al. 2014, Zietsch et al. 2015), which are either unable to generate realistic headshots (see Online Figures S1 and S2a) or cannot markedly change femininity (see Online Figures S2b and S2c) without introducing confounding (e.g., DeBruine et al. 2010), making them unsuitable for our application in the field. More recently, AI methods based on variational autoencoders (VAEs) have been leveraged for controllable stimuli generation (Dew et al. 2022, Sisodia et al. 2025). Unfortunately, we cannot leverage these methods because VAEs are unable to produce photorealistic stimuli, and this would limit the generalizability of any findings to real people. GANs have been used to manipulate headshots by perceived attributes (Peterson et al. 2022), demographics and facial expression (Liang et al. 2023, Athey et al. 2025), and predicted detainability of defendants (Ludwig and Mullainathan 2024). The generative methods in these papers can produce realistic headshots but struggle to completely isolate variation in particular attributes (e.g., see Online Figures S3 to S5). Our approach improves on isolating variation on specific attributes as we manipulate

stimuli using an unconventional and particularly disentangled choice of latent representation. We further contrast our work with Ludwig and Mullainathan (2024). They morph mugshots generated by a StyleGAN2 model by varying their predicted detention probability to generate hypotheses about what visual attributes influence detention likelihood. Being interested in hypothesis generation, Ludwig and Mullainathan (2024) are not particularly concerned with the detention decision being driven by (and, therefore, the headshot stimuli varying along) multiple confounded latent attributes of the mugshot (see Online Figure S6b). In contrast, our approach involves generating stimuli to test specific hypotheses (e.g., how does femininity impact discriminatory behavior). Thus, we designed our methodological approach to create stimuli that isolate variation specifically on the latent attribute of interest (e.g., femininity). Furthermore, the Ludwig and Mullainathan (2024) approach relies on the availability of a large sample of labeled images (e.g., by detention decision). In contrast, our approach allows the researcher to simply specify the attribute of interest (e.g., femininity) using natural text to automatically identify the relevant latent dimensions to be manipulated.

Third, we contribute to the literature on field experimentation with images. Such experiments have been run in the context of labor markets (e.g., Evsyukova et al. 2025), online marketplaces (e.g., Acquisti and Fong 2020), and advertising (e.g., Lambrecht and Tucker 2013). We propose an approach leveraging generative AI to more cleanly identify effects of treatments derived from images in experimental settings. Namely, we use particularly disentangled representations from deep generative models to induce exogenous variation in the latent treatment of interest, keeping constant many if not all potential latent confounding attributes (Fong and Grimmer 2023). In contrast to laboratory studies, field experiments are crucial for eliciting how such treatments affect behavior in a way that reflects incentive-compatible preferences and avoids response bias, especially on sensitive topics such as discrimination.

3. Generative Model Applied to Hypothesis Testing

Popular and off-the-shelf generative AI tools such as Dall-E 3 are not suitable for our particular application (see Online Appendix D.3). Instead, to generate stimuli, we leveraged a particular generative adversarial network model (see Online Appendix A for an overview) called StyleGAN2 (Karras et al. 2020), which was pretrained on a collection of 70,000 high-quality and relatively diverse headshots from Flickr. This model is capable of generating realistic headshots that are indistinguishable from real ones (Nightingale and Farid 2022). The key features of StyleGAN2 arise from the design of its generator: the input vector $\mathbf{z} \in \mathbb{R}^{512}$ (in the latent space $\mathcal{Z}$) is nonlinearly transformed into an intermediate vector $\mathbf{w} \in \mathbb{R}^{512}$ (in the latent space $\mathcal{W}$). A synthesis network generates images from $\mathbf{w}$ and is composed of a sequence of style blocks, each capturing different levels of abstraction. Earlier blocks control coarse-grained features such as face shape and pose; middle blocks, features such as facial expression and hairstyle; and later blocks, fine-grained features such as color scheme and microstructure. Each style block is a single convolutional layer whose weights are scaled by style parameters $\mathbf{s} \in \mathbb{R}^{9088}$ (in the latent space $\mathcal{S}$), an affine transformation of $\mathbf{w}$ learned for each block.

Of all the latent spaces in StyleGAN2, $\mathcal{S}$ is particularly disentangled (i.e., composed of latent dimensions that control distinct visual attributes; Wu et al. 2021). To identify the dimensions that control femininity, we use textual descriptions (e.g., a photo of a female face, a photo of a male face) to query $\mathcal{S}$ with embeddings from CLIP, a pretrained multimodal model that maps images and text into the same representation space (Patashnik et al. 2021; see Online Appendix C). Because CLIP was trained on 400 million image–text pairs, it is capable of connecting natural language to visual concepts, such as the features that compose femininity (which we verify with human raters in Online Appendices D.2.3 and D.2.5).

To manipulate femininity, we directly intervene on the relevant latent dimensions. In our context, femininity corresponds to just two latent disentangled dimensions, which means that 9,086 other latent dimensions (corresponding to attributes such as pose, general facial expression, background, hairstyle, and clothes) are, by construction, left unchanged when varying femininity. Our unconventional choice of latent space $\mathcal{S}$ (rather than $\mathcal{Z}$ or $\mathcal{W}$) as well as intervening on just a few of its latent dimensions allows us to more cleanly isolate variation in femininity relative to prior literature (e.g., Scott et al. 2014, Athey et al. 2025) when generating our stimuli.

We view our methodological approach as a general-purpose tool that can be readily applied to other contexts. For example, in Online Figure S7, we show a stimulus headshot being manipulated on three marketing-relevant attributes: charisma, attractiveness (Feng et al. 2025), and smile (Zhang et al. 2025). In our framework, leaving other steps unchanged, generating such stimuli simply requires specifying the target attribute using a short phrase (e.g., a photo of a charismatic face; see Online Appendix C). To emphasize, the key advantage of our methodological approach is that we can readily identify using natural text which latent dimensions correspond to some latent attribute (related to a hypothesis of interest) and then perturb those dimensions to generate realistic headshots that vary

controllably in that attribute. As off-the-shelf tools improve, they may also allow researchers to generate realistic images that vary along latent attributes of interest. Nevertheless, our approach still offers transparency, interpretability, and the ability to quantifiably control the intensity of the manipulation, which may not be possible with text-based prompting.

