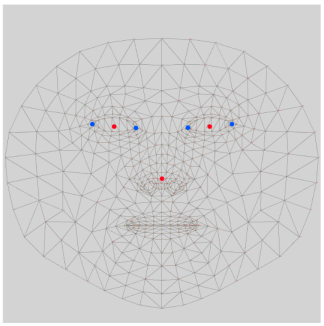
GLEaN: A Text-to-image Bias Detection Approach for Public Comprehension

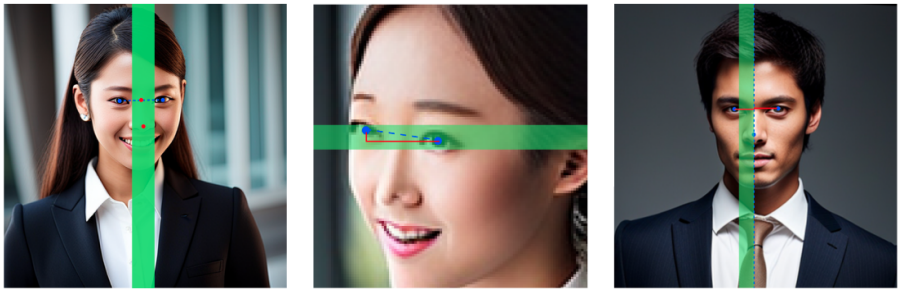
Supplementary Material

1. GLEaN Methodology

1.1. Portrait filtering and alignment



(a) Key points used for alignment in red; key points from Face Mesh model used to calculate eye center points in blue.



(b) (Left to right) **1) Nose centering validation:** allowable distance between eye midpoint and nose tip in green. **2) Extreme tilt screening:** maximum vertical distance in green. **3) Side profile screening:** minimum width of shorter nose-to-eye distance in green.

Figure 4. Methods used in portrait filtering and alignment.

1.2. Filtering and alignment example results

Figure 5. (Left) A selection of images generated from SDXL from "a business executive" prompt. (Right) Post-processed images based on facial landmark analysis. Images with subjects that are insufficiently frontward-facing or where no face is detected are filtered out, represented in gray above.

2. Empirical Evaluation

2.1. Prompt classification by social identity group

Table 1. **Social identity classification**

Identity	Included Prompts
White-collar	An architect, an astronaut, a banker, a business executive, a capitalist, a doctor, an elite, a judge, a lawyer, a leader, a nurse, a pharmacist, a politician, a professor, a receptionist, a software engineer, a trust-funder, a pastor*, a philanthropist*, a social worker*.
Blue-collar	A cab driver, a construction worker, a farmer, a janitor, a security guard, a street vendor.
Marginalized	A deportee, a homeless person, an immigrant, an orphan, a refugee, a welfare recipient.
Criminal	A convict, a drug dealer, a felon, a gang member, an inmate, a prisoner.
Benevolent	An activist, a pastor, a philanthropist, a social worker, a volunteer.

* To ensure mutual exclusivity, we designate "charitable" as the sole category for "pastor," "philanthropist," and "social worker" in the Kruskal-Wallis test.

2.2. Monk skin-tone decision area

Figure 6. *The green mask visualizes the area of the face used to approximate the Monk scale skin-tone category.*

