Web Appendix for "Discrimination Against Femininity in Headshots: A Field Experiment with AI-Enabled Controllable Stimuli Generation"

Lan E. Luo
Yale School of Management

Olivier Toubia
Columbia Business School

July 27, 2026

Web Appendix A Overview of Generative Adversarial Networks

Generative models are instances of unsupervised learning, or learning from unlabelled data. Conceptually, they learn to represent a probability distribution, p_{model}, that estimates the training data's distribution, p_{data}, in order to generate new samples (Goodfellow 2017). GANs are a class of deep generative models that address implicit density estimation, sampling from p_{model} without explicitly defining it.[1] GANs are composed of two neural networks: a generator $G_{\boldsymbol{\phi}}$ and a discriminator $D_{\boldsymbol{\omega}}$, parameterized by $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$, respectively (Goodfellow et al. 2014). The generator receives as input a vector $\mathbf{z}$ sampled from a standard normal prior $p(\mathbf{z})$ and outputs an image, as an array of pixel values.[2] This image is fed into the discriminator, which determines the probability that the image is real (i.e., from the dataset $\mathbf{x}$ rather than the generator). During training, the objective function corresponds to a minimax (or adversarial) game, in which the generator tries to fool the discriminator by producing increasingly realistic images:

$$\min_{\boldsymbol{\phi}} \max_{\boldsymbol{\omega}} \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{\text{data}}} \Big[\log \underbrace{D_{\boldsymbol{\omega}}(\mathbf{x})}_{\substack{\text{recognize real} \\ \text{images as "real"}}} \Big] + \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})} \Big[\log \Big(\underbrace{1 - D_{\boldsymbol{\omega}}\big(\overbrace{G_{\boldsymbol{\phi}}(\mathbf{z})}^{\substack{\text{generate fake images} \\ \text{that look "real" to } D_{\boldsymbol{\omega}}}}\big)}_{\substack{\text{recognize fake} \\ \text{images as "fake"}}} \Big) \Big] \tag{1}$$

This game's solution is a local differential Nash equilibrium (Ratliff et al. 2013), in which ideally $G_{\boldsymbol{\phi}^*}(\mathbf{z})$ is drawn from the same distribution as the training data and $D_{\boldsymbol{\omega}^*}(\mathbf{x}) = 1/2$. In other words, in equilibrium, the generator creates photorealistic images that are equally likely to be classified as real or fake by the discriminator. In practice, the objective function is modified using heuristics for better results (e.g., Salimans et al. 2016). GANs can also be used for other applications (Dew et al. 2024), such as generating artificial data (Athey et al. 2021; Anand and Lee 2023); estimating price markups and targeting consumers (Anand and Lee 2023); and estimating structural models (Kaji et al. 2023).

[1] This is in contrast to models of explicit density estimation like variational autoencoders (Kingma and Welling 2013), which define and solve for p_{model} (e.g., with an approximate density).

[2] In StyleGAN2 specifically, $\mathbf{z} \in \mathbb{R}^{512}$ (in the latent space $\mathcal{Z}$); and several images are outputted, with dimensions of $4 \times 4 \times 3$ to $1024 \times 1024 \times 3$, in increasing powers of two for the image resolution.

Web Appendix B Examples of Controllable Stimuli Generation in Prior Literature

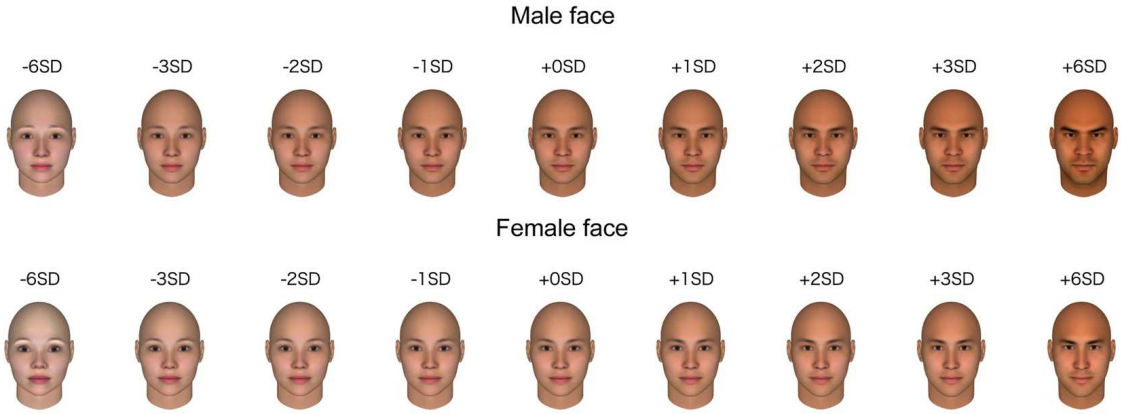
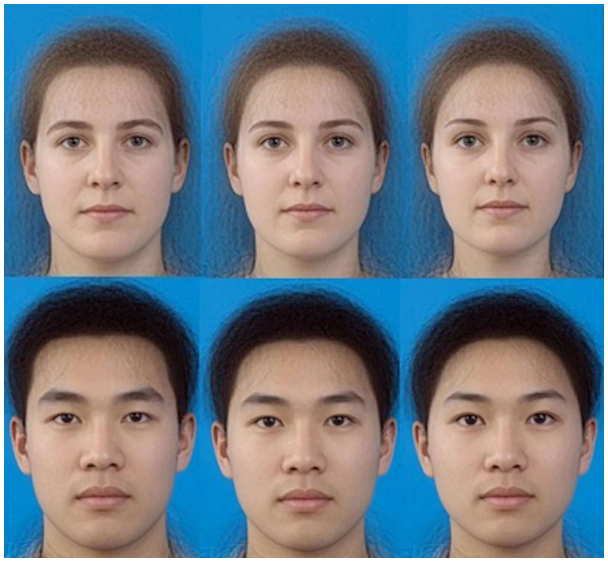
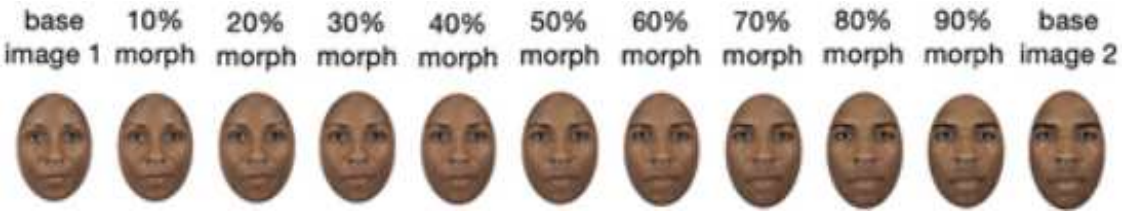


Figure S1: Example of Using FaceGen Modeller to Manipulate Femininity
(From Figure 4 of Nakamura and Watanabe 2020)



(a) From Figure 1 of Scott et al. (2014)

(b) From Figure 1 of Zietsch et al. (2015)



(c) From Figure 1 of Atwood et al. (2024)

Figure S2: Examples of Using Psychomorph to Manipulate Femininity

Figure S3: Manipulating Age (From Figure S11 of Peterson et al. 2022)

Figure S4: Manipulating Demographics (From Figure 1 of Liang et al. 2023)

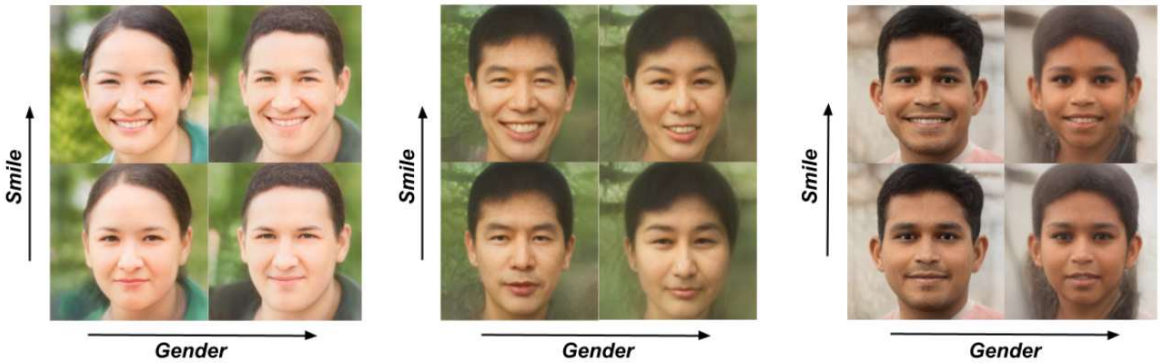
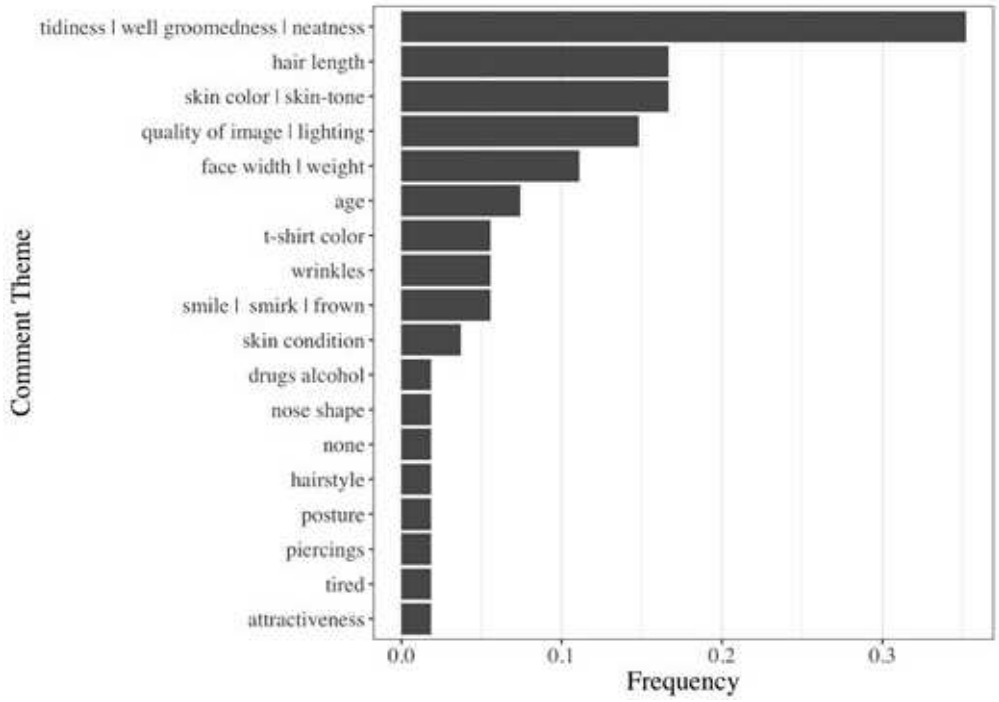


Figure S5: Manipulating Smile and Femininity, Before Post-Processing (From Figure 15 of Athey et al. 2023)

(a) Manipulating Mugshots on Predicted Detention Probability (From Figure VII)



(b) Themes Identified By Humans on How Stimuli Differ (From Figure VIII(B))

Figure S6: Morphing Generated Mugshots in the 1st Hypothesis Generation Step (From Ludwig and Mullainathan 2024)

Web Appendix C Identifying Relevant Dimensions in $\mathcal{S}$

We adapt Patashnik et al. (2021) to our context, using textual descriptions to query $\mathcal{S}$ with embeddings from CLIP (Radford et al. 2021), an extensively pretrained multimodal model. CLIP has two main components: a text encoder (with a Transformer architecture; Vaswani et al. 2017) and an image encoder (with a Vision Transformer ViT-B/32 architecture; Dosovitskiy et al. 2021) that both map to the same 512-dimensional representation space (henceforth, CLIP space). In other words, raw text (images, as an array of pixel values) can be fed as input into the text (image) encoder component of CLIP, which will then output a 512-dimensional numerical vector. CLIP aligns images and text in this common, low-dimensional representation such that a given piece of text (image) will be embedded near conceptually similar texts and images (images and texts).

For style parameters $\mathbf{s} \in \mathbb{R}^{9088}$ (in the latent space $\mathcal{S}$), let $G(\mathbf{s})$ denote the corresponding generated image. We use text to specify a target and baseline attribute: "a photo of a female face" and "a photo of a male face," respectively. For robustness, we augment these textual prompts with a bank of 80 templates (e.g., "a cropped photo of…", "a close-up photo of…", "a rendition of…") from Radford et al. (2021). We use CLIP's text encoder to extract an average embedding for the target and baseline attributes and then compute their normalized difference. We define this normalized difference as the textual query $\Delta\mathbf{t} \in \mathbb{R}^{512}$ (in CLIP space).

We assume $\Delta\mathbf{t}$ is collinear with the difference between the corresponding target and baseline *visual* attributes over images in CLIP space, which we define as $\Delta\mathbf{i} \in \mathbb{R}^{512}$. Then, we can identify a direction $\Delta\mathbf{s} \in \mathbb{R}^{9088}$ (in the latent space $\mathcal{S}$) which manipulates femininity by assessing the relevance of each dimension in $\mathcal{S}$ to the textual query $\Delta\mathbf{t}$. To do so, for each dimension j in $\mathcal{S}$, we generate 100 random image pairs that perturb only that dimension: $G(\mathbf{s} \pm \alpha\Delta\mathbf{p}_j)$, where α determines the strength of the perturbation (and is set to $\alpha = 5$ here) and $\Delta\mathbf{p}_j \in \mathbb{R}^{9088}$ (in the latent space $\mathcal{S}$) is a vector with the jth dimension set to its standard deviation and zeros elsewhere. For each dimension j, we use CLIP's image encoder to extract an average embedding for the $G(\mathbf{s} - \alpha\Delta\mathbf{p}_j)$ images and $G(\mathbf{s} + \alpha\Delta\mathbf{p}_j)$ images and then compute their difference. We define this difference as $\Delta\mathbf{i}_j \in \mathbb{R}^{512}$ (in CLIP space). Then, for some disentanglement threshold β and for each dimension

j in $\mathcal{S}$, we define:

$$\Delta \mathbf{s}_j = \begin{cases} \Delta \mathbf{t} \cdot \Delta \mathbf{i}_j & \text{if } |\Delta \mathbf{t} \cdot \Delta \mathbf{i}_j| \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

$\Delta \mathbf{s}$ sets values for only the latent dimensions relevant for femininity (i.e., those with a large enough dot product between the textual query and the visual manipulations for each latent dimension). Higher β results in more disentangled manipulations, since fewer latent dimensions would be identified. We set $\beta = .4$ (which Patashnik et al. 2021 define as high), identifying two relevant latent dimensions in our setting. Of the 18 layers of the generator, these latent dimensions are at the 5th and 8th layers, operating at 16×16 and 32×32 levels of image resolution, respectively.

