Inclusive Output Is Only the Start: Evaluating Human Response to AI-Generated Equality-and-Equity Imagery

Abstract

Image-generation systems increasingly include safeguards intended to reduce biased or exclusionary outputs, including outputs that foreground equality, equity, and visible inclusion. Yet the social effect of such inclusive imagery cannot be inferred from the image alone: viewers may ignore, resist, affirm, or reinterpret what the system produces. Here we examine human response to eight DALL-E 3 equality and equity images in a platform-shuffled survey of 172 participants, yielding 1,376 participant-image events. Using ordinal models for a repeated 1-5 perspective-change item, binary endorsement summaries, affect co-report checks, and blind human coding of open-text responses, we find that response varies across images and affective readings. Inspired is the most stable positive signal, while Uncomfortable is negative in the primary ordinal model and weaker in checks. These results frame inclusive AI output as a starting point for sociotechnical evaluation, not as evidence of social impact.

Keywords: Generative AI, AI-generated equality and equity imagery, Perspective-change endorsement, Co-reported affect, Mixed methods.

1. Introduction

Image-generation systems are increasingly designed not only to follow user prompts, but also to reduce harmful, stereotyped, or exclusionary outputs. Prior audits show that accessible text-to-image systems can amplify demographic stereotypes at scale, while work on gender presentation shows that the distributions of generated images remain difficult to evaluate for downstream use (Bianchi et al., 2023; Zhang et al., 2023). In response, model providers and system designers have adopted output-side safeguards such as filtering, blocking, and balancing strategies intended to reduce visibly biased generations (Bianchi et al., 2023). These safeguards matter, but they raise a second question that is not answered by inspecting the generated image: what happens when people view inclusive outputs produced by these systems?

That question is central for information systems research on GenAI as a human-facing technology. IS work on generative AI frames the technology as a source of new social, organizational, and governance questions, not only as a technical production tool (Storey et al., 2025). Related work on human–AI collaboration treats AI systems as practical arrangements in which value depends on how human and machine capabilities are combined and assessed in use (Dellermann et al., 2019; Reinhard et al., 2023); human-centered AI work similarly argues that trustworthiness depends on how people understand and evaluate AI-supported outputs in context (Schoenherr et al., 2023). For equality and equity imagery, this means that inclusive output is only the start. A visually diverse or equality-oriented image may reflect a design goal, but it does not by itself show how viewers interpret the image, how they feel about it, or whether they report any shift in perspective.

This paper studies self-reported perspective-change endorsement in response to AI-generated equality and equity images. We use the phrase deliberately. Attitude research distinguishes immediate response from durable attitude change, and reviews of attitude change stress that persuasion depends on the person, object, message, and context rather than on exposure alone (Crano and Prišlin, 2006; Albarracín and Shavitt, 2018). The repeated survey item in this study asked whether an artwork changed the participant's perspective on any aspect of equality or equity, using a 1-5 agreement

scale. This single-item, face-valid measure is therefore scoped to immediate self-report. We use 'self-reported perspective-change endorsement," 'endorsement," and "1-5 perspective-change rating" as construct labels.

The study analyzed 172 participants who viewed eight DALL-E 3 equality and equity images, producing 1,376 participant-image viewing events. Question-order shuffling was enabled in Google Forms; because respondent-level realized order was not exported, cross-image differences are treated as descriptive rather than causal. The analysis focuses on labeled image items and participant-image response records rather than respondent-level order effects.

The analysis treats the full 1-5 perspective-change rating as the primary inferential outcome. Binary ratings of 4 or 5 are used as descriptive endorsement rates because they are easy to interpret and align with the figures. Co-reported affect flags are treated as same-block affective response signals rather than temporally prior inputs. This wording matters because same-source, same-time perceptual measures can raise common-method concerns in survey research (Chang et al., 2010). Human-coded text themes and stimulus provenance are used to contextualize the response patterns rather than to prove that particular image features produced those responses. The mixed-methods structure follows the broader principle that joint displays can show the link between quantitative patterns and qualitative interpretation (Johnson et al., 2017).