To note, there are some boundaries to the current implementation of our methodological approach when applied to headshots. First, the training data used for our StyleGAN2 model is not fully representative (e.g., composed of photos from before 2019 and biased toward white people; see Online Figure S8a). Thus, attributes specific to more recent trends or underrepresented groups may not have been properly learned, limiting the ability of our implementation to identify such attributes. Second, CLIP must be capable of connecting the natural text description of the hypothesis-relevant attribute to the image, and this may limit the identification of certain niche or domain-specific attributes. Third, because of limitations in its architecture and/or training process, StyleGAN2 may occasionally produce images with unrealistic properties (e.g., unnatural distortions or unrealistic objects in the headshot background). Fourth, in general (and beyond headshots), GANs are better suited for modeling relatively homogeneous data distributions (e.g., headshots) but struggle with highly diverse distributions (e.g., display ads) because of mode collapse, and this hinders their ability to learn rich and exhaustive features in such contexts.

4. Field Experiment

4.1. Experimental Setting and Design

We assess the impact of femininity and gender pronouns on real-life outcomes in an educational setting (preregistered using AsPredicted, #166345).[1] To do so, we created an online tutoring website, advertised through Meta to users who were located in the United States, were at least 18 years old, and spoke English.[2] We recruited participants through this Meta ad campaign, which ran for 50 days (March 15–May 4, 2024). The campaign achieved 468,941 impressions (i.e., number of times the ad was on screen for the first time) and reached 238,702 unique user accounts who saw the ad at least once ($M_{age} \approx 55.6$; 65% women, 35% men; 95% Facebook, 5% Instagram; 99.8% mobile, 0.2% PC). The ad itself was generic, was identical in content across all impressions, and did not contain any headshots or gender pronouns.

Upon clicking the ad, participants ($N = 2,880$ defined at the persistent cookie level;[3] $M_{age} \approx 58.2$; 63% women, 37% men; 97% Facebook, 3% Instagram; 100% mobile) were directed to a website we developed that randomly assigned participants to 1 of 20 conditions (5 headshot–pronoun pairings × 4 synthetic identities) in a between-subjects design. We present randomization checks in Section 4.4.

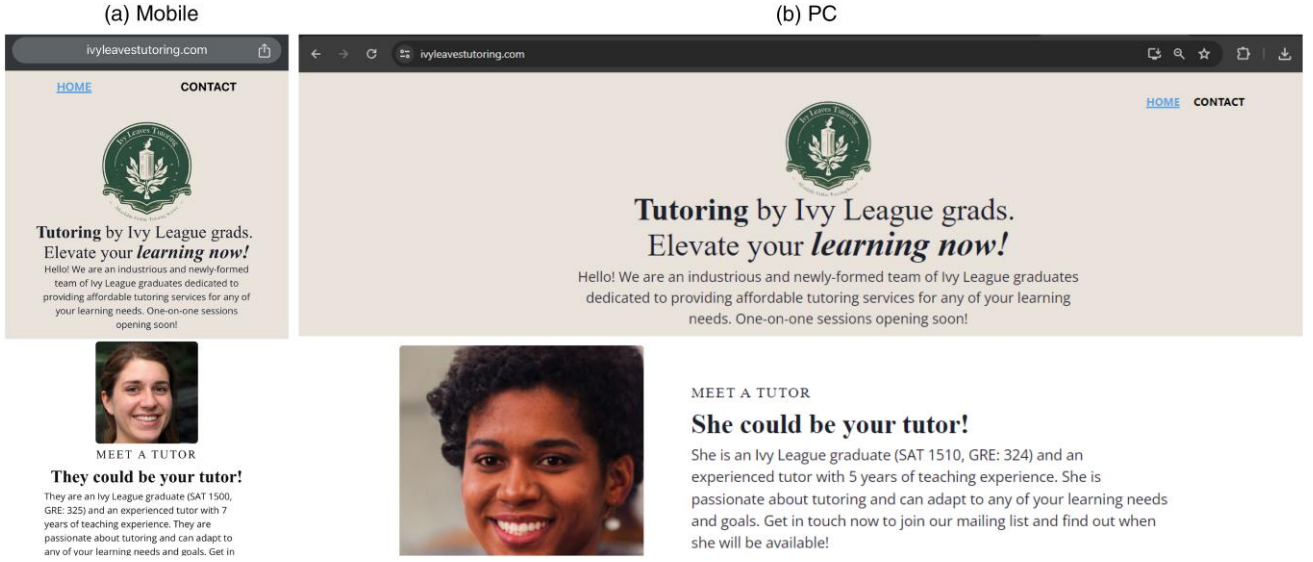
Each condition involved displaying a potential tutor profile with a headshot and a short description that subtly disclosed gender pronouns (e.g., They could be your tutor!) as well as some credentials to signal similarly high-quality, observed productivity characteristics (i.e., Ivy League graduate, relatively high SAT/GRE scores, and a few years of teaching experience; the latter two were set to be the same within synthetic identities and varied only slightly across synthetic identities). Participants could then scroll down (reflecting some interest in the program) and decide whether to click "Learn More" (which redirected to more information about the company, reflecting even greater interest in the program) and to click "Get in Touch" (which redirected to a short survey, reflecting desire and intention to sign up). Each of these decisions reflects the consumer journeying further down the purchase funnel, respectively. We also recorded a measure for how long consumers spent on the website, reflecting overall engagement. We consider scrolling down (to the bottom of the homepage), clicking "Get in Touch," clicking "Learn More," and time spent on the website (in seconds) as our preregistered dependent variables of interest.[4] All participants complied with treatment because, upon clicking, the headshot and at least three mentions of the gender pronoun were immediately visible on both mobile and PC (see Figure 2). See Online Appendix F.1 for details.

Note that we did not use Meta A/B testing tools because its algorithms optimize campaigns over time so that different users are eventually targeted across conditions, violating internal validity (Braun et al. 2024). Moreover, in doing so, these algorithms could propagate detrimental gender biases (Lambrecht and Tucker 2019). Instead, our design allowed us to fully control random assignment across conditions and ensure unconfoundedness. Thus, even within naturalistic advertising environments, our design presents a means to identify valid, stimuli-relevant treatment effects. In doing so, we most notably trade off with statistical power concerns (because participants are determined conditional on clicking and ad click-through rates are generally low, 0.69% in our campaign) and generalizability (because our sample is limited to those who decide to click on the initial ad).