To generate the stimuli from Figure S7, we specify the baseline attribute as "a photo of a face" and the target attribute as "a photo of a (charismatic / attractive / smiling) face." For each attribute, we set β to allow for meaningful changes in the headshot, while still maintaining a considerable level of disentanglement. We set α to a level that noticeably manipulates the attribute of interest. To be specific, for charisma, we set $\alpha = 6$ and $\beta = .1$, which identifies 20 relevant latent dimensions at the 1st, 3rd, 4th, 5th, 6th, and 7th layers, operating at 4×4, 8×8, 8×8, 16×16, 16×16, and 32×32 levels of image resolution, respectively. For attractiveness, we set $\alpha = 2$ and $\beta = .1$, which identifies 65 relevant latent dimensions at the 1st, 3rd, 4th, 5th, 6th, 7th, 8th, 9th, 10th, 11th, 12th, 13th, 15th, and 17th layers, operating at 4×4, 8×8, 8×8, 16×16, 16×16, 32×32, 32×32, 64×64, 64×64, 128×128, 128×128, 256×256, 512×512, and 1024×1024 levels of image resolution, respectively. For smile, we set $\alpha = 1$ and $\beta = .3$, which identifies just one relevant latent dimension at the 6th layer, which operates at a 16×16 level of resolution. Intuitively, we find that abstract attributes like attractiveness relate to more course-grained features at higher levels of resolution. More low-level, concrete headshot features such as femininity and smiling relate to a smaller number of features at lower levels of resolution.

(a) Base Image (b) ↑ Charisma (c) ↑ Attractiveness (d) ↑ Smile

Figure S7: Manipulating Headshots on Marketing-Relevant Attributes

Web Appendix D Stimuli Design

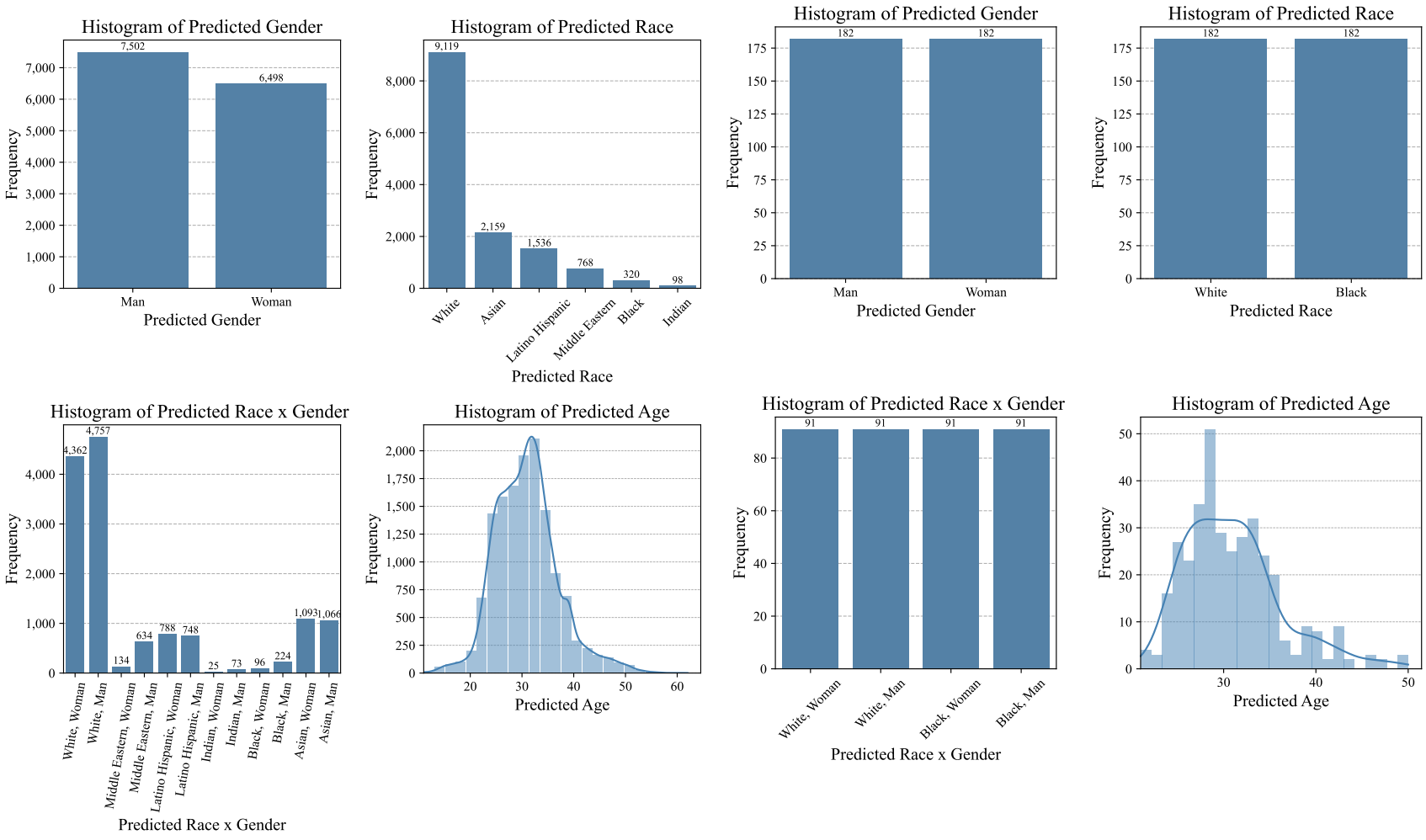
D.1 Determining baseline and manipulated images

We randomly generate seed states to produce a set of 14,000 $\mathbf{z} \in \mathbb{R}^{512}$ vectors (in the latent space $\mathcal{Z}$; see Section 3 for a summary of the generator's latent spaces) sampled from a standard normal prior and generate their corresponding images at full 1024×1024 resolution. Then, we use *deepFace* (Serengil and Ozpinar 2021), a collection of state-of-the-art pretrained neural networks, to predict the gender, race, and age of each image (see Figure S8a). We filter for the headshots between the predicted ages of 21 and 50 and confidently predicted (i.e., softmax probabilities $>$ 0.95) to be either White or Black. Then, we randomly sample to balance by predicted gender and race, leaving 364 potential synthetic identities (see Figure S8b).

For each synthetic identity, we create three versions: one baseline image and two manipulated versions. First, we take the filtered set of 364 $\mathbf{z} \in \mathbb{R}^{512}$ vectors (in the latent space $\mathcal{Z}$) and convert them to their corresponding $\mathbf{s} \in \mathbb{R}^{9088}$ vectors (in the latent space $\mathcal{S}$). Then, we manipulate femininity by $G(\mathbf{s} + \alpha\Delta\mathbf{s}_{fem})$, where α determines the strength/direction of the manipulation and $\Delta\mathbf{s}_{fem} \in \mathbb{R}^{9088}$ is a vector with the dimensions identified from Web Appendix C set to their standard deviations and zeros elsewhere. For images predicted as men, we set $\alpha = \{0, 1, 2\}$; and for images predicted as women, we set $\alpha = \{0, -1, -2\}$. In other words, for predicted men (women), we manipulate the baseline image, $\alpha = 0$, by increasing (decreasing) the values of the relevant latent dimensions by one as well as two standard deviations.

We calculate the root-mean-square difference of the color histogram between each baseline image

and its most manipulated counterpart and filter out any synthetic identities that show very little manipulation (i.e., RMS Difference ≤ 13). The first author manually filtered any synthetic identities with unrealistic or unclear properties, as in anything that made the headshot seem not real (e.g., image artifacts like shiny splotches or fluorescent patches, unnatural distortions or unrealistic objects in the headshot background, severely asymmetric accessories such as eyeglass frames). Finally, we randomly sample to balance by predicted gender and race. This leaves us with our final set of twenty baseline stimuli, balanced equally by the interaction of predicted gender (man, woman) and predicted race (White, Black). We display the full set of stimuli in Figure S9, with the strength of the manipulation (α) on the x-axis and the random seed associated with each synthetic identity on the y-axis (bolded if used in field experiment). This general process and the same exact full set of stimuli were pre-registered using AsPredicted (#119258). In Web Appendix D.2, we employ human raters to verify the predicted demographics as well as other aspects of our stimuli.



(a) Initial

(b) After Filtering and Balancing

Figure S8: Histogram of deepFace Predicted Demographics

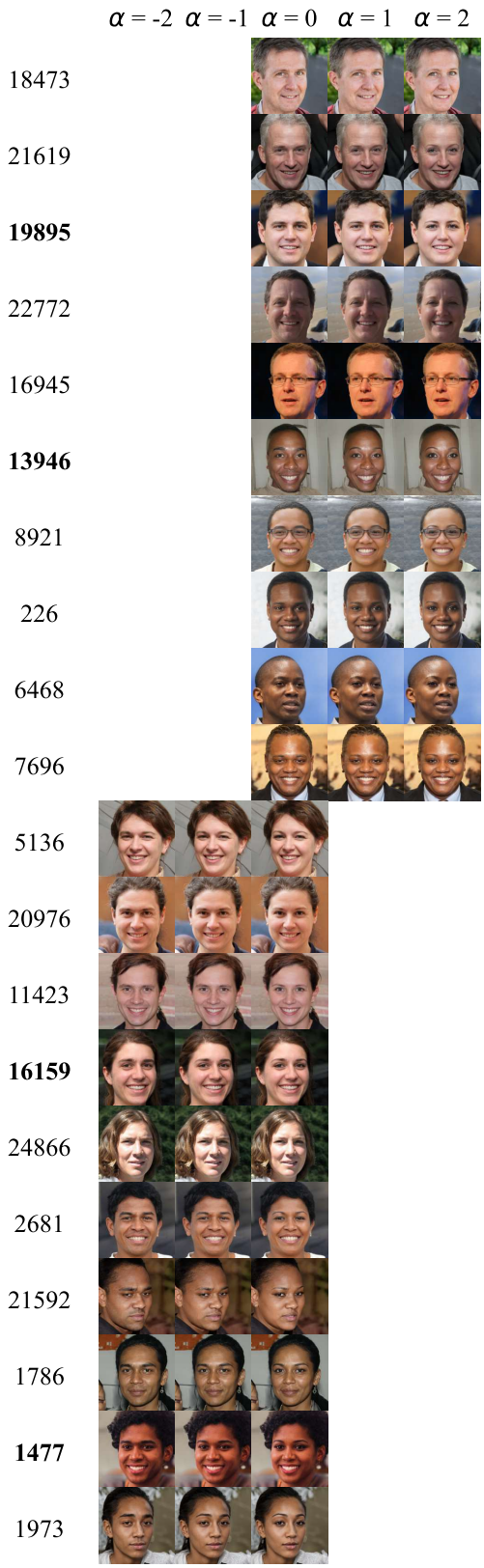


Figure S9: Full Set of Stimuli (row bolded if used in field experiment)

D.2 Validating stimuli with human raters

The following studies in this subsection are IRB compliant (Protocol Number: AAAT4855).

D.2.1 Measuring femininity, validating predicted race/realism

This study was pre-registered using AsPredicted (#119258). We recruited a sample of participants from Amazon Mechanical Turk (MTurk), who were required to be from the United States; have greater than a 95% approval rate for Human Intelligence Tasks (HITs); have more than 50 approved HITs; and have not completed another study from this research project (e.g., pilot studies). Through Qualtrics, we filter out people who do not live in the United States, are using a mobile phone, do not consent to the study, or are not at least 18 years old. Before the analyses, in the following order, we exclude participants if they are missing an MTurk ID, were just previewing the survey, were not a unique response by MTurk ID (only keep last response), were not a unique response by IP address (only keep last response), did not finish the survey, did not submit the correct survey completion code, were likely to be a bot (based on reCaptcha v3), failed either of two attention checks presented before the main study, failed a final attention check screener at the end of the study, or if the time they spent taking the survey was outside 2 standard deviations of the median. As in the main text, we are then left with our final sample ($N = 317$; $M_{\text{age}} = 34.8$, $SD_{\text{age}} = 10.3$; 175 men, 142 women). The exclusion criteria fully align with our pre-registration.

This study was a within-subject design, in which each participant saw all sixty headshots in counterbalanced order. Participants were first shown a headshot and then asked to rate its femininity: "On a continuous scale, **please rate the above face on appearance.**" (with anchors at -1: *extremely masculine*, 0: *androgynous*,[3] 1: *extremely feminine*; anchored at the midpoint).[4] They were then asked the following multiple choice questions to assess the stimulus' perceived race: "**What race do you think the above person is?**" (Black or African American; White; Other), allowing for multiple answers; and to assess whether the stimulus was perceived to be a fake person (in an indirect way to reduce potential demand characteristics): "**Have you seen this person before on TV?**" (Yes; No; This is not a real person).

[3] Participants were provided beforehand with the Oxford Languages definition of androgynous: "Note that an ***androgynous*** face is partly male and partly female in appearance (i.e., of indeterminate sex)."

[4] This scale was a slider scale which appeared continuous to the respondent but was really a 2001-point scale.

We found fair inter-rater reliability[5] for our measure of femininity ($\alpha_{\text{interval}} = .38$) and substantial inter-rater reliability for perceived race ($\alpha_{\text{nominal}} = .65$). For the synthetic identities predicted to be Black (see Web Appendix D.1), for every stimulus ($N = 30$), a majority of participants perceived them as "Black or African American." On average, 79.4% of the ratings for the relevant stimuli indicated "Black or African American," and 1.5% indicated "Black or African American" with other options selected as well. For the synthetic identities predicted to be White, for every stimulus ($N = 30$), a majority of participants perceived them as "White." On average, 97.3% of the ratings for the relevant stimuli indicated "White," and 0.1% indicated "White" with other options selected as well. Despite the heavy-handed (and repeatedly asked) question about realisticness, on average, only 2.3% of ratings indicated "This is not a real person." For every stimulus, the proportion of participants who selected "This is not a real person" never exceeded 6.5%, even though participants saw many similar stimuli from the same synthetic identities.

D.2.2 Validating baseline predicted gender

We recruited a sample of participants from MTurk in two waves. In the first wave, we required MTurkers to have not completed another study from this research. In the second wave, we required MTurkers to be from the United States; have greater than a 98% approval rate for HITs; have more than 10,000 approved HITs; and have not completed another study from this research. Through Qualtrics, in both waves, we filter out people who do not live in the United States, are using a mobile phone, do not consent to the study, or are not at least 18 years old. The remaining exclusion criteria are exactly as in Web Appendix D.2.1. We are then left with our final sample ($N = 182$; $M_{\text{age}} = 38.3$, $SD_{\text{age}} = 11.4$; 124 men, 57 women, 1 other).