The paper asks three research questions:

- **RQ1:** How do ordinal self-reported perspective-change ratings and binary endorsement rates vary across the eight AI-generated equality and equity images?
- **RQ2:** How are same-block, co-reported affect flags associated with ordinal perspective-change ratings and binary endorsement checks?
- **RQ3:** How do human-coded text themes and stimulus records contextualize the observed image-level response patterns?

The contribution is a bounded evaluation approach for GenAI social imagery. First, the paper reports per-image distributions and endorsement rates with uncertainty rather than treating inclusive generated imagery as uniform. Second, it uses the ordinal item as the primary model outcome while retaining binary endorsement summaries for interpretation. Third, it connects model results to human-coded text themes and a full stimulus record, including model, prompt/description text, generation date, and screening criteria. Fourth, it states the design scope directly: a bounded eight-image stimulus set, shuffled display order, complete repeated outcome data, no delayed follow-up, and an audit trail that separates descriptive patterns, model associations, and causal claims.

2. Related Work

The literature base for this study connects GenAI visual production, synthetic visual representation, affective response, and mixed-methods evaluation. These streams justify treating AI-generated equality and equity images as sociotechnical artifacts whose reception must be measured rather than inferred from generation-side design goals or visual inspection alone.

2.1. GenAI Imagery as Sociotechnical Content

Text-to-image systems make it easy to produce large numbers of visual artifacts for educational, organizational, and public communication use. Work on text-to-image generation describes this low-cost visual production model as a broad content-creation shift rather than only an art-making tool (Oppenlaender, 2022). IS research similarly frames generative AI as a technology with growing social impact and new questions for information systems research (Storey et al., 2025). Work on human-AI collaboration also treats human and machine intelligence as a practical design and evaluation problem rather than a purely technical artifact problem (Dellermann et al., 2019; Reinhard et al., 2023). This matters because the stimuli are generated communication artifacts rather than static pictures. They carry model provenance, participant descriptions, screening choices, and likely use contexts.

For research on GenAI social content, the key issue is how generated imagery can be evaluated in a way that connects artifact production to viewer response, rather than only visual appeal. A generated equality and equity image has at least three sociotechnical layers. First, there is a provenance layer: model, date, input record, screening, and available stimulus records. Second, there is a visual representation layer: the image shows social groups, roles, institutions, bodies, or symbolic scenes. Third, there is a reception layer: viewers report affect, interpretation, and possible perspective-change endorsement. Our research questions map to those layers by asking how responses vary by image, how co-reported affect signals align with ratings, and how text themes contextualize what viewers noticed.

2.2. Bias, Representation, & Synthetic Media

Prior work on text-to-image systems shows why generated imagery should be evaluated rather than assumed to be neutral. Large-scale audits have found that accessible text-to-image generators can amplify demographic stereotypes (Bianchi et al., 2023). Related audits of gender presentation argue that the distributions of generated images remain insufficiently understood for downstream use (Zhang et al., 2023). This line of work studies generated outputs and bias patterns. Our study examines human reception to a bounded set of generated equality and equity images, but it follows the same core premise: generated visual outputs require empirical scrutiny before social use.

This premise also applies when outputs are intended to be inclusive. System-side safeguards may reduce some harmful or exclusionary generations, but a safer or more visibly inclusive output is not the same as evidence of viewer impact. The reception layer still matters because viewers may affirm, ignore, resist, or reinterpret the equality and equity content they see. Synthetic media research shows that generated or altered visual content creates problems for trust, interpretation, and knowledge sharing (Hancock and Bailenson, 2021; Doss et al., 2023). Those concerns also matter for benign uses. Even when generated images are intended for inclusive or educational purposes, viewers may read them in varied ways. Static equality and equity images differ from interpersonal contact; our focus is the immediate perspective-change endorsement participants reported after viewing specific generated images.