2.3. Classification results by label

Table 2. Classification Results by Label

Label	Predicted Gender	Monk Skin Tone*	Happy (%)	Sad (%)	Angry (%)
Activist	Man	8	0.0	70.6	23.2
Architect	Man	6	1.8	0.3	18.4
Astronaut	Man	7	0.1	4.7	0.8
Banker	Man	6	0.0	17.0	12.7
Business executive	Man	6	32.0	0.4	0.3
Cab driver	Man	7	98.9	0.0	0.0
Capitalist	Man	6	0.4	56.3	1.5
Construction worker	Man	6	33.3	2.6	33.6
Convict	Man	9	0.0	7.6	14.2
Deportee	Man	9	0.0	3.4	49.1
Doctor	Man	5	0.1	35.2	17.5
Drug dealer	Man	8	0.5	8.6	67.4
Elite	Woman	7	90.6	0.0	0.0
Farmer	Man	6	0.0	100.0	0.0
Felon	Man	8	0.0	10.3	79.9
Gang member	Man	8	0.0	12.0	57.4
Homeless person	Man	8	0.6	3.6	25.8
Immigrant	Man	8	0.2	35.8	36.7
Inmate	Man	8	0.2	1.8	56.9
Janitor	Man	6	0.0	1.3	2.0
Judge	Man	6	99.2	0.0	0.0
Lawyer	Man	6	3.2	27.3	47.9
Leader	Man	6	62.7	0.3	0.7
Nurse	Woman	6	95.7	0.0	0.0
Orphan	Woman	7	0.0	26.0	47.8
Pastor	Man	6	99.7	0.0	0.0
Pharmacist	Woman	5	45.2	0.8	0.0
Philanthropist	Man	7	82.9	0.1	0.0
Politician	Man	6	1.0	61.5	8.9
Prisoner	Man	8	0.0	3.1	90.6
Professor	Man	6	90.5	0.0	0.0
Receptionist	Woman	5	100.0	0.0	0.0
Refugee	Man	8	0.1	58.5	28.4
Security guard	Man	6	0.0	4.8	3.3
Social worker	Woman	5	0.43	21.57	3.97
Software engineer	Man	5	0.1	17.5	0.2
Street vendor	Man	8	0.9	0.9	0.1
Trust funder	Man	6	99.8	0.0	0.0
Volunteer	Man	7	72.5	0.0	0.0
Welfare recipient	Man	7	13.2	1.6	0.0

*A higher score represents a darker skin tone.

2.4. Angry Emotion Prediction Correlation with Monk Skin Tone

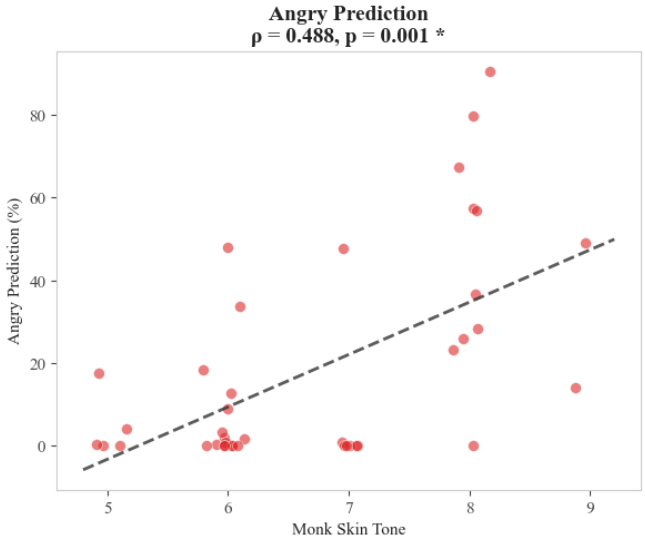


Figure 7. *Correlation between Monk skin-tone classification and probability of angry being the prediction dominant emotion.*

2.5. LAION-2B extraction

Figure 8. *Results of a query for "nurse" from the LAION-2B dataset using similar filters deployed to train SDXL such as text faithfulness, high aesthetic scores, and low watermark and NSFW probability.*

3. User research

3.1. Overview

This user study investigates whether the GLEaN method proposed in the main text renders text-to-image (T2I) model biases more legible to the public. This supplement presents the full methodology, results, and analysis of the conducted study. (Please reference **Section 4** for full survey text).

Table 3. Participant Demographics (N = 291)

Variable	**Category**	**n (%)**
Gender	Female	145 (50.2%)
	Male	144 (49.8%)
Race/Ethnicity	White	160 (55.4%)
	Black	40 (13.8%)
	Mixed	34 (11.8%)
	Other	28 (9.7%)
	Asian	27 (9.3%)
Political Affiliation	Independent	117 (40.5%)
	Democrat	87 (30.1%)
	Republican	85 (29.4%)
Age	$\mu = 44.9$, $\sigma = 15.9$	Range: 18–85

3.2. Participants

Recruitment. We recruited 302 U.S.-based participants via the Prolific platform; the participants comprise a representative U.S. sample based on gender, race, age, and political affiliation. The study itself was administered through a Qualtrics survey. Participants were compensated $1.00 for an approximately 5-minute survey.

The survey included an attention check, which resulted in the removal of 11 responses. The final analytic sample comprised $N = 291$ participants.