This study was a within-subject design, in which each participant saw all sixty headshots in counterbalanced order. Participants were first shown a headshot and then asked to assess the stimulus' perceived gender identity: "Based on your impression, **what gender does the above person identify with?**" (Man; Woman; Other (e.g., non-binary); Not Sure).

We found moderate inter-rater reliability for perceived gender identity ($\alpha_{\text{nominal}} = .48$). For

[5] We measure inter-rater reliability using Krippendorff's alpha (Hayes and Krippendorff 2007), a flexible measure that handles missing data as well as binary, nominal, ordinal, interval, and ratio scales; adjusts to small sample sizes; and can be compared across any number of coders, values, and sample sizes. It generalizes Fleiss' κ, Spearman's rank correlation coefficient ρ, and Pearson's intraclass correlation coefficient r_{ii}, among others.

the baseline stimuli predicted to be men (see Web Appendix D.1), for every stimulus ($N = 10$), a majority of participants perceived them as identifying as a "Man." On average, 85.5% of the ratings for the relevant stimuli indicated "Man." For the baseline stimuli predicted to be women, for every stimulus ($N = 10$), a majority of participants perceived them as identifying as a "Woman." On average, 95.1% of the ratings for the relevant stimuli indicated "Woman."

We observed that manipulating femininity also systematically and markedly changed perceptions of gender identity for both baseline perceived men and women. Manipulating the baseline perceived male (female) stimuli by one and then two standard deviations changed the average proportion of ratings indicating "Man" ("Woman") from 85.5% to 50.5% to 15.4% (95.1% to 58.3% to 21.7%). This observation motivated the construction of our novel experimental design (see Section 4.2), which explicitly discloses the gender identity of each synthetic person, as proxied via gender pronouns. See Web Appendix D.2.5 for further motivation.

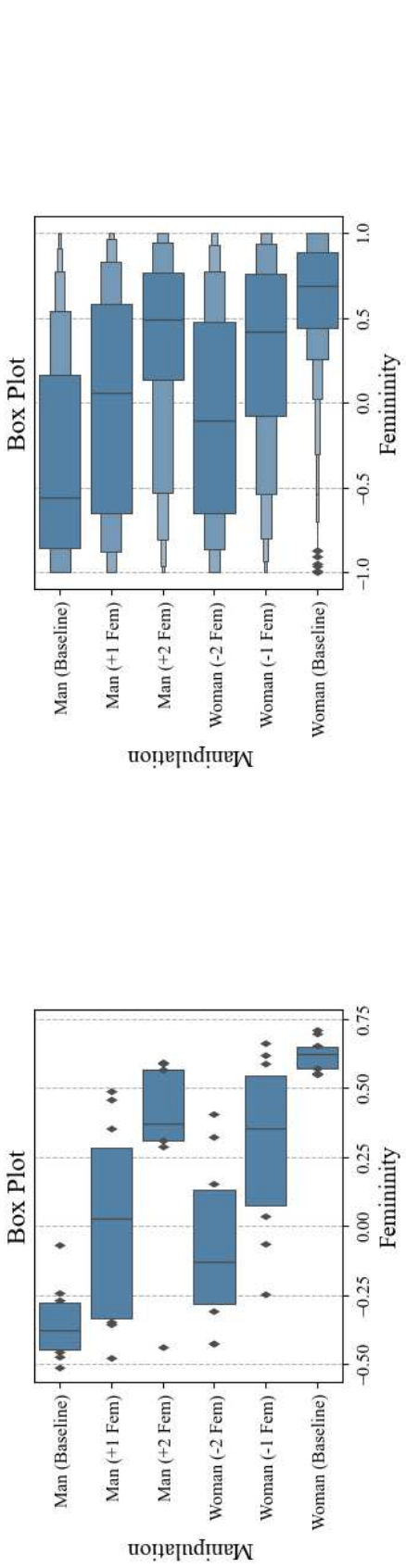
D.2.3 Manipulation checks for femininity

This section analyzes data from Web Appendix D.2.1's pre-registered study ($N = 317$)[6] to assess whether StyleGAN2 successfully manipulated femininity. Plotting perceptions of femininity by manipulation category (with perceived baseline gender, as validated in Web Appendix D.2.2) provides great face validity, aligning as expected directionally with our intended manipulations (see Figure S10).

Next, we inspect the percentage of femininity comparisons over stimuli that align with our intended manipulations. For example, an aligning comparison at the participant and stimuli level would be a single MTurker who rates a *Man (Baseline)* stimuli as having less femininity than its corresponding *Man (+1 SD femininity)* stimuli. We find a substantial degree of alignment at each of three different different ways of mean aggregation: (1) by participant and stimuli (i.e., individual ratings): 13,380/19,020 (70.3%), (2) by participant and manipulation category: 1,673/1,902 (88.0%), and (3) by stimuli and manipulation category: 60/60 (100%).

Finally, to assess statistical significance, we conduct paired samples one-sided t-tests. In short, the differences in means all align directionally with our intended manipulations and are all statistically significant at the $\alpha = 0.1\%$ level. For synthetic identities with baseline perceived men, the baseline stimuli ($M = -.35,\ SD = 0.61$) had significantly less femininity than the versions manipulated by $+1\ SD$ ($M = -.00,\ SD = 0.66$), $t(3169) = -30.47$, $p < .001$, which themselves had significantly less femininity than the versions manipulated by $+2\ SD$ ($M = .35,\ SD = 0.56$), $t(3169) = -31.29$, $p < .001$. For synthetic identities with baseline perceived women, the baseline stimuli ($M = .62,\ SD = 0.35$) had significantly more femininity than the versions manipulated by $-1\ SD$ ($M = .29,\ SD = 0.57$), $t(3169) = 33.78$, $p < .001$, which themselves had significantly more femininity than the versions manipulated by $-2\ SD$ ($M = -.07,\ SD = 0.63$), $t(3169) = 31.09$, $p < .001$.

[6] The analyses in this section (apart from the t-tests) were pre-registered using the same AsPredicted (#119258) as before.



(a) Averaging over Stimuli ($N = 60$)

(b) Individual MTurk Ratings ($N = 19,020$)

Figure S10: Fem(ininity) Perceptions by Manipulation Category

D.2.4 Variation in femininity perceptions for field experiment

Figure S11 plots variation in individual MTurk-level femininity perceptions from Web Appendix D.2.1's study by femininity treatment levels (according to the stimuli generation process), for each synthetic identity selected for the field experiment. Plots display kernel density estimation plots overlaid on histograms. All consecutive comparisons of means of femininity treatment levels (i.e., Low vs. Medium; Medium vs. High) within synthetic identities are statistically significant and align directionally with our intended manipulations at the $\alpha = 0.1\%$ level (according to paired samples one-sided t-tests).

To note, the underlying treatment of interest is how feminine a consumer in the field experiment actually perceives the tutor to be. This underlying continuous construct is measured with some error, since we proxy by an average of MTurk ratings from Web Appendix D.2.1's study to circumvent demand characteristics and unnaturalness/low ecological validity that would arise from asking the consumers in the field experiment to directly provide their femininity perceptions. Thus, discretizing based on the femininity score will help reduce measurement error, since the resulting discrete categories are much more likely to align between our proxy and true unobserved femininity perceptions. Reducing such measurement error may be especially relevant given the substantial overlap in distributions within each synthetic identity, suggesting that consumers may vary in their subjective femininity perceptions. Indeed, the inter-rater reliability for the continuous measure of femininity from Web Appendix D.2.1's study is fair but not perfect ($\alpha_{\text{interval}} = .38$).

In addition, the intended femininity manipulation is not homogeneous across synthetic identities (e.g., for seed 19895, the low and medium femininity treatment levels are much closer in femininity score than those for seed 13946). As a result, femininity treatment levels do not perfectly correspond to femininity tertiles (see Figure 3). We opt to lead with results for discretized tertiles, because each tertile has an absolute meaning (in terms of femininity) that can be compared across treatment conditions. For instance, for seed 19895 (corresponding to the top left headshots in Figure 3), the low and medium femininity treatment levels both bucket into the low femininity tertile—thus, in the tertile regressions, those two headshots are sensibly grouped together when doing comparisons, as they have similar femininity in *absolute* terms. For that same seed, the high femininity treatment level buckets into the medium femininity tertile, such that this synthetic identity has no high

femininity tertile—thus, in the tertile regressions, this synthetic identity sensibly does not contribute to high femininity treatment effects, as no headshot for this seed has a high femininity in *absolute* terms. In any case, results are largely consistent when operationalizing femininity via treatment levels (see Web Appendix F.5).

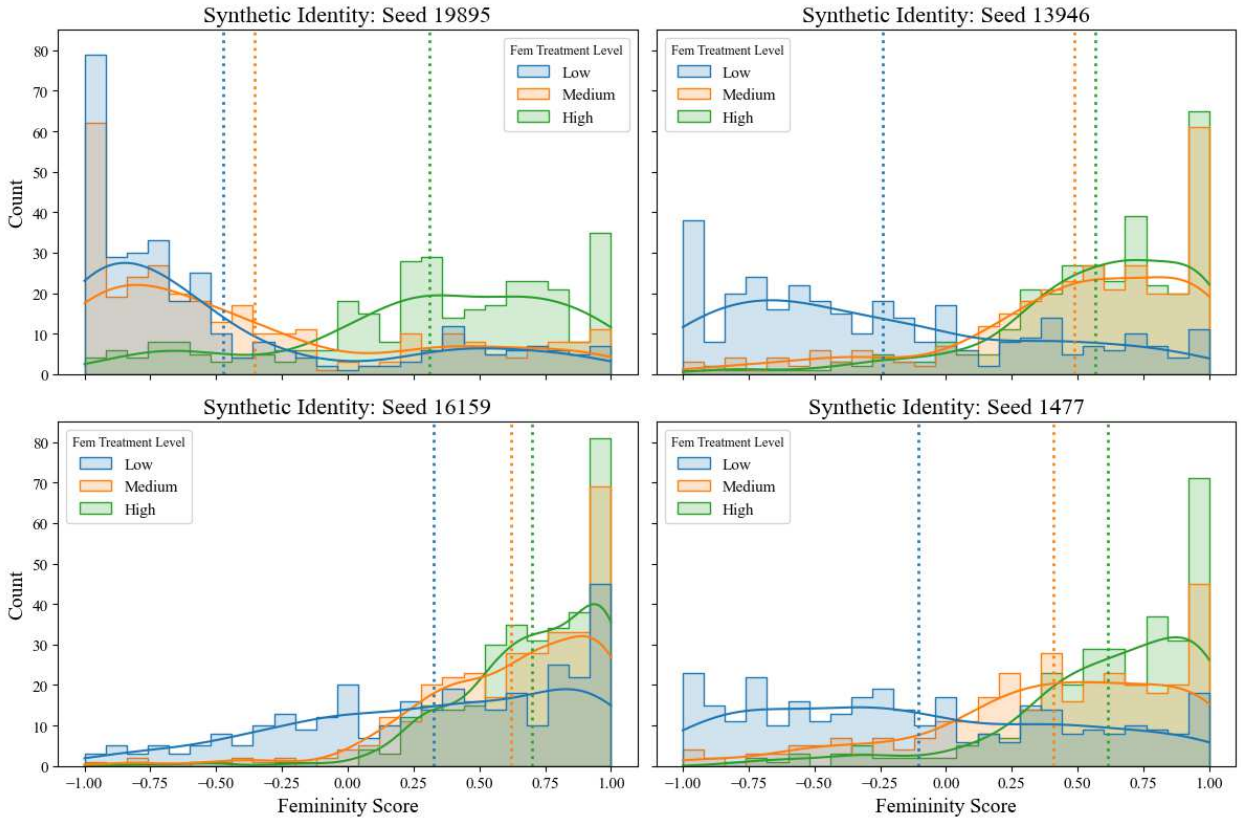
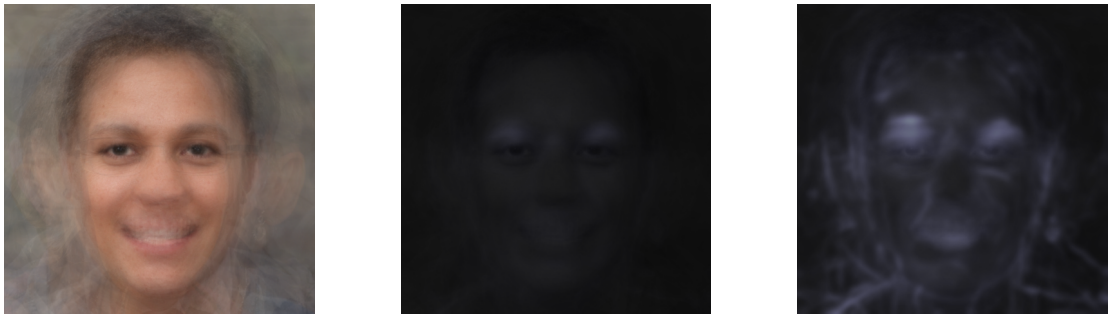


Figure S11: Fem(ininity) Perceptions by Treatment Level for Synthetic Identities Selected for Field Experiment

D.2.5 Assessing latent confounding

Heat Maps. We take pixel-wise absolute differences between baseline stimuli (i.e., $\alpha = 0$) and their most manipulated counterparts (i.e., $\alpha = -2$ or 2), with stimuli being converted to grayscale. To aggregate across these twenty differences, we both average and take the maximum. After applying a Gaussian blur,[7] we plot these aggregated differences in heat maps (see Figure S12), next to the average headshot across all sixty stimuli.

From the average heat map, we first observe that any changes are overall subtle and, as desired, manifest on localized headshot features rather than other attributes (e.g., pose, background, and hairstyle). femininity is operationalized through the eyebrows as well as subtler features like the shape of the mouth, eyes, ears, and jawline. In case this process averages out problematic latent confounds (e.g., in the background), we plot the maximum heatmap as a stronger test, which shows similar results, with some stimuli also minutely changing the shape of shoulders. To note, the average heat map more accurately depicts our treatment variable for femininity, since our regressions primarily estimate *average* treatment effects.



(a) Average Headshot (b) Averaged Differences (c) Maximum of Differences

Figure S12: Absolute Pixel-wise Differences between Extreme Femininity Manipulations

Notes: Heat maps in (b) and (c) are superimposed over the grayscaled average headshot. This process involves linearly blending the two images, with the heat maps receiving 85% of the weight.