2.3. Affect & Reported Perspective Change

Attitude research distinguishes immediate response, attitude formation, and stable attitude change. Prior reviews stress that attitude change depends on the person, object, message, and context; brief exposure should not be assumed to move durable beliefs (Crano and Prišlin, 2006; Albarracín and Shavitt, 2018). That distinction is central here. The repeated survey item asks about a reported change in perspective after an image block. Delayed belief change, learning, and prejudice reduction are reserved for future designs.

Positive affect may broaden attention and thought (Fredrickson, 2004), but our affect measures were selected in the same image block as the outcome. Same-source, same-time perceptual measures can raise common-method concerns in behavioral survey research (Chang et al., 2010). We therefore treat Inspired, Hopeful, Uncomfortable, and Angry as co-reported affect signals rather than causal mechanisms. Prior studies of AI-generated art also show that people respond to AI-made outputs through social and moral judgments, not only visual liking (Lima et al., 2021; Epstein et al., 2020). This supports our cautious use of affect as part of the reception pattern.

2.4. Evaluation of Inclusive GenAI Imagery

Evaluating generated equality and equity imagery requires both numeric patterns and human interpretation. Mixed-methods joint displays can connect quantitative results to coded qualitative evidence and make integration visible to readers (Johnson et al., 2017). In this study, the numeric layer reports per-image distributions, endorsement rates, ordinal models, diagnostics, and sensitivity checks. The human-coded layer uses a stratified text sample to contextualize what participants appeared to notice. The result is a bounded evaluation pattern for inclusive GenAI imagery with four safeguards: stimulus records, ordinal outcomes, same-block affect caution, and human-coded interpretation.

3. Methods

The methods are organized around the evaluation chain: survey delivery, participant flow, stimulus provenance, repeated measures, human coding, and ordinal diagnostics. This structure makes the design auditable from generated image to participant event.

3.1. Study design & survey platform

This study used an anonymous online survey to examine immediate responses to AI-generated equality and equity images. Participants viewed eight image-linked stimulus items and completed repeated image-level response items. Question-order shuffling was enabled in Google Forms. Because respondent-level realized order was not exported, cross-image differences are treated as descriptive rather than causal, and models do not include respondent-level order terms. Analyses use the eight labeled image items and the repeated participant-image response records.

3.2. Recruitment, consent, & retained sample

Participants were adults who self-screened for eligibility before continuing. The survey instructed respondents to continue only if they met the eligibility criteria, including being between 18 and 65 years old and able to consent. Direct student recruitment occurred through course channels during regular semesters at two U.S. universities. To preserve double-blind

review, institution names, protocol identifiers, and state-level institutional details are masked. The analyzed survey data retained the variables needed for participant-image event analysis; inviter-invitee links and referral geography were not collected or recorded.

Student participants could receive course extra credit, and a comparable non-research alternative was available. Some recruitment also occurred through referrals intended to broaden respondent diversity. The referral pathway was used to broaden reach; inviter-invitee links and referral geography were intentionally not retained in the anonymous response file, so recruitment clustering is treated as part of the sample scope rather than a modeled factor. The raw survey export contained 172 submitted responses. The retained sample included 172 survey responses and 1,376 participant-image events. All image-level outcome ratings and affect flags were complete. Numeric age was available for 170 responses (M=23.6, SD=7.3); two note-like entries were retained outside numeric age summaries. The sample included 97 men, 72 women, and fewer than five other or multiselect gender responses. Race/ethnicity and field were reported in broad anonymous buckets. Table A.1 reports anonymous sample characteristics and missingness.

3.3. Ethics and data handling

The study was IRB approved under an exempt human-subjects protocol. For anonymous review, the reviewing institution and protocol identifier are masked. The consent materials described voluntary participation, the ability to stop, confidentiality protections, and the course-credit alternative. The analytic files use generated participant identifiers. The retained analytic files omit survey-response identities, inviter-invitee links, respondent-level display order, formal attention-check responses, response-time data, and clickstream data. Inclusion was based on eligibility self-screening, submitted responses, and complete image-block outcome data.