4.2. Stimuli Generation

4.2.1. Initial Set of Headshots. We generate 20 synthetic identities, balanced equally by baseline perceived gender (man, woman) and race (White, Black), using StyleGAN2. For each synthetic identity, we create three versions: one baseline image and two manipulated versions. For baseline perceived men (women), we use our

Figure 2. Examples of Website (Immediately Upon Loading)



methodological approach to manipulate the generated image by increasing (decreasing) the values of the disentangled latent dimensions relevant to femininity by one as well as two standard deviations. Therefore, in our initial set, we have 60 headshots in total (see Online Appendix D.1).

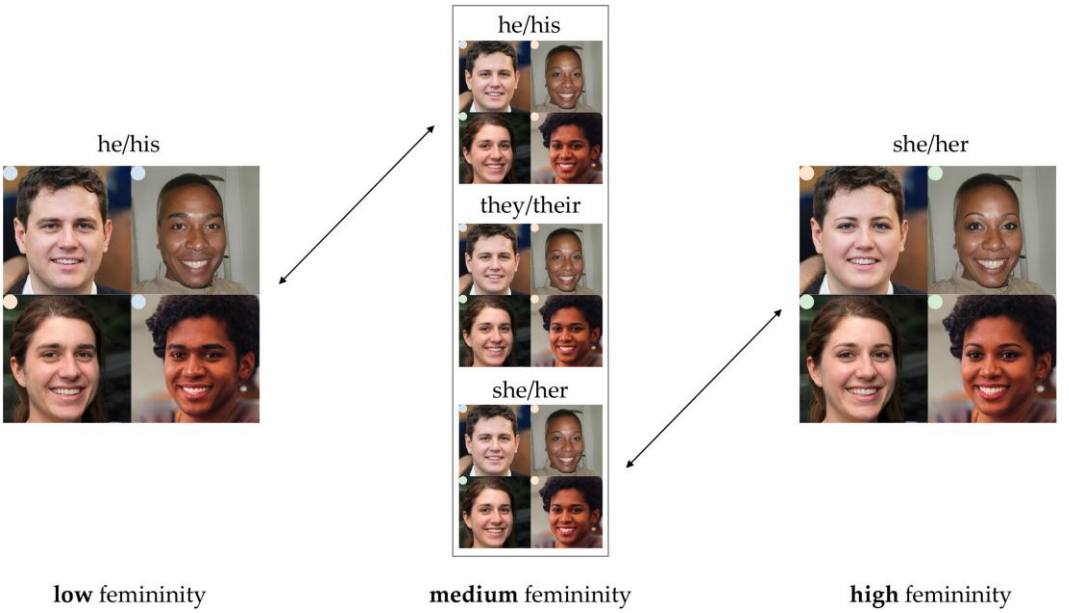
4.2.2. Headshot–Pronoun Pairings. For each synthetic identity, we pair the headshot with lowest femininity with the gender pronouns he/his and the headshot with highest femininity with she/her. For the headshot with medium femininity, we create three versions (using the same exact image), each of which is paired with different gender pronouns: he/his, she/her, or they/their. This fractional factorial design efficiently permits (separate) identification of the impact of femininity and gender identity, proxied by gender pronouns, on outcomes of interest. See Figure 1 for intuition. Femininity and pronoun interaction effects are identified by variation in femininity across synthetic identities.

4.2.3. Femininity Scores. Although we could use the femininity score derived from our generative model in subsequent analyses (e.g., regressing market outcomes on femininity), to ensure ecological validity, we measure the femininity of each image using participants from Amazon Mechanical Turk ($N = 317$; $M_{age} = 34.8$, $SD_{age} = 10.3$; 175 men, 142 women). This preregistered study (AsPredicted, #119258) was a within-subject design, in which each participant saw all 60 headshots in counterbalanced order. Participants rated each headshot on appearance on a continuous scale (with anchors at −1: *extremely masculine*, 0: *androgynous*, 1: *extremely feminine*).[5] We take the average of these ratings to extract a continuous femininity score for each headshot. We discuss details such as validating the full stimuli set in Online Appendices D.2.1 and D.2.2. We find that this measure of femininity is consistent with the measure from our generative model (see Online Appendix D.2.3). See Online Figure S10a for the distribution of femininity across all stimuli.

4.2.4. Stimuli Selection by Psychological Perceptions. In an exploratory laboratory study (AsPredicted, #119290), we estimate the impact of femininity and gender pronouns on context-independent, self-reported perceptions of trustworthiness and competence (see Online Appendix E). We do so in part to briefly speak to the extant psychology literature on how faces affect such perceptions (Todorov et al. 2015) as well as to explore whether our visually subtle manipulations affect any outcomes at all. In short, we find that femininity increased self-reported perceptions of trustworthiness and competence for all gender pronouns considered. These findings differ from prior literature (Sutherland et al. 2015; Oh et al. 2019, 2020), perhaps because of how our femininity effects control for gender pronouns; our unique approach to generating and manipulating stimuli (see Section 2); and/or our more recent and perhaps more progressive sample.

We also use this study to select four synthetic identities for the field experiment; for budget reasons, it was infeasible for us to run a well-powered experiment with all 20 synthetic identities. We selected one synthetic identity within each baseline perceived gender and race category with relatively high average perceived trustworthiness (to increase the likelihood that

Figure 3. Full Set of Headshot and Pronoun Pairings Used in Field Experiment



Notes. Circles on the top left of each headshot denote to which discretized femininity tertile (see Section 5.1) that headshot belongs. Blue represents the low femininity tertile; orange, medium; and green, high.

consumers perceive our education company and employees to be legitimate) and a markedly significant difference in means of perceived competence (to select for cases in which we expect treatment effects to be larger; see Online Appendix E.3). We display all stimuli corresponding to the 20 conditions used in the field experiment in Figure 3.

4.2.5. Manipulation Checks. We assess to what extent our controllable stimuli generation process isolates variation along femininity. From heat maps (see Online Figure S12), our manipulations appear to primarily involve changing the eyebrows as well as subtler features such as the shape of the mouth, eyes, and jawline, which human participants also observe in a preregistered study (AsPredicted, #162310) with closed-ended questions (see Online Appendix D.2.5).

To investigate further, specifically on the four synthetic identities selected for the field experiment, we asked participants from Prolific ($N = 72$; $M_{age} = 41.5$, $SD_{age} = 12.7$; 52 women, 20 men) to describe openendedly how the three headshots within a synthetic identity (displayed side by side) differ for all four synthetic identities in counterbalanced order (see Online Appendix D.2.5 for details). Descriptions from this preregistered study (AsPredicted, #220148) suggest that many observed differences relate to femininity (e.g., eyebrows, skin, chin/jaw, eyes, nose, and lips). However, participants also note changes in five other attributes, which we treat as potential latent confounders: makeup, facial hair, earrings, smile, and age. We view those attributes as latent confounders because they are not inherently features of femininity and are not relatively stable attributes apart from age.