[7] Gaussian blur applies Gaussian smoothing to the input, a standard process for visualization purposes. We use a function from the OpenCV Python package: `cv2.GaussianBlur()`. Kernel standard deviations were set to 10 along the X- and Y-axes. Kernel size was set automatically based on the kernel standard deviations.

Human Raters (Closed-Ended). This study was pre-registered using AsPredicted (#162310). We recruited a sample of participants from Prolific and used the same recruitment and exclusion criteria as in Web Appendix D.2.1, except participants were now required to have at least 95% approval rate and 100 previous submissions on Prolific.[8] We are then left with our final sample ($N = 114$; $M_{\text{age}} = 43.0$, $SD_{\text{age}} = 14.2$; 77 women, 37 men).

Participants were asked to view twenty images, each of which is the three generated headshots for a synthetic identity displayed side by side (as in Figure S9). This study was a within-subject design, in which each participant saw all twenty images in counterbalanced order. Participants were first asked: "**In how many of the above three faces is the person...**" This question's type was a matrix table with columns (All of them; Only some of them; None of them) and rows (smiling?; frowning?; happy?; sad?; afraid?; disgusted?; angry?; surprised?). Rows were presented in counterbalanced order. Then, they were asked to: "**Select all items in which you see any differences across the above (same) three people. If there are none, please select *None*.**" (race; photo background; pose; hairstyle; clothes; accessories (e.g., glasses); eyebrows; mouth; eyes; ears; jawline; None). Multiple answers were allowed, and "None" was an exclusive answer. In an open-ended question, they were finally asked to: "**Please describe any other differences (other than the items above) you notice here. If there are none, leave this blank.**" To note, this design is a particularly adversarial test of latent confounding, since unlike in the main studies, participants are shown *all* three versions of a single synthetic identity right next to each other, for *all* synthetic identities.

In Table S1, we document the proportion of noted differences in headshot features by individual ratings as well as majority aggregation by image, as pre-registered. A rating for the first matrix table question is 1 if "Only some of them" is selected for a particular image and particular expression; and otherwise 0. As expected, participants most consistently noted changes in eyebrows, but also in eyes, mouth, and jawline—all features that compose femininity in our stimuli, as determined by the heat maps. Participants did not, overall, note changes in ears, since those adjustments are visually subtle. All other attributes considered were noted in less than 10% of the ratings. In the open-ended question, most responses noted changes in "gender" (thus again motivating the use of

[8] We did *not* implement our last pre-registered exclusion step, since none of our participants gave free responses that seemed to be generated with a large language model (Veselovsky et al. 2023).

gender pronouns in the experimental design) and "masculinity, femininity" (thus again verifying our manipulation; see Web Appendix D.2.3 for more direct manipulation checks). Some free responses noted features other than those included in the questions before (e.g., rounded face/size, eyelashes, cheeks, cheekbones, forehead, nose, complexion, facial hair, teeth), which we investigate further in the following study. In summary, we find our manipulations systematically and markedly changed several features that compose femininity.

Table S1: Proportion of Noted Differences in Selected Headshot Features in Manipulated Stimuli

Attribute	Ratings ($N = 2,280$)	By Image ($N = 20$)
smiling?	101 (4.4%)	0
frowning?	101 (4.4%)	0
happy?	152 (6.7%)	0
sad?	94 (4.1%)	0
afraid?	67 (2.9%)	0
disgusted?	103 (4.5%)	0
angry?	78 (3.4%)	0
surprised?	119 (5.2%)	0
race	36 (1.6%)	0
photo background	22 (1.0%)	0
pose	39 (1.7%)	0
hairstyle	115 (5.0%)	0
clothes	19 (0.8%)	0
accessories	225 (9.9%)	0
eyebrows	1751 (76.8%)***	18 (90%)
mouth	897 (39.3%)***	4 (20%)
eyes	1189 (52.2%)***	13 (65%)
ears	156 (6.8%)	0
jawline	709 (31.1%)***	0
None	260 (11.4%)	0

Notes: The *Ratings* column includes significance stars for one-sided, binomial tests ($H_0 : \pi \leq .2$) to assess the likelihood the proportions constitute a sizeable proportion, as a conservative test. As pre-registered, proportions for *By Image* are aggregated by majority vote. As pre-registered, we also conduct one-sided, binomial tests ($H_0 : \pi \leq .5$), for which eyebrows (at the $\alpha = 0.1\%$ level) and eyes (at the $\alpha = 5\%$ level) were significant but none else (at the $\alpha = 10\%$ level).
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Human Raters (Open-Ended). This study was pre-registered using AsPredicted (#220148). We recruited a sample of participants from Prolific and used the same recruitment and exclusion criteria as in Web Appendix D.2.1, except participants were now required to have at least 95% approval rate and 100 previous submissions on Prolific. We are then left with our final sample ($N = 72$; $M_{\text{age}} = 41.5$, $SD_{\text{age}} = 12.7$; 52 women, 20 men).

Participants were asked to view four images, each of which is the three generated headshots for the synthetic identities selected for the field experiment displayed side by side (as in Figure S9). This study was a within-subject design, in which each participant saw all four images in counterbalanced order. Participants were asked: "Please ***list any differences*** you see across the above three photos." They were provided with ten boxes and asked to concisely write only one difference per box, without needing to fill out every box. Across all four images, participants wrote an average of 14.9 textual differences (or 3.7 differences per synthetic identity), with a standard deviation of 5.2 differences. Only one participant exhausted all ten response options (and only for one synthetic identity), suggesting that participants were not constrained by the ten box limit.

We grouped together similar differences to produce a set of sixteen semantic categories using a pre-registered procedure (which we fully complied with) involving sentence embeddings and hierarchical clustering. Namely, we first clean the text by lowercasing and removing periods. Then, we convert each unique cleaned difference into a sentence embedding using the sentence transformer: GIST Embedding v0 (Solatorio 2024). We selected this sentence transformer, since it is currently the state-of-the-art on the Massive Text Embedding Benchmark (English v2; Enevoldsen et al. 2025) among light-weight models (i.e., memory usage under 500MB) and ranked #25/216 among all models. Then, we compute the pairwise cosine similarity matrix, $\mathbf{S}$, for all such differences. We convert $\mathbf{S}$ into a distance matrix by $1 - \mathbf{S}$. Next, we perform agglomerative hierarchical clustering with Ward's method using the distance matrix (such that the unit of observation is a unique textual difference). We determine the number of clusters using silhouette scores (i.e., mean silhouette coefficient over all samples), considering between 3 and 18 clusters as pre-registered. The model with the highest silhouette score (corresponding to 16 clusters; see Figure S13) was selected. We display the full dendrogram from this hierarchical clustering process in Figure S14.

We labeled each cluster using cluster prototypes. Prototypes are determined by cluster prototype

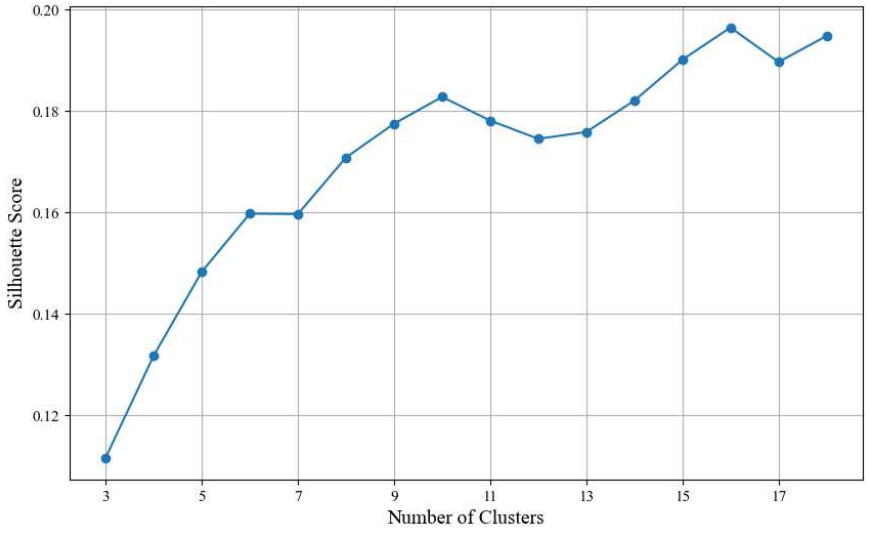


Figure S13: Selecting Number of Clusters with Silhouette Scores

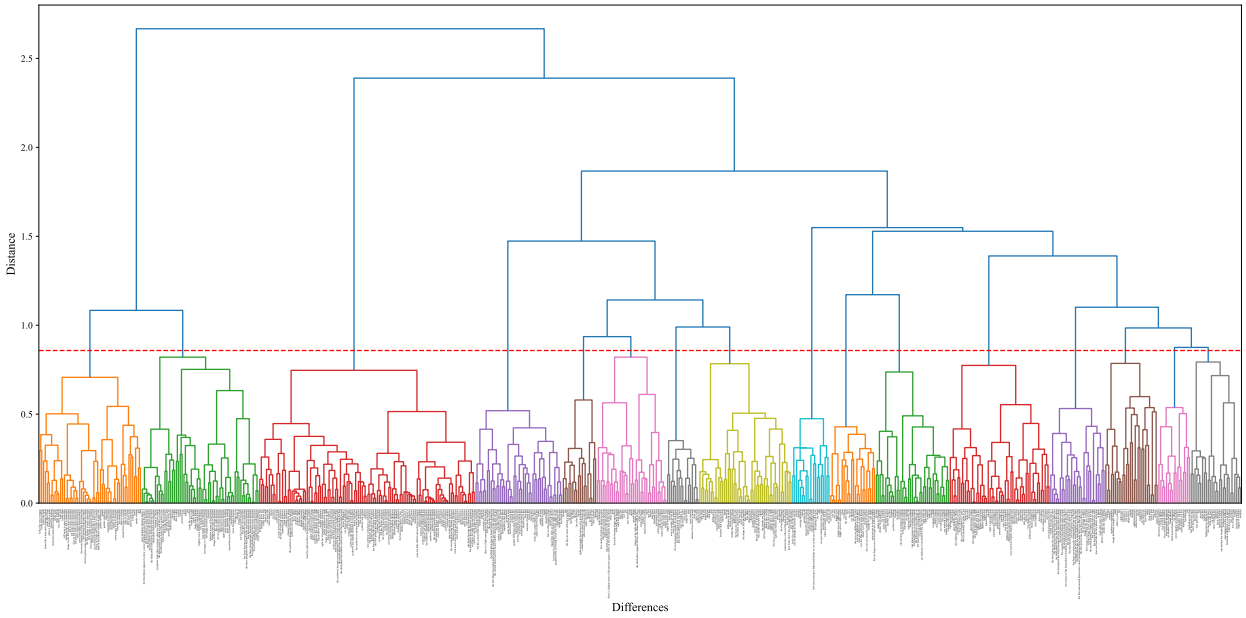


Figure S14: Dendrogram of Hierarchical Clustering of Textual Differences

scores, for which each textual difference's score is determined by: average cosine similarity with differences in the same cluster (excluding self) − average cosine similarity with differences in other clusters. As pre-registered, we label each cluster using an item that scored highly on the prototype score for that cluster and was concise and representative (see bolded items in Table S2). Finally, we mapped each textual difference to its corresponding cluster label.

Table S2: Top Ten Ranked Items by Prototype Score (In Parentheses) for Each Cluster

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
1	**masculine (0.160)**	masculine to feminine (0.153)	masculine / feminine (0.153)	manly to feminine (0.152)	first more masculine (0.148)	the first picture is more masculine (0.147)	first image is more masculine (0.146)	image on far left is masculine (0.145)	somewhat masculine in appearance (0.145)	feminine / masculine (0.145)
2	**feminine (0.155)**	the third picture is quite feminine (0.154)	the third picture looks quite feminine (0.152)	the third photo looks the most feminine (0.146)	feminine in one of the pictures (0.145)	last image is more feminine (0.142)	the third photo looks more feminine than the others (0.142)	more feminine (0.141)	the third photo appears to show a more feminine person than the other 2 (0.141)	image becomes more feminine (0.140)
3	**eyebrows thinner (0.193)**	thinner eyebrows (0.193)	eyebrows are thinner (0.192)	manicure thinner eyebrows (0.191)	thin eye brows (0.191)	thinner eyebrow (0.189)	really thin brows (0.188)	thick eyebrows (0.188)	thickness of eyebrows (0.187)	eyebrow thickness (0.187)
4	**facial hair (0.161)**	beard / stubble (0.160)	facial hair / no facial hair (0.159)	beard / mustache (0.155)	the first photo has beard and mustache stubble not seen in the other photos (0.147)	no facial hair (0.145)	facial hair in left picture (0.144)	facial hair in first photo (0.144)	signs of shaved facial hair (0.142)	beard (0.141)
5	**lip thickness (0.159)**	thickness of lips (0.152)	lip size (0.142)	lip color / fullness (0.136)	lip color (0.136)	lip fullness (0.133)	lips (0.130)	lip plumpness (0.130)	the lips are more defined w lipstick (0.121)	lip shade (0.118)
6	teeth and smile bigger (0.132)	how much of teeth are shown in smile (0.131)	teeth (0.121)	deep smile lines (0.106)	**smile (0.105)**	deeper smile lines (0.105)	more teeth (0.103)	teeth look bigger (0.102)	jaw line (0.102)	smile width (0.100)
7	**nose size (0.201)**	width of nose (0.193)	nose (0.190)	nose width (0.183)	shape of nose (0.178)	nose shape (0.177)	nose is slightly smaller (0.171)	nose structure (0.168)	the nose is wider (0.163)	nose is bigger in 1 (0.160)
8	chin shape (0.117)	**chin / jaw (0.112)**	chin (0.108)	the shape of face and jawline (0.106)	chin size (0.106)	defined chin (0.104)	face shape (0.100)	broader chin in left picture (0.100)	thicker chin (0.099)	shape of face (0.098)
9	**earring in ear (0.211)**	earring (0.201)	earrings (0.185)	ear piercing (0.181)	earring definition (0.179)	one ear lob has an earring (0.163)	ear ring in ear (0.159)	right most has an earring (0.147)	earing in right picture (0.135)	ears (0.133)
10	**size of eyes (0.158)**	eyes (0.158)	eye size (0.155)	eye sizes (0.152)	the eyes (0.150)	the eyes again (0.136)	eye shape (0.136)	bigger eyes (0.135)	bigger eyes in right image (0.132)	bigger eyes in right picture (0.129)
11	**eye makeup / mascara (0.172)**	eyeliner / eye makeup (0.169)	eye makeup (0.169)	eye makeup is applied (0.160)	makeup (0.148)	eyeliner (0.145)	eye liner (0.142)	more eye makeup (0.138)	woman with makeup (0.136)	make up (0.134)
12	smoother skin complexion (0.167)	**smoother skin (0.163)**	skin smoothness (0.156)	smooth skin complexion (0.154)	skin clarity (0.152)	clearer skin (0.151)	smooth skin (0.151)	smoothness of skin (0.150)	clarity of skin (0.146)	very smooth skin (0.145)
13	**the first picture has no edits (0.121)**	the first picture is without any edits (0.118)	does not look like the first photo. ai picture (0.118)	the second picture is lighter than the first (0.114)	the third photo has less face lines than the others (0.104)	photo looks like it been changed (0.102)	the first picture has more lines in their face (0.102)	first image is less filled out less rounded (0.101)	unlike the other photos (0.101)	the faces are more in the foreground on one of the pic (0.101)
14	younger 2 (0.064)	**older to younger (0.063)**	younger (0.057)	older age 1 (0.056)	older (0.052)	youngest 3 (0.050)	age (0.041)	third softer (0.027)	mid focus 2 (0.027)	completion (0.024)
15	**five o'clock shadow (0.110)**	5 oclock shadow (0.102)	afternoon shadow lightens (0.099)	more shadows 1st box (0.094)	faint beard shadow disappears (0.092)	eye shadow (0.091)	beard shadow (0.090)	five o'clock shadow / stubble (0.085)	blemish darkness (0.076)	eye shadow 3rd box (0.072)
16	**more sideburn hair** (0.076)	forehead wrinkles (0.065)	hair (0.061)	sideburns (0.057)	wrinkles (0.056)	thick sideburns (0.054)	thick side burns (0.053)	less wrinkles (0.053)	hair texture (0.050)	degree of wrinkles (0.049)

We display the proportion of textual differences that belong to each semantic category in Figure S15, the proportion of participants who mentioned a semantic category (across all their textual differences) in Figure S16 (as pre-registered), and a model-free word cloud of differences in Figure S17. These results suggest that many described differences relate to femininity (i.e., eyebrows, skin,[9] chin/jaw, eyes, nose, and lips). However, participants also note changes in other attributes (i.e., makeup, facial hair, earrings, smile, and age), which we view as potential latent confounders.