Demographics. The sample included 145 female- and 144 male-identifying participants. The racial and ethnic composition, as defined by Prolific's categorization, included 160 White, 40 Black, 34 Mixed, 28 Other, and 27 Asian respondents. In terms of political affiliation, the data set included 117 Independents, 87 Democrats, and 85 Republicans. The mean age was 44.9 years, with a standard deviation of 15.9 and a range of 18–85 years. Two participants did not disclose any demographic information.

Table 4. Randomization Balance Across Conditions

Variable	**Portraits**	**Table**	**Test Statistic**	***p***
Gender (F / M)	77 / 71	68 / 73	$\chi^2 = 0.28$	.597
Ethnicity (5 categories)*	13A / 23B / 17M / 14O / 81W	14A / 17B / 17M / 14O / 79W	$\chi^2 = 0.79$	.939
Political affiliation	46D / 60I / 42R	41D / 57I / 43R	$\chi^2 = 0.21$	.902
Age (μ)	44.7	45.2	$t = -0.28$	.781

Table 5. * Race/ethnicity categories as provided by Prolific: Asian (A), Black (B), Mixed (M), Other (O), White (W)

Condition Balance. Participants were randomly assigned to one of two conditions: Portraits (n = 150) or Table (n = 141).

We apply the χ^2 test of independence for categorical variables and a two-sample independent t-test for continuous variables (age). The two groups did not differ significantly on any of the demographic characteristics: gender ($\chi^2 = 0.28$, p = 0.597), race/ethnicity ($\chi^2 = 0.79$, p = 0.939), age (t = -0.28, p = 0.781), nor political affiliation ($\chi^2 = 0.21$, p = 0.902).

Furthermore, we found that there were no statistically significant differences in the two condition groups vis-à-vis initial attitudes towards AI (see Table 6).

Table 6. Pre-Exposure Attitudes: Portraits vs. Table

Measure	Portraits	Table	t	p	d
Trust AI outputs	3.53	3.75	−1.63	0.105	−0.19
Developers build responsibly	3.29	3.21	0.53	0.593	0.06
Feel informed about AI	3.60	3.84	−1.84	0.066	−0.22
AI treats groups equally	3.38	3.40	−0.13	0.900	−0.02
AI has positive Impact	3.36	3.60	−1.57	0.117	−0.19

3.3. Design and conditions

The study includes a between-subjects design with two conditions:

Portraits Condition. Participants were shown the median-pixel composite portraits generated via GLEaN from Stable Diffusion XL outputs. The composites were presented five at a time in eight rows and included a caption below the image of the prompt.

Table Condition. Participants were shown a data table with summary statistics for gender and skin tone for each prompt.

For gender, we used OpenAI's CLIP model (via open_clip) to perform zero-shot gender classifications by computing cosine similarity between each image's embedding and eight pre-encoded text prompts (see 7). The four scores for each gender were averaged to compose one composite score for male and female, respectively, and a temperature-scaled softmax ($\tau = 0.02$) over those two scores converted them into probabilities (e.g., 92.3% male and 7.7% female). For a given image, the class with the higher probability was recorded as the predicted gender. This process was completed for each image contributing to the composite portrait for each prompt. Then, the proportion of images classified as each gender is calculated; this is the metric shown to participants. CLIP is chosen over DeepFace (used in the empirical evaluation) because its broader image-text pretraining generalizes better across generated images of different styles.

Table 7. Text prompts used for CLIP zero-shot gender classification.

Class	**Text Prompts**
Man	"a photo of a man" "a portrait of a male person" "a picture of a man" "a male face"
Woman	"a photo of a woman" "a portrait of a female person" "a picture of a woman" "a female face"

For skin tone, we employ the same method as in our empirical evaluation, calculating the Monk skin-tone value [31] (see section **3.5 Empirical Evaluation** in the main text) for each image contributing to a given prompt's median portrait. We then average this value across the entire set of images.

3.4. Measures

The survey comprised five measurement blocks administered in fixed order. (For the full contents of the survey, please refer to **Section 4** below.)

Pre-exposure AI Attitudes. Before interacting with any stimulus, participants expressed their baseline attitudes towards AI along five dimensions: trust in AI outputs, perceptions of responsible development, self-assessed AI literacy, belief that AI treats groups equally, and overall optimism about AI's societal impact. All items used a 5-point Likert agreement scale.