4.2.6. Measuring Potential Latent Confounders. We measure scores for the five potential latent confounders identified in the previous section (makeup, facial hair, earrings, smile, age) to control for them in subsequent regressions. In a preregistered study (AsPredicted, #221255; see Online Appendix D.2.5 for details), using continuous scales, we ask participants from Amazon Mechanical Turk ($N = 308$; $M_{age} = 32.1$, $SD_{age} = 6.2$; 250 men, 58 women) the extent to which each headshot (among the 12 headshots selected for the field experiment in counterbalanced order) has makeup (e.g., mascara, lipstick), facial hair (e.g., mustache, beard, sideburns), earrings, and a smile (with anchors at −1: *not at all*, −0.5: *slight*, 0: *moderate*, 0.5: *considerable*, 1: *extreme*) as well as how old each face looks (with anchors at −1: *very young*, −0.5: *somewhat young*, 0: *middle-aged*, 0.5: *somewhat old*, 1: *very old*). We take the average of these ratings to extract a continuous score for each of the five potential latent confounders for each headshot.

4.3. Summary Statistics

Table 1 displays summary statistics for the continuous femininity score, psychological perceptions, and potential latent confounders (at the headshot level) as well as outcome variables (at the observation level).

Table 1. Summary Statistics for Field Experiment

	Mean	Standard deviation	Minimum	Median	Maximum	*N*
Headshot level						
Femininity score	0.24	0.42	−0.47	0.37	0.70	12
Perceived trustworthiness	0.39	0.09	0.22	0.39	0.54	12
Perceived competence	0.43	0.08	0.28	0.43	0.56	12
Perceived makeup	0.11	0.11	−0.08	0.10	0.31	12
Perceived facial hair	−0.11	0.08	−0.20	−0.14	0.10	12
Perceived earrings	−0.08	0.17	−0.22	−0.19	0.25	12
Perceived smile	0.46	0.06	0.34	0.48	0.52	12
Perceived age	−0.02	0.09	−0.10	−0.05	0.13	12
Observation level						
Whether scrolled to bottom	0.02	0.15	0	0	1	2,880
Whether clicked "Learn More"	0.01	0.10	0	0	1	2,880
Whether clicked "Get in Touch"	0.01	0.09	0	0	1	2,880
Time on site, s	1.51	30.42	0.00	0.00	1,527.69	2,879

Note. Deviating from preregistration, only for time on site, we removed one extreme outlier from all analyses, which had a value of 50,419.73 seconds.

4.4. Randomization Checks

Testing balanced treatment assignment, we fail to reject the null hypothesis at the $\alpha = 10\%$ level that the observed frequency of participants in each condition is the same as the expected frequency under perfectly balanced treatment assignment (i.e., 144 participants in each condition), $\chi^2(19) = 26.42$, $p = 0.12$. Next, we conduct randomization checks for the only individual-level information we were able to collect from the field experiment: browser type (Apple WebKit, Chrome, Edge, Facebook, Safari) and whether the consumer was browsing on mobile. We find no significant association at the $\alpha = 10\%$ level between treatment assignment and both the browser type, $\chi^2(76) = 82.82$, $p = 0.28$, and whether the consumer was browsing on mobile, $\chi^2(19) = 13.45$, $p = 0.81$.

5. Empirical Analyses

5.1. Model-Free Evidence

In Figure 4, we display model-free evidence focusing on the primary hypothesis of interest (how femininity in headshots impacts discriminatory behavior). We include conditional means of outcomes by two different discretizations of femininity, at the headshot level: median split and tertiles. Each femininity median split (below $\in [-0.48, 0.33]$, above $\in [0.40, 0.70]$) corresponds to six headshots. Each femininity tertile (low $\in [-0.48, -0.10]$, medium $\in [0.31, 0.49]$, high $\in [0.56, 0.70]$) corresponds to four headshots. Results are overall estimated imprecisely although the pattern of means across both discretizations suggests that consumers discriminate against femininity for all stages of the purchase funnel considered despite spending comparable amounts of time on the platform across different levels of femininity. Because these results do not account for potential latent confounders or include covariates to improve precision, we turn to model-based evidence using a regression.

5.2. Model-Based Evidence

To describe notation, each observation corresponds to a participant i who sees headshot h (belonging to synthetic identity k) paired with pronoun p. We estimate the following main ordinary least squares (OLS) regression specification[6] with robust standard errors clustered by synthetic identity:[7]

$$\mathrm{DV}_{ihp} = \boldsymbol{\alpha} \cdot \mathbf{Fem(ininity)\ Tertile}_h + \boldsymbol{\psi} \cdot \mathbf{Controls}_h + \boldsymbol{\gamma}_k + \varepsilon_{ihp}. \tag{1}$$