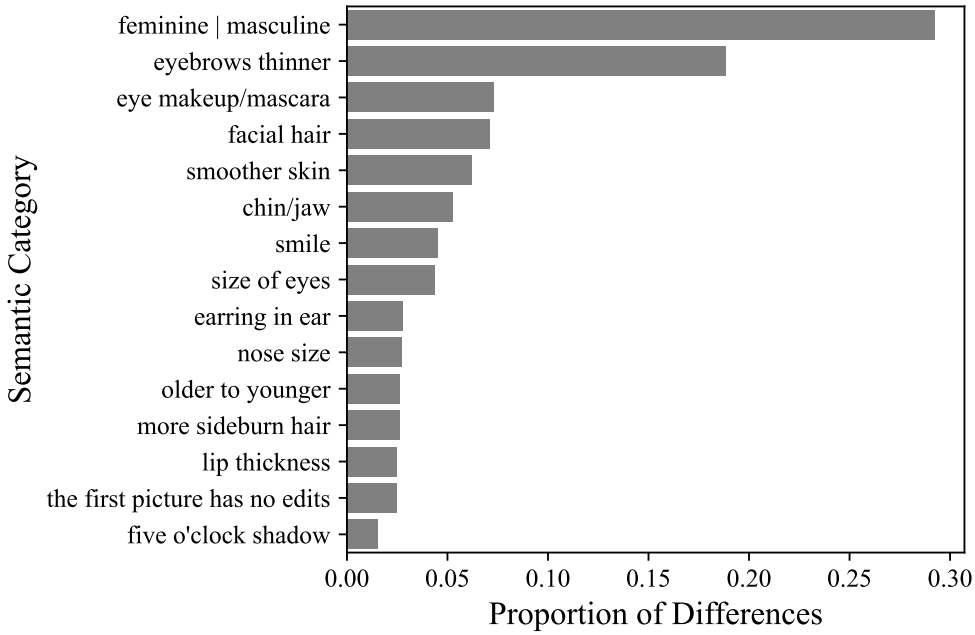


Figure S15: Proportion of Textual Differences by Semantic Category

Notes: "feminine" and "masculine" are two different semantic categories that we combined, since they both relate to femininity.

[9] Following the literature, we consider skin smoothness to be a feature relevant to femininity (Jaeger et al. 2018; Russell 2009), the perception of which may be influenced by makeup (Russell 2009), which we control for in Equation 1 as a potential confounder.

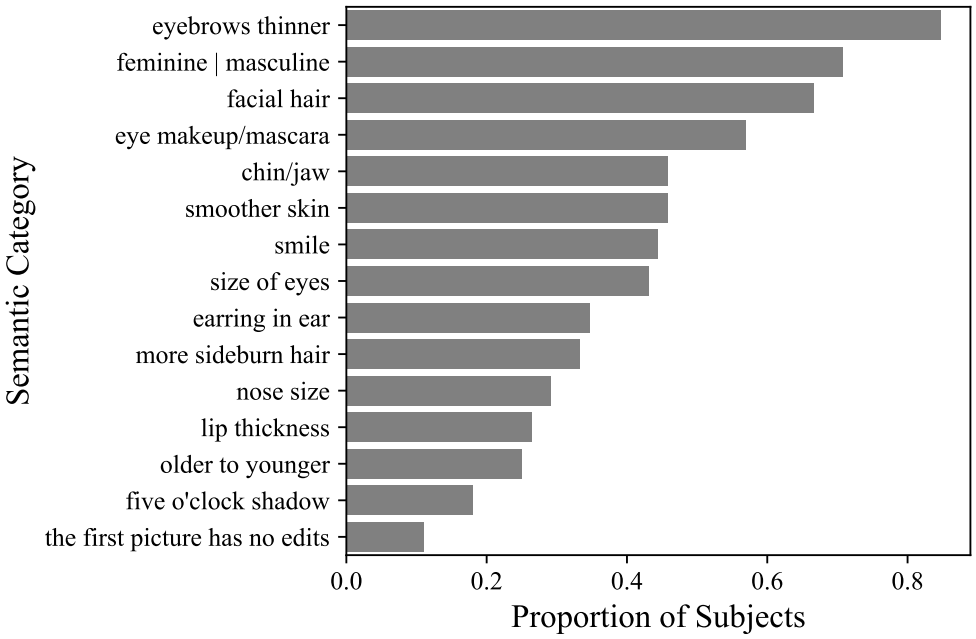


Figure S16: Proportion of Subjects Who Mention Each Semantic Category

Notes: "feminine" and "masculine" are two different semantic categories that we combined, since they both relate to femininity.

Figure S17: Word Cloud of Textual Differences

Notes: We lowercase textual differences, remove stop words, and then combine common consecutive pairs of words to form bigrams. Terms are sized proportional to their frequency. Color progressions are determined by frequency such that reds represent terms that appear more than 60% of times; oranges, between 30% and 60% of times; and blues, otherwise.

Measuring Latent Confounders. This study was pre-registered using AsPredicted (#221255). We recruited a sample of participants from Amazon Mechanical Turk and used the exact same recruitment and exclusion criteria as in Web Appendix D.2.1. We are then left with our final sample ($N = 308$; $M_{\text{age}} = 32.1$, $SD_{\text{age}} = 6.2$; 250 men, 58 women).

Participants were asked to view twelve headshots (i.e., the ones selected for the field experiment). This study was a within-subject design, in which each participant saw all twelve images in counterbalanced order. Participants were first shown a headshot and then asked the following questions, respectively:

1. "**On a continuous scale**, to what extent is **makeup** (e.g., mascara, lipstick) applied on this face?"
2. "**On a continuous scale**, to what extent are **earrings** present on their ears?"
3. "**On a continuous scale**, to what extent is **facial hair** (e.g., mustache, beard, sideburns) present on this face?"
4. "**On a continuous scale**, to what extent is this person **smiling**?"
5. "**On a continuous scale**, how **old** do you think this person looks?"

The first four questions had anchors at -1: *Not at all*, $-.5$: *Slight*, 0: *Moderate*, .5: *Considerable*, 1: *Extreme*. The last question on age had anchors at -1: *Very young*, $-.5$: *Somewhat young*, 0: *Middle-aged*, .5: *Somewhat old*, 1: *Very old*.[10]

We take the average of these ratings to extract a continuous score for each of the five latent confounders. For the twelve headshots, we compute pairwise Pearson correlations with each of these latent confounders' scores with femininity scores (computed in Web Appendix D.2.1), which are included as follows in parentheses: makeup (.69), earrings (.50), facial hair (-.51), smile (.75), and old (.10).

[10] All scales were slider scales which appeared continuous to respondents but were really 2001-point scales.

D.3 Comparing to text-to-image diffusion models

To compare our StyleGAN2-based approach with alternative deep generative image models, we anecdotally assess whether popular text-to-image diffusion models—i.e., Dall-E 3 (Betker et al. 2023), Stable Diffusion XL Base 1.0 (Podell et al. 2023), Midjourney, and GPT-4o (i.e., the version released in March 2025 with native image generation capabilities)—can generate realistic headshots and markedly manipulate femininity in generated headshot photos independently of other attributes, off-the-shelf.[11] In short, we find that no considered models can do so. In fact, many times they fail to produce photorealistic stimuli to begin with, which would induce unwanted uncanny valley or algorithmic aversion effects (De Freitas et al. 2023) if shown to participants. Overall, this qualitative exercise provides further evidence that StyleGAN models have become the "defacto golden standard for the editing of facial images" (Bermano et al. 2022).

To generate images, one for each of the combinations of baseline perceived gender (man, woman) and race (White, Black), we generally used the following base prompt:

> Generate three images side by side. On the left, generate a high-resolution, realistic, and natural headshot of a (white / black) (man / woman) that does NOT look like a cartoon or animation. Towards the right, generate versions that SIGNIFICANTLY and visibly (INCREASE / DECREASE) this person's facial femininity, while keeping all other aspects of the photo like race, pose, expression, background, hairstyle, hair color, clothes, and accessories exactly the same.

We display results for each model type below. Dall-E 3 was queried via ChatGPT-4 between February and April 2024. For Stable Diffusion XL Base 1.0, we used "text, bad anatomy, bad teeth, cartoon, animation" as negative prompts; set number of inference steps to 12; and experimented with guidance scale—for each combination, we generated around twenty versions and selected the best version to display. Midjourney was queried via Discord between March and April 2024. GPT-4o was queried via ChatGPT-4o in May 2025.

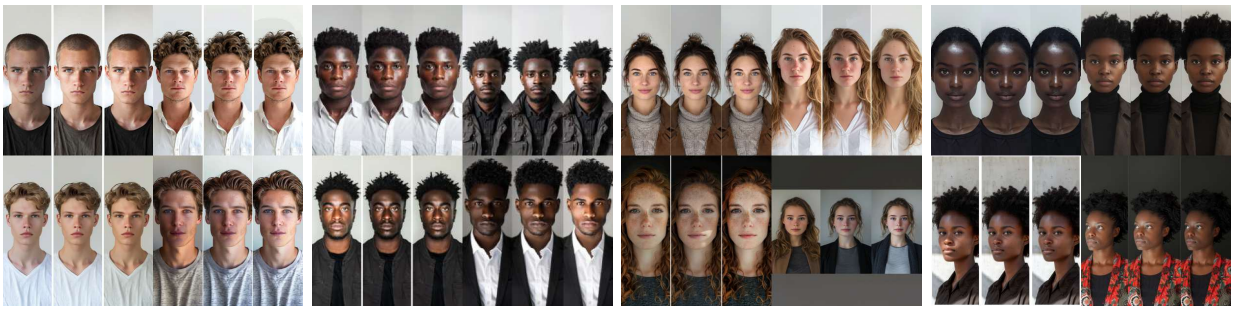
[11] For diffusion models, there has been some non off-the-shelf work in directly examining or encouraging disentanglement, with varying degrees of face validity (see Huang et al. 2024 for a review).

(a) White Man (b) Black Man (c) White Woman (d) Black Woman

Figure S18: Dall-E 3: Femininity Manipulations

(a) White Man (b) Black Man (c) White Woman (d) Black Woman

Figure S19: Stable Diffusion XL: Femininity Manipulations



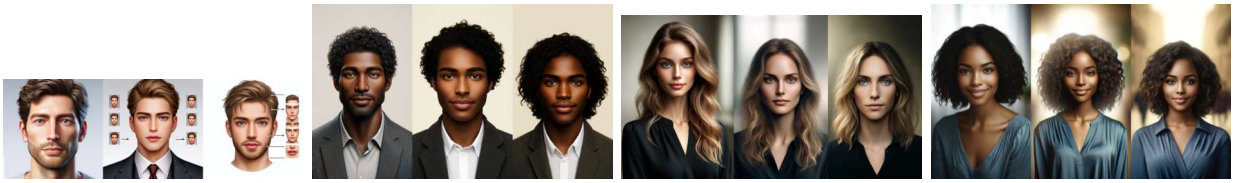
(a) White Man (b) Black Man (c) White Woman (d) Black Woman

Figure S20: Midjourney: Femininity Manipulations

(a) White Man (b) Black Man (c) White Woman (d) Black Woman

Figure S21: GPT-4o: Femininity Manipulations

We also try text guided image-to-image, by starting with a base image and then specifying to manipulate femininity using the same general prompting strategy as before:

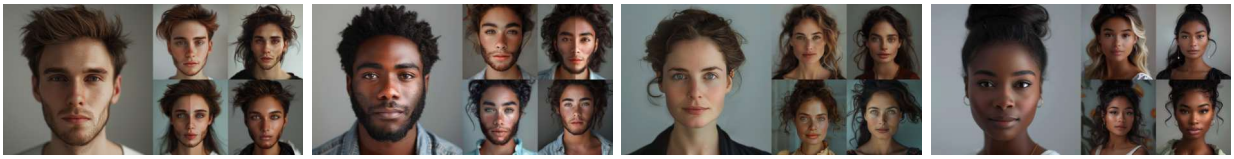


(a) White Man (b) Black Man (c) White Woman (d) Black Woman

Figure S22: Dall-E 3 Image-to-Image: Femininity Manipulations

(a) White Man (b) Black Man (c) White Woman (d) Black Woman

Figure S23: Stable Diffusion XL Image-to-Image: Femininity Manipulations



(a) White Man (b) Black Man (c) White Woman (d) Black Woman

Figure S24: Midjourney Image-to-Image: Femininity Manipulations

(a) White Man (b) Black Man (c) White Woman (d) Black Woman

Figure S25: GPT-4o Image-to-Image: Femininity Manipulations

Web Appendix E Lab Experiments for Psychological Perceptions

Evidence suggests that facial trustworthiness may affect decisions in peer-to-peer lending (Duarte et al. 2012), firm audits (Hsieh et al. 2020), crowdfunding (Duan et al. 2020), and sharing economies (Ert et al. 2016). Facial competence predicts voting behavior (Olivola and Todorov 2010) and the selection and compensation of CEOs (Rule and Ambady 2008; Graham et al. 2017). We use this literature to guide the selection of the two dependent variables of interest in our lab experiments: perceived trustworthiness and competence.