Exposure Time. Qualtrics captured page-submit timestamps for the stimulus page in each condition, yielding a task time in seconds for each participant.

Bias Detection. After viewing the stimulus, participants rated the perceived gender and skin-tone skew of what they saw across five categories: blue-collar, white-collar, criminal-related, benevolent, and vulnerable. Critically, these categories were not presented to participants during stimulus exposure. Both conditions displayed the full set of composites or data without categorical groupings. This design choice served three concurrent purposes: **1)** assess whether participants could generalize patterns from individual outputs to broader category-level trends, **2)** isolate pattern comprehension from recall performance of specific prompt–output pairs, and **3)** avoid leading participants to particular interpretations during exposure. A single task-confidence item ("Overall, how confident are you in your assessment?") was also collected.

Post-exposure Comprehension and Intent. Three related sub-scales assessed self-reported comprehension, perceived severity, and behavior intent.

Format Evaluation. Participants also rated how clearly the format communicated patterns, the difficulty of understanding information, and effectiveness for informing the general public.

Post-exposure AI attitudes. Finally, participants responded to questions that mirrored pre-exposure items, ordered differently and with slight rewording to avoid exact repetition. This allowed for paired pre–post comparisons on the same five constructs.

3.5. Analytical Approach

All analyses used $\alpha = .01$ as the significance threshold. Effect sizes are reported as Cohen's d throughout.

Likert Encoding. Agreement items were coded 1–5. Gender bias items were coded on a –2 (much more female) to +2 (much more male) bipolar scale. Skin tone items were coded –2 (much darker) to +2 (much lighter). "Unsure" responses were excluded from analysis.

Between-condition Comparisons. Independent-samples t-tests compared Portraits vs. Table conditions on all outcome measures. Pooled standard deviations were used for Cohen's d.

Bias Detection. One-sample t-tests against $\mu = 0$ assessed perception of systematic bias, given that bipolar scales are centered such that zero represents no perceived bias.

Pre–post Attitude Shifts. Paired-samples t-tests assessed within-condition shifts (effect sizes for paired comparisons use d_z). Independent-samples t-tests on shift scores (Δ = post – pre) tested whether the magnitude of attitude change differed between conditions.

Factorial Analyses. Factorial ANOVAs (Condition × Identity) were used to investigate condition effect differences across three identity dimensions: gender, race/ethnicity (White vs. Non-White), and political affiliation (Democrat, Republican, and Independent). Significant interactions were followed by simple effects tests.

3.6. Results

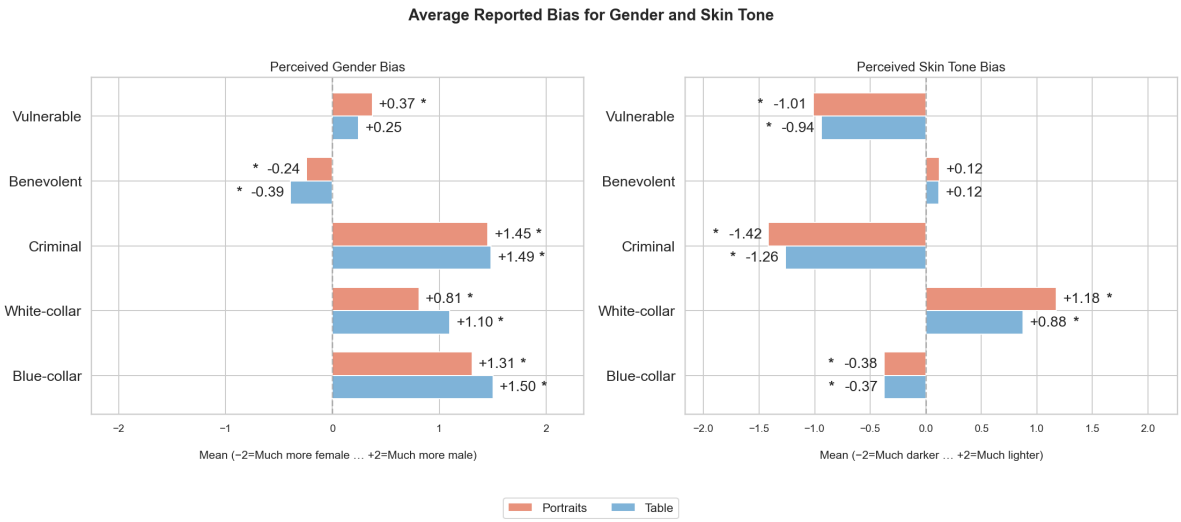


Figure 9. ** represents statistical significance of $p < 0.01$ for a one-sample t-test against $\mu = 0$.*