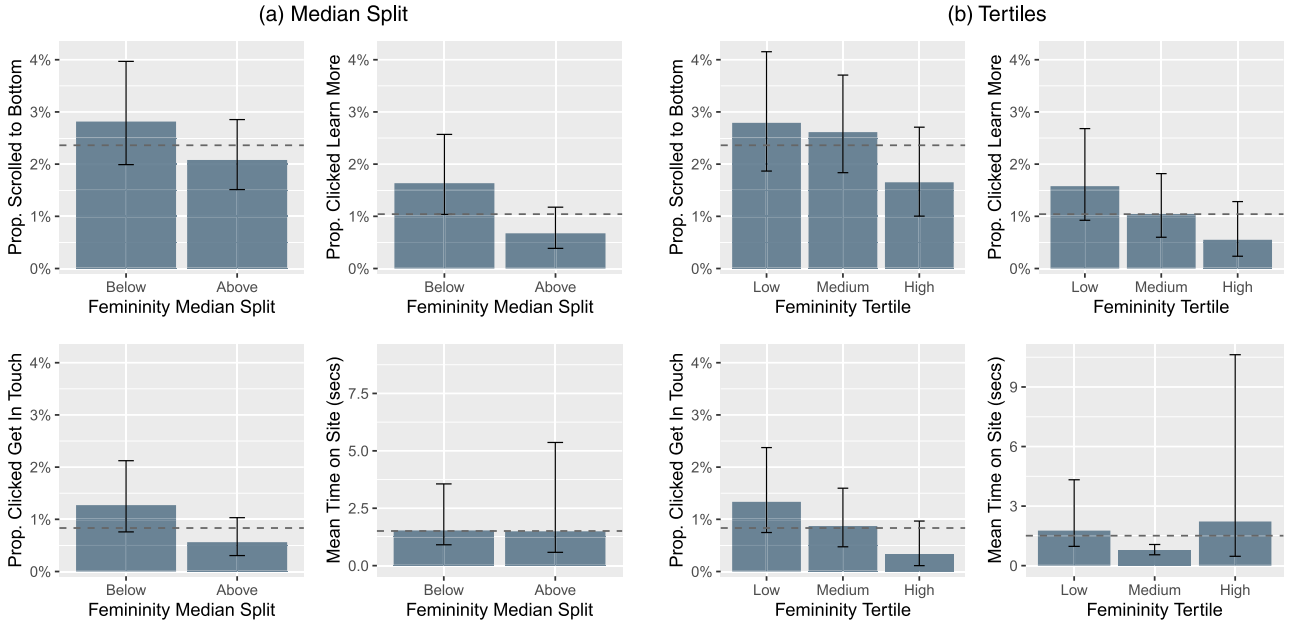
We discretize the continuous femininity measure into tertiles to allow for nonparametric treatment effects (e.g., instead of assuming a linear functional form) and to reduce measurement error (see Online Appendix D.2.4).[8] We use $\mathbf{Controls}_h$ to further control for the other potential confounders discovered in Section 4.2.5 (i.e., makeup, facial hair, earrings, smile, and age) that vary within synthetic identities. We also control for perceived trustworthiness (operationalized as headshot-level averages using ratings from Section 4.2.4) to distinguish the effect of femininity from trustworthiness.[9] As preregistered, the fixed effect $\boldsymbol{\gamma}_k$ is included as a covariate to improve precision. Some synthetic identities do not span all three tertiles (see Figure 3), in which case those synthetic identities are only relevant for certain femininity treatment comparisons (e.g., only low versus medium; see Online Appendix D.2.4).

The full OLS regression specification also involves gender pronouns:

$$\begin{aligned}\mathrm{DV}_{ihp} = {} & \boldsymbol{\alpha} \cdot \mathbf{Fem(ininity)\ Tertile}_h + \boldsymbol{\beta} \cdot \mathbf{Pronoun}_p \\ & + \boldsymbol{\theta} \cdot \mathbf{Fem\ Tertile}_h \times \mathbf{Pronoun}_p \\ & + \boldsymbol{\psi} \cdot \mathbf{Controls}_h + \boldsymbol{\gamma}_k + \varepsilon_{ihp}.\end{aligned} \tag{2}$$

We consider medium femininity and "he" pronouns as baselines.

Figure 4. Means of Outcomes by Femininity Discretizations



Notes. Error bars represent 95% confidence intervals, which are computed by Wilson score intervals for binary outcomes and by bias-corrected and accelerated bootstrap for continuous outcomes (using 5,000 resamples). Unconditional means are included as gray dotted lines.

We report regression results in Table 2. As shorthand, we refer to "he" pronouns as male tutors, "she" pronouns as female tutors, and "they" pronouns as nonbinary tutors. To note when interpreting results, discrimination occurs when "members of a minority group … are treated differentially (less favorably) than members of a majority group with otherwise identical characteristics in similar circumstances" (Bertrand and Duflo 2017, p. 310). In our context, discrimination may occur if consumers decline to interact with the education service upon seeing a tutor who scores high on femininity but not a tutor who scores low on this measure despite those tutors having identical qualifications and looking otherwise identical. This discrimination may occur, for example, because of a pure distaste for interacting with such tutors (Bertrand and Duflo 2017) or because of beliefs (whether accurate or not) about average group differences in unobserved productivity (e.g., teaching skills, abilities to create rapport) (Coffman et al. 2021).

In terms of main effects (reflecting Equation (1)), femininity significantly decreased the probability of scrolling (to the bottom of the homepage), clicking "Learn More," and clicking "Get in Touch," considering both high relative to medium femininity and medium relative to low femininity (columns (1), (3), and (5)). Such discrimination occurs even though consumers spend significantly more time on the platform as femininity increases, considering both high relative to medium femininity and medium relative to low femininity (column (7)). These effects are somewhat small but meaningful in magnitude: relative to low femininity, high femininity significantly decreases the probability of scrolling by 4%, clicking "Learn More" by 2.3%, and clicking "Get in Touch" by 2.9% (such that treatment effect sizes generally diminish with purchase funnel depth) and significantly increases time spent on the platform by 20.1 seconds. Although corresponding R^2 values are low (columns (1), (3), (5), and (7)), the improvement in variation explained relative to specifications using only controls (see Online Table S5, columns (1), (3), (5), and (7)) is on par with the improvement for specifications that just adds gender pronouns (see Online Table S5; columns (2), (4), (6), and (8)). This suggests that, benchmarked against gender pronouns, femininity explains variation in behavior to a comparable (albeit smaller) degree.

We now turn to the Equation (2) specification. For male tutors, relative to medium femininity, high femininity (i.e., High Fem Tertile main effects) significantly decreased the probability of clicking "Learn More" (column (4)); effects are directionally consistent but not significant for scrolling (column (2)) and clicking "Get in Touch" (column (6)). For male tutors, relative to medium femininity, low femininity (i.e., Low Fem Tertile main effects) significantly increased the probability of scrolling (column (2)) and clicking "Get in Touch" (column (6)); effects are directionally consistent but not significant for clicking "Learn More" (column (4)). These findings contrast with prior (laboratory-based) psychology literature (Sutherland et al. 2015), which finds that counter-stereotypical male faces were not negatively evaluated.