E.1 Measuring self-reported perceptions of trustworthiness and competence

This study was partially pre-registered[12] using AsPredicted (#119290). We recruited a sample of participants from MTurk and used the same recruitment and exclusion criteria as in Web Appendix D.2.1. We are then left with our final sample ($N = 1,818$; $M_{\text{age}} = 38.6$, $SD_{\text{age}} = 11.5$; 997 men, 807 women, 8 other, 6 prefer not to say). We also collected participant demographics on race/ethnicity (1,303 White, 184 Black or African American, 125 Asian, 91 Hispanic or Latino, 25 prefer not to say, 19 Hispanic or Latino and White) and political orientation (376 very conservative, 279 very liberal, 274 moderately liberal, 263 moderately conservative, 235 neither liberal nor conservative, 193 slightly liberal, 187 slightly conservative, 11 prefer not to say) to study heterogeneous treatment effects.

Participants were randomly assigned to the conditions of a 5 (headshot-pronoun pairing; fractional factorial) × 2 (DV: trustworthiness, competence) × 20 (synthetic identities) mixed design, with headshot-pronoun and DV as between-subject factors[13] and synthetic identities as a within-subject factor. In other words, each participant saw exactly one headshot-pronoun pairing (see Figure 1) for each of the twenty synthetic identities in counterbalanced order and rated them all on the same one dependent variable. In the intro, we mentioned: "Whenever possible, we will use people's preferred gender pronouns." We use this phrasing over a version without the first two

[12] To explain further, we had already collected data on perceptions of trustworthiness and competence but had yet to run the study in Web Appendix D.2.1 on perceptions of femininity, a key treatment variable. Therefore, we pre-registered our analysis plan (e.g., regression specifications, exclusion criteria) for complete transparency.

[13] We selected a between-subject design for headshot-pronoun to ensure participants do not see similar headshots from the same synthetic identity, which may inflate demand effects and cue participants that the headshots are AI-generated; and for DV to minimize potential spillover/correlation between our outcomes of interest.

words to add extra emphasis to participants that we are aware of people's preferred gender pronouns when we display them (rather than just guessing). Then, we asked: "Please rate the person below on (his / her / their) (**trustworthiness** / **competence**)." We displayed the headshot and then asked "How (**trustworthy** / **competent**) do you think (he is / she is / they are)?" on a continuous scale (with anchors at -1: *not at all*, 1: *extremely*; anchored at the midpoint).[14]

E.2 Regression specification and results

For participant i who sees headshot h (belonging to synthetic identity k) paired with pronoun p, we estimate the following OLS regression specification—with robust standard errors clustered by participant:

$$\text{Perception}_{ihp} = \boldsymbol{\alpha} \cdot \textbf{Fem(ininity) Tertile}_h + \boldsymbol{\beta} \cdot \textbf{Pronoun}_p + \boldsymbol{\theta} \cdot \textbf{Fem Tertile}_h \times \textbf{Pronoun}_p + \boldsymbol{\lambda}_i + \boldsymbol{\gamma}_k + \varepsilon_{ihp} \tag{2}$$

As shorthand, we will refer to He/His pronouns as He pronouns; She/Her, as She pronouns; and They/Their as They pronouns. Given reviewer feedback, we deviate from pre-registration by using nonparametric femininity tertiles (such that each tertile corresponds to twenty headshots; Low $\in [-.52, -.15]$, Medium $\in [-.11, .41]$, High $\in (.41, .72]$), as in Footnote 9, instead of assuming that femininity enters linearly; and by mean-centering covariates such that main average treatment effects can be interpreted in the same way when including covariate treatment interaction effects (as recommended by Lin 2013; Athey and Imbens 2017). We consider medium femininity and He pronouns as baselines.

We report regression results in Table S3. First, we consider Column 1. For He pronouns, relative to medium femininity, high femininity (i.e., High Fem Tertile main effect) significantly increased perceived trustworthiness, an effect which persisted for She ($\chi^2(1) = 65.50$, $p < .001$; High Fem Tertile main effect plus High Fem Tertile $\times$ She interaction term) and They pronouns ($\chi^2(1) = 24.03$, $p < .001$; High Fem Tertile main effect plus High Fem Tertile $\times$ They interaction term). For He pronouns, relative to medium femininity, low femininity (i.e., Low Fem Tertile main effect) significantly decreased perceived trustworthiness, an effect which persisted for She

[14] This scale was a slider scale which appeared continuous to the respondent but was really a 2001-point scale.

($\chi^2(1) = 13.15$, $p < .001$; Low Fem Tertile main effect plus Low Fem Tertile × She interaction term) and They pronouns ($\chi^2(1) = 9.67$, $p = .002$; Low Fem Tertile main effect plus Low Fem Tertile × They interaction term).

Next, we consider Column 3. For He pronouns, relative to medium femininity, high femininity (i.e., High Fem Tertile main effect) significantly increased perceived competence, an effect which persisted for She ($\chi^2(1) = 99.40$, $p < .001$; High Fem Tertile main effect plus High Fem Tertile × She interaction term) and They pronouns ($\chi^2(1) = 32.86$, $p < .001$; High Fem Tertile main effect plus High Fem Tertile × They interaction term). Relative to medium femininity, low femininity did not significantly decrease perceived trustworthiness for He pronouns (i.e., Low Fem Tertile main effect) or They pronouns ($\chi^2(1) = 0.20$, $p = .654$; Low Fem Tertile main effect plus Low Fem Tertile × They interaction term), but did for She pronouns ($\chi^2(1) = 9.64$, $p = .002$; Low Fem Tertile main effect plus Low Fem Tertile × She interaction term). Among the headshots with medium femininity, relative to He pronouns, She pronouns (i.e., She main effect) significantly increased perceived competence. Altogether, results from Columns 1 and 3 suggest that femininity increases perceptions of trustworthiness and competence, for all gender pronouns considered.

We also study a number of pre-registered heterogeneous treatment effects based on stimuli and rater characteristics using interaction terms and exploratory cluster-robust Wald tests (reported with χ^2-statistics). Overall, there appears to be little evidence of heterogeneous treatment effects. To note, the effect of low femininity relative to medium femininity significantly decreased perceived trustworthiness and competence for female raters (reflecting homophily; Columns 2 and 4, Low Fem Tertile × Female (Rater) interaction term) and for liberal raters (Columns 2 and 4, Low Fem Tertile × Conservativeness (Rater)[15] interaction term). For perceived competence, the positive effect of She pronouns relative to He pronouns (among the headshots with medium femininity) significantly attenuated for conservatives (Column 4, She × Conservativeness (Rater) interaction term).

[15] Political orientation (i.e., "Conservativeness (Rater)" in Table S3) is coded on a 7-point scale: -3 = Very Liberal, -2 = Moderately Liberal, -1 = Slightly Liberal, 0 = Neither Liberal nor Conservative, 1 = Slightly Conservative, 2 = Moderately Conservative, 3 = Very Conservative.

Table S3: Impact of Femininity and Gender Pronouns on Self-Reported Psychological Perceptions (OLS)

	Dependent variable:			
	Perceived Trustworthiness		Perceived Competence	
	(1)	(2)	(3)	(4)
Low Fem(ininity) Tertile	−0.047***	−0.049***	0.010	0.013
	(0.010)	(0.011)	(0.011)	(0.011)
High Fem Tertile	0.061***	0.062***	0.068***	0.068***
	(0.013)	(0.013)	(0.015)	(0.015)
She	0.012	0.012	0.027**	0.027**
	(0.010)	(0.010)	(0.010)	(0.010)
They	0.008	0.007	0.010	0.010
	(0.012)	(0.012)	(0.011)	(0.011)
Low Fem Tertile x She	−0.010	−0.014	−0.057**	−0.054**
	(0.017)	(0.019)	(0.018)	(0.019)
Low Fem Tertile x They	−0.011	−0.009	−0.017	−0.015
	(0.019)	(0.022)	(0.017)	(0.020)
High Fem Tertile x She	0.028	0.028	0.032	0.030
	(0.015)	(0.015)	(0.016)	(0.016)
High Fem Tertile x They	0.014	0.014	0.015	0.015
	(0.017)	(0.018)	(0.018)	(0.018)
Low Fem Tertile x Black (Stimuli)		0.001		−0.014
		(0.019)		(0.018)
High Fem Tertile x Black (Stimuli)		−0.005		0.008
		(0.018)		(0.016)
She x Black (Stimuli)		−0.015		0.017
		(0.016)		(0.014)
They x Black (Stimuli)		−0.003		0.012
		(0.017)		(0.015)
Low Fem Tertile x Female (Rater)		−0.054***		−0.070***
		(0.016)		(0.013)
Low Fem Tertile x Black (Rater)		0.043		−0.016
		(0.029)		(0.023)
Low Fem Tertile x Conservativeness (Rater)		0.028***		0.012***
		(0.004)		(0.003)
High Fem Tertile x Female (Rater)		0.008		−0.005
		(0.012)		(0.011)
High Fem Tertile x Black (Rater)		−0.021		−0.002
		(0.018)		(0.020)
High Fem Tertile x Conservativeness (Rater)		−0.0004		0.003
		(0.003)		(0.003)
She x Female (Rater)		0.014		0.011
		(0.013)		(0.012)
She x Black (Rater)		0.020		−0.026
		(0.023)		(0.018)
She x Conservativeness (Rater)		0.003		−0.006*
		(0.003)		(0.003)
They x Female (Rater)		0.004		−0.014
		(0.016)		(0.014)
They x Black (Rater)		0.033		−0.025
		(0.025)		(0.024)
They x Conservativeness (Rater)		−0.0002		−0.00000
		(0.004)		(0.003)
Participant FE	✓	✓	✓	✓
Synthetic Identity FE	✓	✓	✓	✓
Observations	18,480	18,379	17,880	17,790
R^2	0.462	0.467	0.460	0.462
Adjusted R^2	0.433	0.438	0.431	0.432

Notes: Robust standard errors clustered by participant are reported in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

E.3 Stimuli selection by psychological perceptions

We selected one synthetic identity within each baseline perceived gender and race category with relatively high average perceived trustworthiness (to increase the likelihood that 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), across the five experimental conditions within the same synthetic identity. To describe our selection process in more detail, we first only consider synthetic identities with low p-values (i.e., $p < .025$) from an F-test (not assuming equal variances) for the difference in means of perceived competence. Then, among those synthetic identities, we select the one within each baseline perceived gender and race category with the highest mean perceived trustworthiness. We display every statistic we used for this stimuli selection process in Table S4, for which "Seed" denotes the random seed associated with each synthetic identity (in the same order as in Figure S9), and rows are bolded if the corresponding synthetic identity was selected for the field experiment. The four exact synthetic identities selected by this process as well as the general process itself were pre-registered before we ran our field experiment (AsPredicted, #166345; see included Google Drive link, with all files last modified before the pre-registration date).

Table S4: Trustworthiness and Competence Statistics for Each Synthetic Identity

Baseline Perceived Gender x Race	Seed	Mean Trustworthiness	Difference in Means of Competence
White Man	18473	0.30	$F(4, 443.93) = 1.11$, $p = .349$
	21619	0.31	$F(4, 442.31) = 1.29$, $p = .275$
	19895	**0.30**	$\boldsymbol{F(4, 442.75) = 2.87, p = .023}$
	22772	0.34	$F(4, 441.05) = 0.98$, $p = .420$
	16945	0.16	$F(4, 442.06) = 0.52$, $p = .720$
Black Man	**13946**	**0.36**	$\boldsymbol{F(4, 441.47) = 3.34, p = .010}$
	8921	0.48	$F(4, 440.16) = 2.07$, $p = .084$
	226	0.44	$F(4, 442.5) = 2.70$, $p = .030$
	6468	0.17	$F(4, 443.72) = 1.01$, $p = .404$
	7696	0.29	$F(4, 443.41) = 1.77$, $p = .134$
White Woman	5136	0.45	$F(4, 440.22) = 3.24$, $p = .012$
	20976	0.35	$F(4, 442.44) = 4.15$, $p = .003$
	11423	0.39	$F(4, 437.68) = 5.26$, $p < .001$
	16159	**0.49**	$\boldsymbol{F(4, 437.7) = 6.10, p < .001}$
	24866	0.22	$F(4, 443.72) = 2.34$, $p = .054$
Black Woman	2681	0.36	$F(4, 442.25) = 9.38$, $p < .001$
	21592	-0.05	$F(4, 443.12) = 8.68$, $p < .001$
	1786	0.25	$F(4, 440.18) = 7.50$, $p < .001$
	1477	**0.43**	$\boldsymbol{F(4, 444.34) = 3.17, p = .014}$
	1973	0.20	$F(4, 442.35) = 2.39$, $p = .050$

Web Appendix F Field Experiment (Miscellaneous)

F.1 Setting up the tutoring website

To make the website as realistic as possible, we designed a logo, created a company email, and set up a social media (i.e., Facebook) account with over twenty followers and several posts (e.g., "Happy Thursday students. Our motto of the day is from Snoopy! 'Learn from yesterday, live for today, look to tomorrow, rest this afternoon.'"). The Facebook page's intro read: "New affordable online tutoring service by Ivy League graduates. Founded by a Columbia Business School student." We designed ads for recruitment (see Figure S26). Upon clicking the ad, participants were linked to a website (see Figure 2), where they were immediately exposed to treatment (i.e., a headshot-pronoun pairing), leading to full compliance. In particular, the headshot and at least three mentions of the gender pronoun were immediately visible upon clicking. SAT/GRE scores and years of teaching experience were set to be identical within a synthetic identity (but varied only slightly across identities).

Figure S26: Sample Recruitment Ad

Other pages on the website are displayed in Figures S27 to S31.