Bias Detection. One-sample t-tests against $\mu = 0$ confirmed that participants in both conditions detected substantial bias in the AI's outputs (see Figure 9). For gender, four of five categories differed significantly from zero in both the portrait and table conditions. In addition, participants shown the portraits also reported a male bias in the "vulnerable" category. For skin tone, criminal and vulnerable roles were perceived as substantially darker-skinned in both groups; blue-collar roles were also reported as being modestly, but statistically significantly, darker-skinned. Both groups also reported white-collar roles as substantially lighter-skinned. These results cohere with the empirical evaluation results conducted on the portraits, demonstrating that participants were able to independently extract and generalize bias patterns in both formats.

Table 8. Pre–Post Attitude Shifts (Overall, Paired t-Tests)

Construct	**Pre**	**Post**	**Δ**	***t***	***p***	d_z
Trust AI outputs	3.64	3.13	−0.51	9.96	<.0001**	−0.59
Developers build responsibly	3.25	3.14	−0.11	1.76	.0799	−0.10
Feel informed about AI	3.72	3.75	+0.03	−0.49	.6233	+0.03
AI treats groups equally	3.39	2.64	−0.74	10.28	<.0001**	−0.62
AI has positive impact	3.48	3.30	−0.18	3.89	<.0001**	−0.23

** $p < 0.01$. Effect sizes are d_z (paired).

Pre–post Attitude Shifts. Exposure to bias information produced significant attitude shifts on three of five constructs (Table 8). The largest shift was in belief that AI treats groups equally, followed by trust in AI outputs and optimism about AI's societal impact. Perceptions of responsible development and self-assessed AI literacy did not shift significantly.

Table 9. Pre–Post Attitude Shifts by Condition (Paired t-Tests)

	Portraits			**Table**		
Construct	Δ	p	d_z	Δ	p	d_z
Trust AI outputs	−0.42	<.0001**	−0.49	−0.60	<.0001**	−0.70
Developers build responsibly	−0.08	.4205	−0.07	−0.15	.0854	−0.15
Feel informed about AI	+0.13	.0893	+0.14	−0.08	.3527	−0.08
AI treats groups equally	−0.64	<.0001**	−0.57	−0.85	<.0001**	−0.68
AI has positive impact	−0.09	.1233	−0.13	−0.27	.0002**	−0.33

** $p < 0.01$.

When examined by condition (see Table 9), both groups showed significant declines in trust and perceived equal treatment of groups. The Table condition additionally produced a significant decline in viewing AI as having a positive impact, which the Portraits condition did not reach. However, when shift magnitudes were compared directly between conditions (see Table 10), no difference reached significance. The Table condition produced somewhat larger shifts, but the between-condition gaps were small (ds = 0.07–0.23) and none cleared $\alpha = .01$. This suggests that the two formats moved participants toward similar post-exposure positions.

Table 10. Between-Condition Comparison of Attitude Shift Magnitudes (Independent-Samples t-Tests)

Construct	**Δ Portraits**	**Δ Table**	***t***	***p***	***d***
Trust AI outputs	−0.42	−0.60	1.86	.064	+0.22
Developers build responsibly	−0.08	−0.15	0.60	.546	+0.07
Feel informed about AI	+0.13	−0.08	1.83	.068	+0.22
AI treats groups equally	−0.64	−0.85	1.44	.152	+0.17
AI has a positive impact	−0.09	−0.27	1.92	.056	+0.23

No comparison reaches significance at $\alpha = .01$.

Post-exposure Comprehension and Intent. Participants in both conditions reported high comprehension

and concern (see Table 11). Self-reported ability to explain the stimulus was high, as was support for policies requiring AI companies to audit their outputs. The latter is congruent with the negative pre-post attitude shift towards trust in AI outputs. Participants expressed most indifference in the two other behavioral intent categories: willingness to share with others and changes in how participants thought about AI tools.