Table 2. Impact of Femininity Tertiles and Gender Pronouns on Real-Life Outcomes (OLS)

	Scrolling to bottom		Clicking "Learn More"		Clicking "Get in Touch"		Time on site, s	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Low Fem(ininity) Tertile	0.006***	0.038*	0.015***	0.016	0.018***	0.034***	−6.65***	1.30***
	(3.5×10^{-14})	(0.009)	(9.19×10^{-15})	(0.009)	(1.22×10^{-14})	(0.001)	(1.87×10^{-11})	(0.091)
High Fem Tertile	−0.034***	−0.046	−0.008***	−0.005**	−0.010***	−0.018	13.5***	15.0***
	(1.11×10^{-13})	(0.017)	(2.36×10^{-14})	(0.0008)	(2.27×10^{-14})	(0.010)	(5.41×10^{-11})	(0.336)
She		−0.011*		−0.007		0.003		−0.004
		(0.002)		(0.005)		(0.003)		(0.017)
They		0.0004		−0.011		0.009		−0.002
		(0.024)		(0.008)		(0.008)		(0.397)
Low Fem Tertile × She		−0.033***		−0.008		−0.006		−1.14***
		(0.002)		(0.005)		(0.003)		(0.017)
Low Fem Tertile × They		−0.045		−0.012		−0.005		−2.07*
		(0.024)		(0.008)		(0.008)		(0.397)
High Fem Tertile × She		0.0002		0.008		−0.009*		−8.84***
		(0.002)		(0.005)		(0.003)		(0.017)
High Fem Tertile × They		−0.001		0.013		−0.008		−9.05***
		(0.024)		(0.008)		(0.008)		(0.397)
Synthetic Identity fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Perceived makeup control	✓	✓	✓	✓	✓	✓	✓	✓
Perceived facial hair control	✓	✓	✓	✓	✓	✓	✓	✓
Perceived earrings control	✓	✓	✓	✓	✓	✓	✓	✓
Perceived smile control	✓	✓	✓	✓	✓	✓	✓	✓
Perceived age control	✓	✓	✓	✓	✓	✓	✓	✓
Perceived trustworthiness control	✓	✓	✓	✓	✓	✓	✓	✓
Observations	2,880	2,880	2,880	2,880	2,880	2,880	2,879	2,879
R^2	0.004	0.007	0.004	0.006	0.004	0.005	0.002	0.005
Within R^2	0.002	0.006	0.001	0.003	0.002	0.003	0.001	0.005

Note. Robust standard errors clustered by synthetic identity are reported in parentheses.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

For female tutors, the negative effects of high femininity become significant for scrolling and clicking "Get in Touch" (see Online Appendix F.3). For female tutors, the positive effects of low femininity become insignificant for scrolling (but are directionally consistent), become significant for clicking "Learn More" and remain significant for clicking "Get in Touch" (see Online Appendix F.3). These findings contrast with Sutherland et al. (2015) and Oh et al. (2020), who find that stereotypical female faces were more positively evaluated.

For nonbinary tutors, the effects of high femininity are not significant. For nonbinary tutors, the positive effects of low femininity become significant for clicking "Learn More" and remain significant for clicking "Get in Touch" (see Online Appendix F.3).

Altogether, these results suggest that consumers consistently discriminate against femininity at various stages of the purchase funnel to a largely similar extent for men, women, and nonbinary people. These overall findings contrast with prior psychology literature based on laboratory studies (Perrett et al. 1998, Said and Todorov 2011) and our own laboratory studies (in which we find that femininity significantly increased self-reported perceptions of trustworthiness and competence for all gender pronouns considered). In part, this discrepancy may demonstrate how pernicious discrimination can be: people can express certain views in surveys yet act with prejudice in consequential settings. Despite their high cost and logistical complexity, field experiments are useful for shedding more light on such effects.

Furthermore, high femininity significantly increased time spent on the platform for male tutors relative to medium (column (8), High Fem Tertile main effect) and low femininity (see Online Appendix F.3). This pattern of results persists for female and nonbinary tutors (see Online Appendix F.3). Altogether, these findings suggest that consumers take the longest when considering tutors with high femininity (for male, female, and nonbinary tutors) and, even so, discriminate against them.

Finally, although treatment effects seem predominately related to femininity, consumers still attend to gender pronouns. For example, among tutors with medium femininity, female tutors are discriminated against in terms of scrolling relative to male tutors (column (2), "she" main effect), an effect that disappears with the metrics further down the purchase funnel.

To note, we find evidence of gender-based discrimination even though our sample was primarily female

(63% women). As a caveat, we are unable to formally investigate heterogeneous treatment effects with consumer characteristics (e.g., to test homophily) because Meta only provides aggregate, campaign-level information. Our results may be driven by a general bias to perceive feminine faces as less competent (Oh et al. 2019) as well as gender stereotypes (Coffman et al. 2021, Heilman et al. 2024), conscious prejudices (Becker 1957), and/or unconscious biases (Bertrand et al. 2005) that may exist in our specific online tutoring context. These findings pose important considerations for equal opportunity of workers in labor markets. Namely, worker characteristics such as gender pronouns and femininity (even involving overall subtle and minimal visual changes) that are orthogonal to one's qualifications and relatively stable over time can lead to discrimination. This could substantially decrease, for example, feminine workers' demand and, thus, productivity, leading managers to be less likely to retain and hire such workers and, thus, promote further employment discrimination (Kelley et al. 2026).

6. Discussion

Characterizing gender discrimination in society and marketplaces is crucial for welfare, equitability, and legal considerations. This research, by leveraging methodologies and approaches at the nexus of marketing, computer science, political science, economics, and psychology, explores an understudied form of gender discrimination through femininity and its interaction with gender identity (proxied by gender pronouns). Future research should assess its prevalence in other substantive and cultural contexts,[10] further refine our understanding of the headshot features that drive femininity effects, and design effective policy solutions. Interventions may involve initially concealing information in marketplaces such as headshots and gender pronouns (that are later revealed) to place more emphasis on credible qualifications (e.g., reviews; Cui et al. 2020) during the consideration stage, "extended and imagined contact" (Paluck et al. 2021, p. 547) by encouraging consumers to envision positive interactions with service providers, and antibias education or diversity training for consumers while noting limitations in their empirical efficacy (Paluck et al. 2021).

We view our methodological approach as a general-purpose tool that can be used to address a variety of other research questions involving latent treatment effects. One could study discrimination, implicit associations, algorithmic bias, or how spokespeople/influencers affect brand impressions along a variety of other headshot attributes (e.g., charisma, attractiveness, smile; see Online Figure S7). More broadly, controllable stimuli generation can be applied to other unstructured data contexts. Specific examples include assessing the impact of image and text attributes in the packaging of vice products such as processed snacks on consumer preferences; of accents or emotional tones of medical AI assistants on drug adoption; and of decreasing toxicity in social media posts (political marketing messages) on well-being (word of mouth).