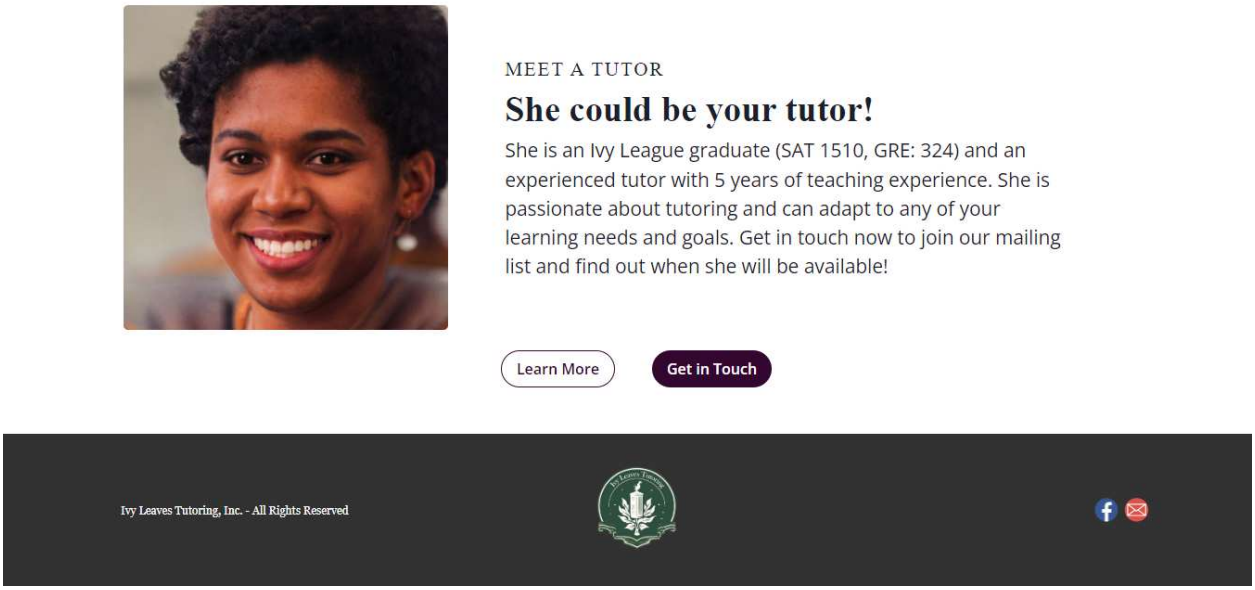


Figure S27: Bottom of Home Page

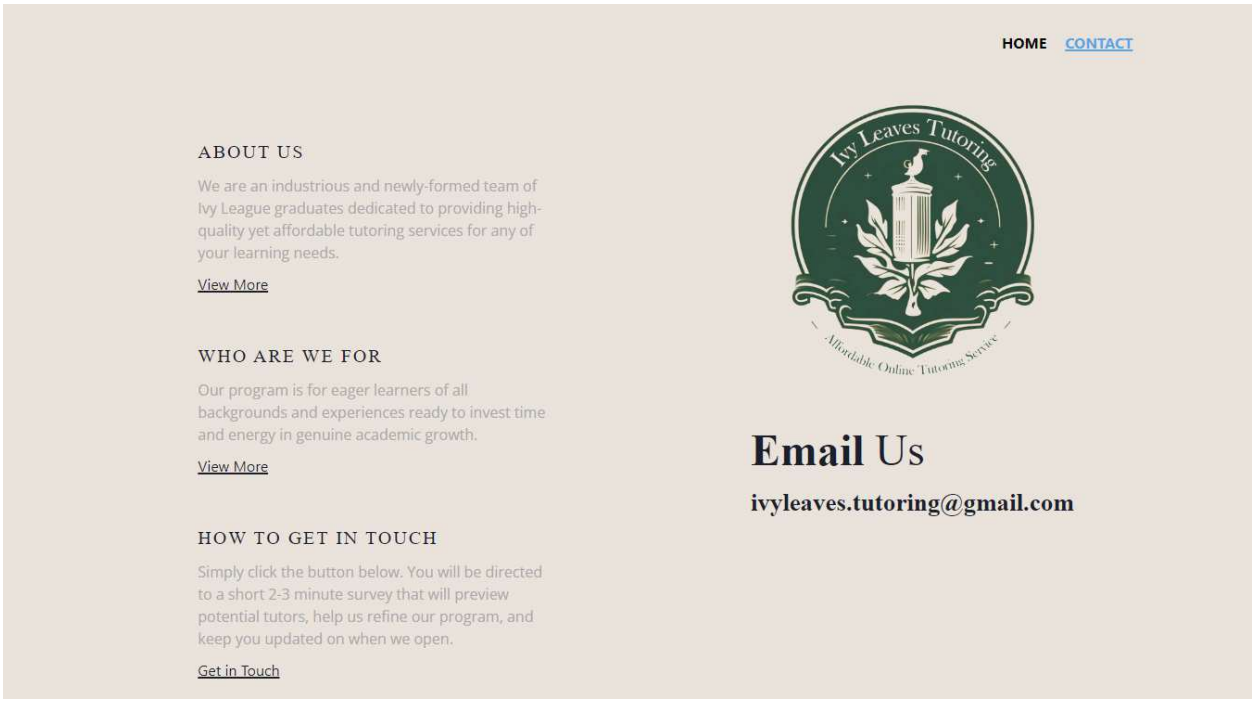


Figure S28: Contact Page—Directed here after clicking CONTACT tab in the top right

Figure S29: Learn More Page (Top)—Directed here after clicking Learn More

Our mission

At Ivy Leaves Tutoring, we're grounded in a simple yet powerful belief: every student deserves access to a high-quality education. To this end, we're dedicated to bringing unparalleled educational experiences online, making tutoring from Ivy League graduates available to all, from big cities to smaller towns. Through our platform, we ensure each student receives not only information, but also a personalized roadmap to academic success, irrespective of their background. Our vision is a world where physical boundaries vanish, and every student, regardless of their circumstances, benefits from the very best in educational guidance and support.

Get In Touch

Our tutoring services open soon! Simply click the button below and you will be directed to a short 2-3 minute survey that will preview potential tutors, help us refine our program, and keep you updated on when we open.

Get in Touch

MORE ABOUT US

Experienced Ivy League tutors with ***5+ years of experience***

This program was founded by a Columbia Business School student dedicated to providing accessible, high-quality, and affordable tutoring to students, regardless of their background. Our founder brought together a team of Ivy League graduates with significant teaching experience encompassing a diverse collection of subjects such as writing, math, statistics, history, biology, physics, chemistry, computer science, foreign languages (including Spanish, French, German, Latin, Japanese, Chinese), and standardized test-taking. We are ready to address and adapt to your learning needs and goals, no matter what they may be!

AVAILABLE 7 DAYS A WEEK

Our **program** is for...

1. **Eager learners** of all backgrounds and experiences ready to invest time and energy in genuine academic growth.
2. **Students at any stage**, from primary to university level.
3. **Applicants** to college or graduate programs who are seeking an edge.
4. **Professionals** seeking to upskill or reskill in a rapidly changing world
5. **Lifelong learners** looking to refresh or expand their knowledge; or perhaps even to streamline and improve their study habits.

Figure S30: Learn More Page (Bottom)—Directed here after clicking Learn More

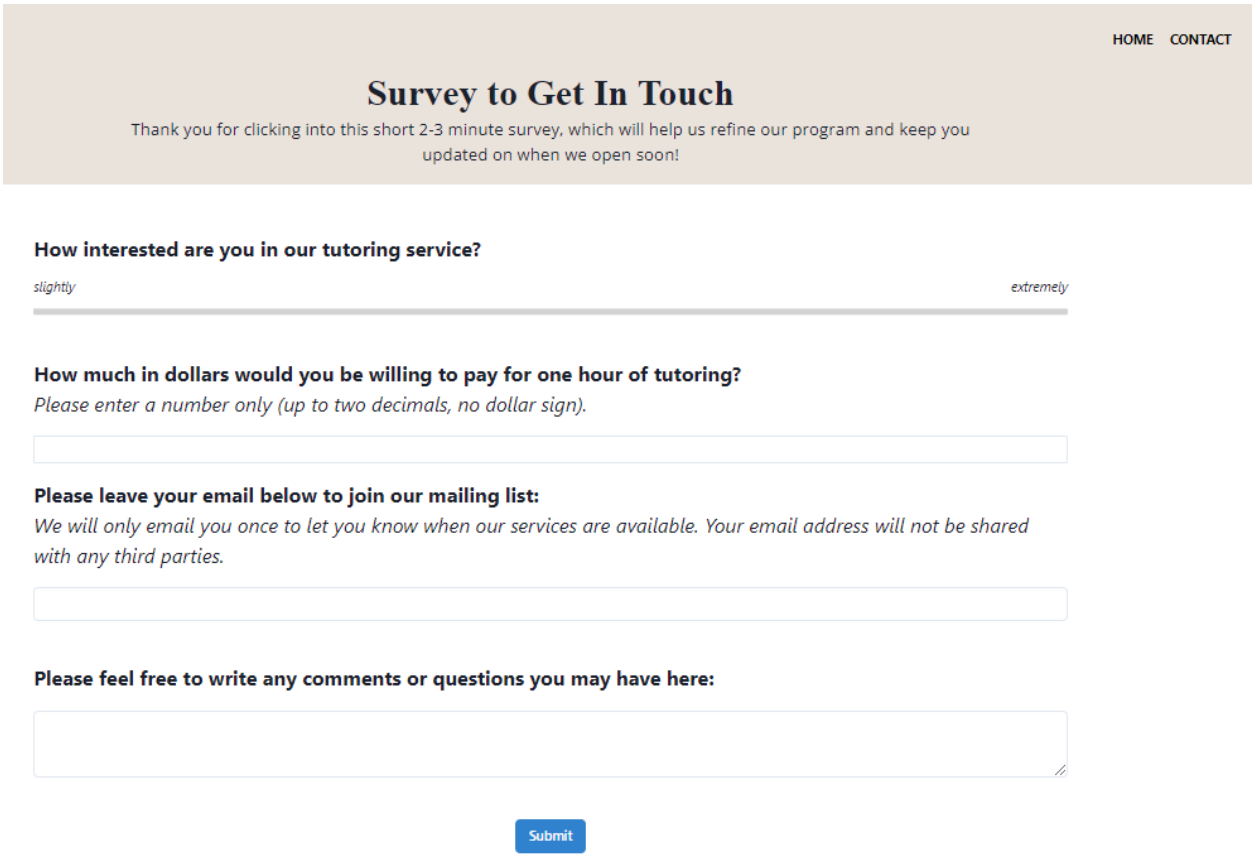


Figure S31: Survey Page—Directed here after clicking Get In Touch

F.2 Benchmarking femininity effects in field experiment with gender pronouns

Table S5: Benchmarking Impact of Femininity and Gender Pronouns on Real-Life Outcomes (OLS)

	Scrolling to Bottom		Clicking Learn More		Clicking Get In Touch		Time on Site (Secs)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
She		-0.019		-0.007		-0.0002		-3.04
		(0.009)		(0.004)		(0.003)		(2.41)
They		-0.012		-0.011		0.005		-2.57
		(0.016)		(0.006)		(0.005)		(2.02)
Synthetic Identity FE	✓	✓	✓	✓	✓	✓	✓	✓
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	0.003	0.005	0.004	0.005	0.004	0.004	0.001	0.003
Within R^2	0.002	0.004	0.001	0.002	0.002	0.002	0.0005	0.002

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

F.3 Significance testing combinations of treatment effects with cluster-robust Wald tests

High femininity effects for binary outcomes for female tutors. For female tutors, relative to medium femininity, high femininity (i.e., High Fem Tertile main effect plus High Fem Tertile × She interaction term) significantly decreases the probability of scrolling ($\beta = -0.046$, $\chi^2(1) = 5.77$, $p = .016$; Column 2) and clicking Get In Touch ($\beta = -0.028$, $\chi^2(1) = 17.63$, $p < .001$; Column 6), but has no significant effect on the probability of clicking Learn More ($\beta = 0.003$, $\chi^2(1) = 0.18$, $p = .670$; Column 4).

Low femininity effects for binary outcomes for female tutors. For female tutors, relative to medium femininity, low femininity (i.e., Low Fem Tertile main effect plus Low Fem Tertile × She interaction term) significantly increases the probability of clicking Learn More ($\beta = 0.008$, $\chi^2(1) = 4.85$, $p = .028$; Column 4) and clicking Get In Touch ($\beta = 0.028$, $\chi^2(1) = 353.37$, $p < .001$; Column 6), but has no significant effect on the probability of scrolling ($\beta = 0.005$, $\chi^2(1) = 0.48$,

$p = .487$; Column 2).

High femininity effects for binary outcomes for non-binary tutors. For non-binary tutors, relative to medium femininity, high femininity (i.e., High Fem Tertile main effect plus High Fem Tertile × They interaction term) has no significant effect on the probability of scrolling ($\beta = -0.047$, $\chi^2(1) = 1.34$, $p = .247$; Column 2), clicking Learn More ($\beta = 0.008$, $\chi^2(1) = 0.78$, $p = .377$; Column 4), or clicking Get In Touch ($\beta = -0.026$, $\chi^2(1) = 2.23$, $p = .136$; Column 6).

Low femininity effects for binary outcomes for non-binary tutors. For non-binary tutors, relative to medium femininity, low femininity (i.e., Low Fem Tertile main effect plus Low Fem Tertile × They interaction term) has a significant positive effect on clicking Learn More ($\beta = 0.004$, $\chi^2(1) = 13.80$, $p < .001$; Column 4) and clicking Get In Touch ($\beta = 0.029$, $\chi^2(1) = 9.26$, $p = .002$; Column 6), but has no significant effect on the probability of scrolling ($\beta = -0.007$, $\chi^2(1) = 0.25$, $p = .621$; Column 2).

High (relative to medium) femininity effects for time spent on website. For female tutors, relative to medium femininity, high femininity significantly increased the time spent on the website ($\beta = 6.20$, $\chi^2(1) = 376.42$, $p < .001$; Column 8, High Fem Tertile main effect plus High Fem Tertile × She interaction term).

For non-binary tutors, relative to medium femininity, high femininity significantly increased the time spent on the website ($\beta = 5.99$, $\chi^2(1) = 66.74$, $p < .001$; Column 8, High Fem Tertile main effect plus High Fem Tertile × They interaction term).

High (relative to low) femininity effects for time spent on website. For male tutors, relative to low femininity, high femininity significantly increased the time spent on the website ($\beta = 13.74$, $\chi^2(1) = 1032.80$, $p < .001$; Column 8, High Fem Tertile main effect minus Low Fem Tertile main effect).

For female tutors, relative to low femininity, high femininity significantly increased the time spent on the website ($\beta = 6.00$, $\chi^2(1) = 199.15$, $p < .001$; Column 8, High Fem Tertile main effect

plus High Fem Tertile × She interaction term minus Low Fem Tertile main effect minus Low Fem Tertile × She interaction term).