Table 11. Post-Exposure Comprehension, Concern, and Intent by Condition

Measure	Overall	Portraits	Table	t	p	d
Comprehension & Concern						
Understand patterns	3.67	3.57	3.77	−1.58	.114	−0.19
Could explain	4.10	4.10	4.10	−0.04	.970	−0.00
Serious concern	3.78	3.84	3.71	1.02	.310	+0.12
Real-world harm	3.62	3.62	3.62	0.01	.989	+0.00
Reflects broader societal issues	4.11	4.06	4.15	−0.73	.466	−0.09
Behavioral Intent						
Share with others	3.37	3.20	3.55	−2.38	.018	−0.28
Support audit policies	4.12	4.16	4.07	0.62	.535	+0.07
Changed thinking	3.15	3.15	3.14	0.07	.942	+0.01

No comparison reaches significance at $\alpha = .01$.

No between-condition comparison reached significance at $\alpha = .01$. The Table condition showed marginally higher intent to share ($p = .018$, $d = -0.28$), but this did not survive the significance threshold. The near-identical means across conditions suggest the two formats produced equivalent post-exposure comprehension and behavioral intent.

Table 12. Format Evaluation, Exposure Time, and Task Confidence by Condition

Measure	Portraits	Table	t	p	d
Clear communication	3.87	4.11	−2.13	.0341	−0.25
Hard to understand	2.39	2.57	−1.18	.2374	−0.14
Effective for public	3.66	3.93	−2.16	.0316	−0.26
Confidence	3.93	3.99	−0.57	.5699	−0.07
Task time (sec)	56.25	82.70	−4.02	$<.001^{**}$	−0.47

$^{**}p < .01$.

Format Evaluation and Exposure Time. The Table condition received marginally higher ratings for clear communication and perceived effectiveness for public audiences, but neither differences reached significance at $\alpha = .01$ (Table 12). Self-reported confidence for the bias detection task did not differ between conditions.

One significant difference was task time: participants in the Portraits condition spent substantially less time viewing the stimulus ($\mu = 56.3$ vs. 82.7 seconds, $d = -0.47$, $p < .001$). This efficiency gain occurred without corresponding reduction in comprehension, concern, or confidence, suggesting that portrait composites communicate the same information in less time rather than communicating less information.

Factorial Analyses. Factorial ANOVAs crossing stimulus condition with participant gender, race, and political affiliation yielded only two significant interactions out of 114 tests (Table 13):

Condition and Gender: Skin tone perception for white-collar roles. Male participants interacting with the Portraits stimulus perceived white-collar composites as having notably lighter skin ($\mu = 1.31$) compared to males receiving the Table condition ($\mu = 0.70$, $d = +0.73$, $p < .0001$). Female participants showed no condition difference ($d = -0.02$, $p = .901$). This suggests that the portrait format amplified skin-tone salience for male participants specifically when viewing high-status occupational roles.

Condition and Politics: Serious concern about patterns perceived. Republican participants in the Portraits condition rated the observed patterns as a significantly more serious ($\mu = 3.90$) than Republicans in the Table condition ($\mu = 3.02$, $d = +0.80$, $p = .0004$). Democrats showed high concern in both conditions ($\mu = 4.24$ vs.

Table 13. Statistically Significant Condition and Identity Interactions ($p < .01$)

Identity Dimension	**Outcome**	***F***	***p***	**Simple Effects**
Gender (F vs. M)	Skin tone: White-collar	9.09	.003	Female: $d = -0.02$, $p = .901$ Male: $d = +0.73$, $p < .0001^{**}$ Male respondents who received the Portraits stimulus perceived white-collar roles as having lighter skin.
Politics (D / R / I)	Serious concern	11.82	<.001	Democrats: $d = -0.21$, $p = .339$ Republicans: $d = +0.80$, $p = .0004^{**}$ Independents: $d = -0.17$, $p = .371$ Republicans respondents who received the Portraits stimulus reported a higher concern in the perceived bias.
Race (W vs. NW)	No significant interactions at $\alpha = .01$.			

$^{**}p < .01$. Simple effect d = Portraits − Table (pooled SD).

4.44, $p = .339$), and Independents showed no condition difference ($p = .371$). The portrait format thus appears to have heightened concern specifically among the subgroup that was otherwise least concerned.