This research presents a few limitations. First, our approach necessitates an experimental design which requires laboratory studies (that may entail some degree of response bias) or costly field experiments such as the one we report. Second, deep generative models capable of manipulating latent treatments with high fidelity (e.g., GANs, diffusion models, normalizing flows) are typically black box and, thus, not well understood. Third, controllable stimuli generation may also change imperceptible latent attributes (e.g., subtle adjustments to the background) that affect the outcome. No latent confounding is an ultimately untestable assumption (Fong and Grimmer 2023) though we believe it to be convincing in our analyses and plausibly numerous others. At the very least, using our conceptual approach to control for many, if not all, latent confounds is a significant improvement over extant options. We envision a fruitful line of research leveraging controllable stimuli generation and experimentation across scientific disciplines that will be enriched by concurrent developments in statistics and computer science.

Acknowledgments

The authors thank Shin Oblander, Andrey Simonov, Hortense Fong, Melanie Brucks, George Gui, Christopher Olivola, Kohei Onzo, Eric Park, Sonia Kim, and participants in the 12th Triennial Invitational Choice Symposium in the Probabilistic Machine Learning session at INSEAD and the Quantitative Marketing Lab at Columbia for their helpful comments. The authors thank Daniel Joseph Merlau and Kameron Bobbitt for excellent research assistance. The corresponding author's legal name is Lan Luo. This paper is based on part of the corresponding author's doctoral dissertation at Columbia Business School. The authors have no financial or nonfinancial interests to disclose in the subject matter or materials discussed in this manuscript.

Endnotes

[1] This study was approved by the Columbia University institutional review board (protocol number: AAAT4855). In particular, we debriefed (with some delay in timing) all participants in pilots or the main study who left an email, informing them that the tutoring program does not exist.

[2] In designing the website, we made it similar to and conceptually representative of the numerous other popular, general tutoring services available online.

[3] Although we use persistent cookies, participants are initially determined at the Meta user account level. User accounts who landed on the website (as determined by a Meta pixel we installed on the page) were excluded from viewing the ad again.

[4] We were unable to examine three preregistered dependent variables (i.e., willingness to pay, interest in service, and whether or not

someone left an email) because no participant answered the survey after clicking "Get in Touch."

[5] We conceptualize femininity and masculinity as two ends of the same continuum following prior psychology literature (e.g., Perrett et al. 1998, Scott et al. 2014, Zietsch et al. 2015). Using two separate scales for femininity and masculinity could be an interesting future direction to operationalize this construct (as Hester et al. 2021 suggest; to note, their key finding is the revelation of facial androgyny as a consequence of using two scales, which we incorporate directly as the labeled midpoint of our single scale).

[6] We deviate from preregistration by opting for a linear probability model instead of assuming a logit functional form because only the former is directly interpretable and unbiased for average treatment effects. We thank an anonymous reviewer for this recommendation.

[7] Clustering is motivated by a sampling mechanism (Abadie et al. 2023), in which the first stage selects clusters (i.e., synthetic identities) at random from an infinite population (mirroring how StyleGAN2 mechanically produces synthetic identities by draws from a Gaussian distribution), followed by a second stage of selecting units (i.e., particular headshots) from the sampled clusters.

[8] Results are largely consistent when using a more continuous measure of femininity (see Online Appendix F.4). We opt to discretize irrespective of synthetic identity so as to be able to interpret femininity effects in absolute terms (see Online Appendix D.2.4) although results are largely consistent when operationalizing femininity via treatment levels (see Online Appendix F.5).

[9] We deviate from preregistration by using femininity tertiles, controlling for latent confounders, and controlling for perceived trustworthiness. We thank the reviewers for these improvements to our specification.

[10] To note, real workers may sort away from discriminating customers so that discrimination could be competed away in general equilibrium (Becker 1957). Future research could potentially quantify the amount of discrimination that happens by gender identity and femininity to real, employed workers (Kelley et al. 2026).

References

Abadie A, Athey S, Imbens GW, Wooldridge JM (2023) When should you adjust standard errors for clustering? *Quart. J. Econom.* 138(1):1–35.

Acquisti A, Fong C (2020) An experiment in hiring discrimination via online social networks. *Management Sci.* 66(3):1005–1024.

Athey S, Karlan D, Palikot E, Yuan Y (2025) Smiles in profiles: Improving fairness and efficiency using estimates of user preferences in online marketplaces. NBER Working Paper No. 30633, National Bureau of Economic Research, Cambridge, MA.

Becker GS (1957) *The Economics of Discrimination* (University of Chicago Press, Chicago).

Bertrand M, Duflo E (2017) Field experiments on discrimination. Banerjee AV, Duflo E, eds. *Handbook of Economic Field Experiments*, vol. 1 (North-Holland, Amsterdam), 309–393.

Bertrand M, Chugh D, Mullainathan S (2005) Implicit discrimination. *Amer. Econom. Rev.* 95(2):94–98.

Braun M, de Langhe B, Puntoni S, Schwartz EM (2024) Leveraging digital advertising platforms for consumer research. *J. Consumer Res.* 51(1):119–128.

Brock JM, De Haas R (2023) Discriminatory lending: Evidence from bankers in the lab. *Amer. Econom. J. Appl. Econom.* 15(2):31–68.

Busse MR, Israeli A, Zettelmeyer F (2017) Repairing the damage: The effect of price knowledge and gender on auto repair price quotes. *J. Marketing Res.* 54(1):75–95.

Chan J, Wang J (2018) Hiring preferences in online labor markets: Evidence of a female hiring bias. *Management Sci.* 64(7):2973–2994.

Coffman KB, Exley CL, Niederle M (2021) The role of beliefs in driving gender discrimination. *Management Sci.* 67(6):3551–3569.

Cui R, Li J, Zhang DJ (2020) Reducing discrimination with reviews in the sharing economy: Evidence from field experiments on Airbnb. *Management Sci.* 66(3):1071–1094.

DeBruine LM, Jones BC, Smith FG, Little AC (2010) Are attractive men's faces masculine or feminine? The importance of controlling confounds in face stimuli. *J. Experiment. Psych. Human Perception Performance* 36(3):751–758.

Dew R, Ansari A, Toubia O (2022) Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Sci.* 41(2):401–425.

Evsyukova Y, Rusche F, Mill W (2025) LinkedOut? A field experiment on discrimination in job network formation. *Quart. J. Econom.* 140(1):283–334.

Feng X, Zhang S, Liu X, Srinivasan K, Lamberton C (2025) An AI method to score celebrity visual potential. *J. Marketing Res.* 62(5): 757–775.