For non-binary tutors, relative to low femininity, high femininity significantly increased the time spent on the website ($\beta = 6.76$, $\chi^2(1) = 249.98$, $p < .001$; Column 8, High Fem Tertile main effect plus High Fem Tertile × They interaction term minus Low Fem Tertile main effect minus Low Fem Tertile × They interaction term).

F.4 Robustness to operationalizing femininity more continuously (using mean substituted tertiles)

As a robustness check, we display results for a specification that operationalizes femininity more continuously in Table S6. We opt to substitute each of the three femininity tertiles with their corresponding mean values (as opposed to directly using the femininity score) to help reduce measurement error, as discussed in Web Appendix D.2.4. When interpreting results, a one unit increase in femininity corresponds to, for example, a shift in the underlying femininity scale from androgynous to extremely feminine (see Section 4.2).

The main effects of femininity (Columns 1, 3, 5, 7) are significant and directionally consistent with those from Table 2.

For male tutors, a one unit increase in femininity (i.e., Fem main effect) is not significant for scrolling (Column 2) but significantly decreases the probability of clicking Learn More (Column 4) and Get In Touch (Column 6); and significantly increases the time spent on the website (Column 8).

For female tutors, a one unit increase in femininity (i.e., Fem main effect plus Fem × She interaction term) is not significant for scrolling ($\beta = -0.013$, $\chi^2(1) = 0.25$, $p = .615$; Column 2) but significantly decreases the probability of clicking Learn More ($\beta = -0.014$, $\chi^2(1) = 101.24$, $p < .001$; Column 4) and Get In Touch ($\beta = -0.037$, $\chi^2(1) = 7.01$, $p = .008$; Column 6); and significantly increases the time spent on the website ($\beta = 9.56$, $\chi^2(1) = 29.95$, $p < .001$; Column 8).

For non-binary tutors, a one unit increase in femininity (i.e., Fem main effect plus Fem × They interaction term) is not significant for scrolling ($\beta = 0.000$, $\chi^2(1) = 0.001$, $p = .980$; Column 2) but

significantly decreases the probability of clicking Learn More ($\beta = -0.005$, $\chi^2(1) = 8.10$, $p = .004$; Column 4) and Get In Touch ($\beta = -0.037$, $\chi^2(1) = 16.15$, $p < .001$; Column 6); and significantly increases the time spent on the website ($\beta = 10.97$, $\chi^2(1) = 26.52$, $p < .001$; Column 8).

Altogether, these substantive findings are overall consistent with those from Table 2.

Table S6: Impact of Femininity (Mean-Substituted Tertiles) and Gender Pronouns on Real-Life Outcomes (OLS)

	Scrolling to Bottom		Clicking Learn More		Clicking Get In Touch		Time on Site (Secs)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Fem(ininity) Mean-Substituted Tertiles	-0.020*	-0.053	-0.023***	-0.029**	-0.029***	-0.037*	13.7**	15.4*
	(0.006)	(0.027)	(0.0005)	(0.004)	(0.0007)	(0.012)	(2.01)	(3.60)
She		-0.030***		-0.011*		-0.0009		-1.20
		(0.002)		(0.002)		(0.002)		(1.52)
They		-0.026		-0.017*		0.006		-1.64
		(0.009)		(0.003)		(0.003)		(1.58)
Fem Mean-Substituted Tertiles x She		0.040**		0.015*		0.0005		-5.84
		(0.006)		(0.003)		(0.006)		(4.44)
Fem Mean-Substituted Tertiles × They		0.053*		0.024*		0.0005		-4.43
		(0.014)		(0.006)		(0.006)		(4.91)
Synthetic Identity FE	✓	✓	✓	✓	✓	✓	✓	✓
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	0.003	0.007	0.004	0.006	0.004	0.005	0.002	0.004
Within R^2	0.002	0.005	0.001	0.003	0.002	0.003	0.001	0.003

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

F.5 Robustness to operationalizing femininity using discrete treatment levels

Table S7 displays results for specifications that operationalize femininity using femininity treatment levels (according to the stimuli generation process).

The main effects of femininity (Columns 1, 3, 5, 7) are significant and directionally consistent with those from Table 2, apart from scrolling to bottom. To unpack this discrepancy for scrolling further, we posit that femininity treatment effects are heterogeneous based on the femininity of the base headshot of each synthetic identity, because the effect of femininity tertiles is non-linear (see Table 2; the effect of Low Femininity Tertile is different in magnitude from that of High

Femininity Tertile in Columns 1, 3, 5, 7). In Columns 2, 4, 6, and 8 of Table S7, we include treatment interaction terms with the femininity score of the base headshot—we remove synthetic identity fixed effects in order to be able to identify those interaction terms, since otherwise only at most four terms would be able to be identified. As posited, femininity treatment effects then directionally align with those from Table 2, because the main femininity effects are now interpreted as for a base headshot that is aligned in terms of being similarly androgynous. The interaction terms themselves suggest that the more feminine the initial base headshot is, the more consumers discriminate against those tutors (for all femininity treatment levels) in terms of scrolling, clicking Learn More, and clicking Get In Touch (Columns 2, 4, 6), despite spending more time on the platform (Column 8).

Table S7: Impact of Femininity Treatment Levels on Real-Life Outcomes (OLS)

	Scrolling to Bottom		Clicking Learn More		Clicking Get In Touch		Time on Site (Secs)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant		0.087^{***}		0.019^{***}		0.056^{***}		-3.75^{***}
		(1.71×10^{-12})		(6.31×10^{-13})		(1.59×10^{-12})		(2.46×10^{-10})
Low Fem(ininity) Treatment Level	-0.037^{***}	0.201^{***}	0.069^{***}	0.047^{***}	0.085^{***}	0.166^{***}	-10.3^{***}	-15.7^{***}
	(4.38×10^{-12})	(5.94×10^{-12})	(6.9×10^{-12})	(2.19×10^{-12})	(7.88×10^{-12})	(5.56×10^{-12})	(1.22×10^{-9})	(8.63×10^{-10})
High Fem(ininity) Treatment Level	0.029^{***}	-0.173^{***}	-0.031^{***}	-0.025^{***}	-0.039^{***}	-0.122^{***}	0.797^{***}	10.3^{***}
	(2.48×10^{-12})	(4.11×10^{-12})	(3.89×10^{-12})	(1.58×10^{-12})	(4.45×10^{-12})	(3.88×10^{-12})	(6.86×10^{-10})	(6.05×10^{-10})
Low Fem × Base Fem Score		-0.381^{***}		-0.060^{***}		-0.305^{***}		30.4^{***}
		(1.15×10^{-11})		(4.19×10^{-12})		(1.07×10^{-11})		(1.64×10^{-9})
Medium Fem × Base Fem Score		-0.435^{***}		-0.075^{***}		-0.347^{***}		39.8^{***}
		(1.27×10^{-11})		(4.62×10^{-12})		(1.18×10^{-11})		(1.83×10^{-9})
High Fem × Base Fem Score		-0.398^{***}		-0.085^{***}		-0.344^{***}		41.9^{***}
		(1.31×10^{-11})		(4.61×10^{-12})		(1.21×10^{-11})		(1.86×10^{-9})
Synthetic Identity FE	✓		✓		✓		✓	
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	0.004	0.004	0.004	0.004	0.004	0.002	0.002
Within R^2	0.002		0.001		0.002		0.001	

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

References (Web Appendix)

Anand, P. and Lee, C. (2023). Using Deep Learning to Overcome Privacy and Scalability Issues in Customer Data Transfer. *Marketing Science*, 42(1):189–207.

Athey, S. and Imbens, G. W. (2017). The Econometrics of Randomized Experimentsa. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Economic Field Experiments*, volume 1 of *Handbook of Field Experiments*, pages 73–140. North-Holland.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2021). Using Wasserstein Generative Adversarial Networks for the design of Monte Carlo simulations. *Journal of Econometrics*.

Athey, S., Karlan, D., Palikot, E., and Yuan, Y. (2023). Smiles in Profiles: Improving Fairness and Efficiency Using Estimates of User Preferences in Online Marketplaces. *Available at NBER 30633*.

Atwood, S., Gibson, D. J., Briones Ramírez, S., and Olson, K. R. (2024). Flexibility in continuous judgments of gender/sex and race. *Journal of Experimental Psychology: General*, 153(12):2931–2950.

Bermano, A., Gal, R., Alaluf, Y., Mokady, R., Nitzan, Y., Tov, O., Patashnik, O., and Cohen-Or, D. (2022). State-of-the-Art in the Architecture, Methods and Applications of StyleGAN. *Computer Graphics Forum*, 41(2):591–611.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., and Ramesh, A. (2023). Improving Image Generation with Better Captions. *https://cdn.openai.com/papers/dall-e-3.pdf*.

De Freitas, J., Agarwal, S., Schmitt, B., and Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, 7(11):1845–1854.

Dew, R., Padilla, N., Luo, L. E., Oblander, S., Ansari, A., Boughanmi, K., Braun, M., Feinberg, F. M., Liu, J., Otter, T., Tian, L., Wang, Y., and Yin, M. (2024). Probabilistic Machine Learning: New Frontiers for Modeling Consumers and their Choices. *Available at SSRN 4790799*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*.

Duan, Y., Hsieh, T.-S., Wang, R. R., and Wang, Z. (2020). Entrepreneurs' facial trustworthiness, gender, and crowdfunding success. *Journal of Corporate Finance*, 64:101693.

Duarte, J., Siegel, S., and Young, L. (2012). Trust and Credit: The Role of Appearance in Peer-to-peer Lending. *The Review of Financial Studies*, 25(8):2455–2484.

Enevoldsen, K., Chung, I., Kerboua, I., Kardos, M., Mathur, A., Stap, D., Gala, J., Siblini, W., Krzemiński, D., Winata, G. I., Sturua, S., Utpala, S., Ciancone, M., Schaeffer, M., Sequeira, G., Misra, D., Dhakal, S., Rystrøm, J., Solomatin, R., Çağatan, Ö., Kundu, A., Bernstorff, M., Xiao, S., Sukhlecha, A., Pahwa, B., Poświata, R., GV, K. K., Ashraf, S., Auras, D., Plüster, B., Harries, J. P., Magne, L., Mohr, I., Hendriksen, M., Zhu, D., Gisserot-Boukhlef, H., Aarsen, T., Kostkan, J., Wojtasik, K., Lee, T., Šuppa, M., Zhang, C., Rocca, R., Hamdy, M., Michail, A., Yang, J., Faysse, M., Vatolin, A., Thakur, N., Dey, M., Vasani, D., Chitale, P., Tedeschi,

S., Tai, N., Snegirev, A., Günther, M., Xia, M., Shi, W., Lù, X. H., Clive, J., Krishnakumar, G., Maksimova, A., Wehrli, S., Tikhonova, M., Panchal, H., Abramov, A., Ostendorff, M., Liu, Z., Clematide, S., Miranda, L. J., Fenogenova, A., Song, G., Safi, R. B., Li, W.-D., Borghini, A., Cassano, F., Su, H., Lin, J., Yen, H., Hansen, L., Hooker, S., Xiao, C., Adlakha, V., Weller, O., Reddy, S., and Muennighoff, N. (2025). MMTEB: Massive Multilingual Text Embedding Benchmark. *arXiv:2502.13595*.

Ert, E., Fleischer, A., and Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management*, 55:62–73.

Goodfellow, I. (2017). NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv:1701.00160*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27.

Graham, J. R., Harvey, C. R., and Puri, M. (2017). A Corporate Beauty Contest. *Management Science*, 63(9):3044–3056.

Hayes, A. F. and Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1):77–89.

Hsieh, T.-S., Kim, J.-B., Wang, R. R., and Wang, Z. (2020). Seeing is believing? Executives' facial trustworthiness, auditor tenure, and audit fees. *Journal of Accounting and Economics*, 69(1):101260.

Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Chen, S., and Cao, L. (2024). Diffusion Model-Based Image Editing: A Survey. *arXiv:2402.17525*.

Jaeger, B., Wagemans, F. M. A., Evans, A. M., and van Beest, I. (2018). Effects of Facial Skin Smoothness and Blemishes on Trait Impressions. *Perception*, 47(6):608–625.

Kaji, T., Manresa, E., and Pouliot, G. (2023). An Adversarial Approach to Structural Estimation. *Econometrica*, 91(6):2041–2063.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv:1312.6114*.

Liang, H., Perona, P., and Balakrishnan, G. (2023). Benchmarking Algorithmic Bias in Face Recognition: An Experimental Approach Using Synthetic Faces and Human Evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4977–4987.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

Ludwig, J. and Mullainathan, S. (2024). Machine Learning as a Tool for Hypothesis Generation. *The Quarterly Journal of Economics*, page qjad055.

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

Olivola, C. Y. and Todorov, A. (2010). Elected in 100 milliseconds: Appearance-Based Trait Inferences and Voting. *Journal of Nonverbal Behavior*, 34(2):83–110.

Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., and Lischinski, D. (2021). StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094.

Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., and Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, 119(17):e2115228119.

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:2307.01952*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Ratliff, L. J., Burden, S. A., and Sastry, S. S. (2013). Characterization and computation of local Nash equilibria in continuous games. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 917–924.

Rule, N. O. and Ambady, N. (2008). The Face of Success: Inferences From Chief Executive Officers' Appearance Predict Company Profits. *Psychological Science*, 19(2):109–111.

Russell, R. (2009). A Sex Difference in Facial Contrast and its Exaggeration by Cosmetics. *Perception*, 38(8):1211–1219.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. (2016). Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Scott, I. M., Clark, A. P., Josephson, S. C., Boyette, A. H., Cuthill, I. C., Fried, R. L., Gibson, M. A., Hewlett, B. S., Jamieson, M., Jankowiak, W., Honey, P. L., Huang, Z., Liebert, M. A., Purzycki, B. G., Shaver, J. H., Snodgrass, J. J., Sosis, R., Sugiyama, L. S., Swami, V., Yu, D. W., Zhao, Y., and Penton-Voak, I. S. (2014). Human preferences for sexually dimorphic faces may be evolutionarily novel. *Proceedings of the National Academy of Sciences*, 111(40):14388–14393.

Serengil, S. I. and Ozpinar, A. (2021). HyperExtended LightFace: A Facial Attribute Analysis Framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4.

Solatorio, A. V. (2024). GISTEmbed: Guided In-sample Selection of Training Negatives for Text Embedding Fine-tuning. *arXiv.2402.16829*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ukasz Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Veselovsky, V., Ribeiro, M. H., and West, R. (2023). Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. *arXiv:2306.07899*.

Zietsch, B. P., Lee, A. J., Sherlock, J. M., and 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. *Psychological Science*, 26(9):1440–1448.