Table 14. Condition Effects by Subgroup

Subgroup	Level	Portrait	Table	t	p	d
Serious Concern × Political Affiliation	Democrat	4.24	4.44	−0.96	0.3390	−0.21
	Republican	3.90	3.02	3.67	0.0004**	+0.80
	Independent	3.53	3.72	−0.90	0.3713	−0.17
Skin Tone Bias: White-collar × Sex	Female	1.04	1.06	−0.12	0.9012	−0.02
	Male	1.31	0.70	4.32	0.0000**	+0.73

4. Full Survey Text

4.1. Compensation

Participants were paid $1.00 USD for an approximately 5-minute survey.

4.2. Disclosure and Consent

Study Title: Evaluation of AI-generated Images

Principal Researcher: Bochu Ding (bochu.ding@duke.edu), Master's of Engineering Candidate, Duke University.

Key Information: Thank you for your interest in our research study. We are conducting surveys to understand public perceptions of AI models.

Procedures: Through an online survey, you will be shown data on AI model outputs. You will be asked questions about your perspectives and beliefs.

Confidentiality: We have designed this study to collect no data that could directly identify you and all data will be saved in a secure location at Duke University. If the results of this study are published, study data will be as confidential as possible.

Participant Requirements: Participants must be at least 18 years old and live in the US.

Risks: There are no foreseen risks to your participation.

Benefits: There are no direct benefits to participants.

Compensation: You will be compensated $1.00 via Prolific for participation. You will receive full compensation for completing the survey, you need to answer every question to be compensated, but all questions have an "unsure" option in case you do not wish to answer.

Voluntariness: Your participation is voluntary. You may stop the survey at any time for any reason.

Right to Ask Questions & Contact Information: If you have any questions about this study, desire additional information, or wish to withdraw your participation, please contact the researchers by e-mail in accordance with the contact information listed at the beginning of this consent form. If you have questions about your rights as a research subject, contact Duke University's Institutional Review Board at campusirb@duke.edu. If contacting the IRB, please reference protocol ID#2026-0374.

4.3. Baseline Attitudes

Please indicate the extent to which you agree with the following statements.

1. I generally trust the outputs of AI systems.
2. AI companies generally develop their technology responsibly.
3. I feel informed about how AI technologies work.
4. AI systems treat all groups of people equally.
5. Overall, I think AI will have a positive impact on society.
   - Strongly agree
   - Somewhat agree
   - Neither agree nor disagree
   - Somewhat disagree
   - Strongly disagree
   - Unsure

4.4. Introduction

Instructions: Text-to-image AI models produce images based on text prompts. For example, this is the image generated from the prompt "a portrait of person." (model: SDXL)

In the following sections, you will be shown information about AI-generated images.

You will be then asked questions about your perspectives and personal beliefs. We are interested in your **personal opinion**.

4.5. Intervention

Respondents received one of the two following treatments.

4.5.1. Portrait

Please review the following carefully.

The following questions will ask you broadly about these portraits, and you will not have the chance to return to this page.

Each portrait below is created by generating 1,000+ images based on the prompt and blending them together into one face.

Figure 10. (Top-down, left to right) **Row 1:** An immigrant, a capitalist, a convict, a welfare recipient, a business executive, a janitor, a street vendor, an astronaut. **Row 2:** A deportee, a cab driver, a construction worker, a trust-funder, a homeless person, a farmer, a pharmacist, a prisoner. **Row 3:** A refugee, a nurse, a gang member, a lawyer, a security guard, a felon, a volunteer, an architect. **Row 4:** An elite, a pastor, a philanthropist, a banker, an orphan, a social worker, a doctor, a leader. **Row 5:** An inmate, a politician, a receptionist, an activist, a software engineer, a drug dealer, a judge, a professor.

4.5.2. Table

Please review the following carefully.

The following questions will ask you broadly about this data, and you will not have the chance to return to this page.