Fong C, Grimmer J (2023) Causal inference with latent treatments. *Amer. J. Political Sci.* 67(2):374–389.

Heilman ME, Caleo S, Manzi F (2024) Women at work: Pathways from gender stereotypes to gender bias and discrimination. *Annual Rev. Organ. Psych. Organ. Behav.* 11(1):165–192.

Hester N, Jones BC, Hehman E (2021) Perceived femininity and masculinity contribute independently to facial impressions. *J. Experiment. Psych. General* 150(6):1147–1164.

Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of StyleGAN. *Proc. 2020 IEEE/CVF Conf. Comput. Vision Pattern Recognition* (IEEE, Piscataway, NJ), 8107–8116.

Kelley E, Lane G, Pecenco M, Rubin E (2026) Customer discrimination in the workplace: Evidence from online sales. *J. Labor Econom.* 44(3):855–889.

Kline P, Rose EK, Walters CR (2022) Systemic discrimination among large U.S. employers. *Quart. J. Econom.* 137(4):1963–2036.

Lambrecht A, Tucker C (2013) When does retargeting work? Information specificity in online advertising. *J. Marketing Res.* 50(5):561–576.

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

Liang H, Perona P, Balakrishnan G (2023) Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation. *Proc. 2023 IEEE/CVF Internat. Conf. Comput. Vision* (IEEE, Piscataway, NJ), 4954–4964.

Ludwig J, Mullainathan S (2024) Machine learning as a tool for hypothesis generation. *Quart. J. Econom.* 139(2):751–827.

Milkman KL, Akinola M, Chugh D (2015) What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *J. Appl. Psych.* 100(6):1678–1712.

Nakamura K, Watanabe K (2020) A new data-driven mathematical model dissociates attractiveness from sexual dimorphism of human faces. *Sci. Rep.* 10(1):16588.

Nightingale SJ, Farid H (2022) AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proc. Natl. Acad. Sci. USA* 119(8):e2120481119.

Oh D, Buck EA, Todorov A (2019) Revealing hidden gender biases in competence impressions of faces. *Psych. Sci.* 30(1):65–79.

Oh D, Dotsch R, Porter J, Todorov A (2020) Gender biases in impressions from faces: Empirical studies and computational models. *J. Experiment. Psych. General* 149(2):323–342.

Olivola CY, Todorov A (2010) Elected in 100 milliseconds: Appearance-based trait inferences and voting. *J. Nonverbal Behav.* 34(2):83–110.

Ozturk OC, He C, Chintagunta PK (2024) Frontiers: Inequalities in dealers' interest rate markups? A gender- and race-based analysis. *Marketing Sci.* 43(1):20–32.

Paluck EL, Porat R, Clark CS, Green DP (2021) Prejudice reduction: Progress and challenges. *Annual Rev. Psych.* 72(1):533–560.

Patashnik O, Wu Z, Shechtman E, Cohen-Or D, Lischinski D (2021) StyleCLIP: Text-driven manipulation of StyleGAN imagery. *Proc. 2021 IEEE/CVF Internat. Conf. Comput. Vision* (IEEE, Piscataway, NJ), 2065–2074.

Perrett DI, Lee KJ, Penton-Voak I, Rowland D, Yoshikawa S, Burt DM, Henzi SP, Castles DL, Akamatsu S (1998) Effects of sexual dimorphism on facial attractiveness. *Nature* 394(6696):884–887.

Peterson JC, Uddenberg S, Griffiths TL, Todorov A, Suchow JW (2022) Deep models of superficial face judgments. *Proc. Natl. Acad. Sci. USA* 119(17):e2115228119.

Proserpio D, Troncoso I, Valsesia F (2021) Does gender matter? The effect of management responses on reviewing behavior. *Marketing Sci.* 40(6):1199–1213.

Said CP, Todorov A (2011) A statistical model of facial attractiveness. *Psych. Sci.* 22(9):1183–1190.

Scott IM, Clark AP, Josephson SC, Boyette AH, Cuthill IC, Fried RL, Gibson MA, et al. (2014) Human preferences for sexually dimorphic faces may be evolutionarily novel. *Proc. Natl. Acad. Sci. USA* 111(40):14388–14393.

Sisodia A, Burnap A, Kumar V (2025) Generative interpretable visual design: Using disentanglement for visual conjoint analysis. *J. Marketing Res.* 62(3):405–428.

Sutherland CAM, Young AW, Mootz CA, Oldmeadow JA (2015) Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British J. Psych.* 106(2):186–208.

Todorov A, Olivola CY, Dotsch R, Mende-Siedlecki P (2015) Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Rev. Psych.* 66(1):519–545.

Troncoso I, Luo L (2023) Look the part? The role of profile pictures in online labor markets. *Marketing Sci.* 42(6):1080–1100.

Wu Z, Lischinski D, Shechtman E (2021) StyleSpace analysis: Disentangled controls for StyleGAN image generation. *Proc. 2021 IEEE/CVF Conf. Comput. Vision Pattern Recognition* (IEEE, Piscataway, NJ), 12858–12867.

Xiao L, Ding M (2014) Just the faces: Exploring the effects of facial features in print advertising. *Marketing Sci.* 33(3):338–352.

Zhang S, Friedman EMS, Srinivasan K, Dhar R, Zhang X (2025) Serving with a smile on Airbnb: Analyzing the economic returns and behavioral underpinnings of the host's smile. *J. Consumer Res.* 51(6):1073–1097.

Zietsch BP, Lee AJ, Sherlock JM, Jern P (2015) Variation in women's preferences regarding male facial masculinity is better explained by genetic differences than by previously identified context-dependent effects. *Psych. Sci.* 26(9):1440–1448.

Lan E. Luo is an assistant professor of marketing at the Yale School of Management. He develops and applies methods at the intersection of interpretable machine learning, applied econometrics, and probabilistic machine learning to draw causal insights from unstructured data such as images and text, in areas including data-driven design, user-generated content, digital advertising, and discrimination. He received his PhD in marketing from Columbia University.

Olivier Toubia is the Glaubinger Professor of Business at Columbia Business School. His research focuses primarily on innovation, customer insights, and creative industries, combining methods from the social sciences and data science to study human processes such as motivation, choice, and creativity. He previously served as editor-in-chief of *Marketing Science*. He received his MS in operations research and his PhD in marketing from the Massachusetts Institute of Technology.