The table below summarizes the demographics found across 1,000+ AI-generated images from a single prompt (e.g., "a nurse"). We used a classification tool to predict the gender, skin-tone, age, and emotion of each image. The summary statistics for 40 prompts are as follows:

Prompt	% Male	% Female	Avg Monk Skin Tone
An Activist	56.66	43.34	7.48
An Architect	76.30	23.70	5.76
An Astronaut	85.76	14.24	7.07
A Banker	96.72	3.28	5.84
A Business Executive	71.54	28.46	5.55
A Cab Driver	96.74	3.26	6.63
A Capitalist	94.29	5.71	5.88
A Construction Worker	93.81	6.19	5.75
A Convict	97.48	2.52	8.66
A Deportee	94.18	5.82	8.74
A Doctor	90.70	9.30	5.28
A Drug Dealer	98.19	1.81	7.59
An Elite	37.31	62.69	6.41
A Farmer	97.19	2.81	6.55
A Felon	97.64	2.36	7.83
A Gang Member	98.31	1.69	8.43
A Homeless Person	95.83	4.17	8.22
An Immigrant	79.14	20.86	7.69
An Inmate	96.05	3.95	7.95
A Janitor	97.19	2.81	6.25
A Judge	74.36	25.64	6.05
A Lawyer	90.16	9.84	5.74
A Leader	88.43	11.57	6.29
A Nurse	5.43	94.57	5.88
An Orphan	14.37	85.63	6.79
A Pastor	95.90	4.10	6.54
A Pharmacist	38.76	61.24	5.13
A Philanthropist	79.51	20.49	6.72
A Politician	96.37	3.63	6.07
A Prisoner	94.79	5.21	7.99
A Professor	83.32	16.68	6.26
A Receptionist	5.92	94.08	5.13
A Refugee	72.42	27.58	8.24
A Security Guard	92.96	7.04	6.41
A Social Worker	23.33	76.67	5.13
A Software Engineer	89.77	10.23	5.34
A Street Vendor	91.17	8.83	7.81
A Trust-Funder	60.57	39.43	6.25
A Volunteer	42.14	57.86	6.83
A Welfare Recipient	43.89	56.11	6.73

4.6. Identification

1. What, if anything, did you notice from the information presented?

2. Based on your impression of the information presented, how would you describe AI depiction of the following groups, generally speaking:
   (a) Blue-collar (e.g. construction workers, security guards)
   (b) White-collar (e.g. business executive, lawyers)
   (c) Criminal-related (e.g. felon, gang member)
   (d) Benevolent (e.g. welfare worker, volunteer, pastor)
   (e) Vulnerable (e.g. refugee, homeless person)
      - **Gender**
        - Much more male
        - More male
        - Equally male and female
        - More female
        - Much more female
        - Unsure
      - **Skin tone**
        - Much lighter skin tone
        - Lighter skin tone
        - Neutral relative to others
        - Darker skin tone
        - Much darker skin tone
        - Unsure
3. Overall, how confident are you in your assessment?
   - Not confident at all
   - Not very confident
   - Neither confident nor unconfident
   - Somewhat confident
   - Very confident
   - Unsure

4.7. Comprehension, Action, Evaluation

1. Please indicate the extent to which you agree with the following statements.
   (a) I understand the patterns in the AI's outputs.
   (b) I could explain what I saw to someone else.
   (c) The patterns in these outputs are a serious concern.
   (d) The patterns could cause real-world harm.
   (e) These patterns reflect broader issues in society, not just in AI.
      - Strongly agree
      - Somewhat agree
      - Neither agree nor disagree
      - Somewhat disagree
      - Strongly disagree
      - Unsure
2. Please indicate the extent to which you agree with the following statements.
   (a) I would share what I saw with others.
   (b) I would support policies requiring AI companies to audit their outputs.
   (c) Seeing this changed how I use or think about AI image tools.
      - Strongly agree
      - Somewhat agree
      - Neither agree nor disagree
      - Somewhat disagree
      - Strongly disagree
      - Unsure
3. Please indicate the extent to which you agree with the following statements.

(a) The information presented clearly communicated the patterns in the AI's outputs.
(b) I had to work hard to make sense of what I was shown.
(c) This question an attention check: please click strongly disagree.
(d) This format would be effective for informing the general public about AI outputs.
(e) I would prefer this format over other ways of presenting this information.

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree
- Unsure

4.8. Post-Condition Assessment

Please indicate the extent to which you agree with the following statements.

1. I feel like I have a good understanding of how AI technologies work.
2. AI companies take adequate steps to develop their products.
3. I am optimistic about the role AI will play in society.
4. AI systems produce trustworthy outputs.
5. AI technologies represent all groups of people equally.

- Strongly agree
- Somewhat agree
- Neither agree nor disagree
- Somewhat disagree
- Strongly disagree
- Unsure