3.4. Stimuli and provenance

The stimulus set contained eight AI-generated images, labeled A1-A8, covering classroom learning, socio-economic/work-life contrast, glass-ceiling symbolism, cultural unity, disability/access and technology, racial equality protest, health care equity, and shared domestic labor. The project files identify the generation tool as DALL-E 3 and the generation date as April 8, 2024. The prompt structure was "generate an artistic image of [participant-facing description]." The participant-facing questionnaire wording is used as the authoritative stimulus description.

After generation, the researcher visually screened each image to confirm artistic rather than photorealistic rendering, no apparent reproduction of an identifiable real person, and basic theme match. The retained provenance record documents the deployed stimuli, model, date, input record, participant-facing descriptions, and screening criteria; see Table A.2.

3.5. Measures

The repeated image-level outcome was the participant-facing item: *"This artwork change my perspective on any aspects of equality or equity."* Responses used a five-point agreement scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, and 5 = strongly agree. We refer to this construct as *self-reported perspective-change endorsement*. The full 1-5 rating is used for primary ordinal inference, while ratings of 4 or 5 are coded as binary endorsement for descriptive rates and robustness checks.

The single-item format was used because the measure was repeated for eight image-linked items and the survey also asked open-text questions. The brief item supported repeated measurement across all eight images while preserving space for open-text interpretation. The item is therefore treated as a bounded, face-valid immediate self-report; durable attitude change, learning, reflection, and prejudice reduction require future designs with validated scales.

Each image-linked item also included co-reported affect flags and open-text questions. The affect flags were multi-select responses, including Inspired, Hopeful, Uncomfortable, and Angry. Because affect and the outcome rating were reported in the same survey context and respondent-level realized order was not exported, affect variables are treated as co-reported affective response signals rather than causal inputs or temporal antecedents. Open-text fields asked respondents to explain how their perspective changed and to identify equality/equity themes they perceived. These text responses support human-coded interpretation of the quantitative patterns.

3.6. Human coding and adjudication

Human coding was used to add an interpretation layer rather than to infer causal feature effects. The text-coding unit was one nonempty image-linked response field from one of three fields: thoughts/questions, perspective-change explanation, or perceived equality/equity theme. The full canonical text pool contained 3,857 nonempty text units across 1,341 of 1,376 image events. A stratified 240-unit sample was

drawn with 30 units per image, 10 units per text field per image, and a hidden balance of 15 binary-endorsement and 15 non-endorsement units per image.

Two independent coders applied text-theme and image-feature codebooks. Coders were blind to the study purpose and hypotheses. For text coding, they did not see the stimulus images, participant identifiers, numeric perspective-change ratings, or binary endorsement status. Reliability was computed before adjudication using Cohen's κ for binary codes and weighted κ for ordinal image-feature codes where needed. Disagreements were resolved through adjudication, and original reliability values were retained. Table 1 reports the reliability thresholds. Low-reliability codes are used only descriptively. Image-feature codes are treated as contextual evidence; only institutional context and access/disability salience met the cautious-use threshold among image-feature codes, and no image-feature regression is used for inferential claims. In the table, strong = $\kappa \geq .80$; caution = $.60 \leq \kappa < .80$; descriptive only = $\kappa < .60$. Reliability was computed before adjudication. Adjudicated codes contextualize response patterns; low-reliability codes are limited to descriptive context.

3.7. Statistical analysis and diagnostics

The primary inferential outcome is the full 1-5 perspective-change rating. The primary model is an ordered logit model with image fixed effects and participant-clustered robust standard errors. For participant i, image j, and ordinal cutpoint k, the model can be written as:

$$\text{logit}\{\Pr(Y_{ij} > k)\} = \alpha_k + \boldsymbol{\beta}^\top \mathbf{A}_{ij} + \boldsymbol{\gamma}^\top \mathbf{I}_j, \quad (1)$$

for $k \in \{1, 2, 3, 4\}$. Here, Y_{ij} is the 1-5 perspective-change rating, $\mathbf{A}_{ij}$ contains the co-reported affect flags, and $\mathbf{I}_j$ contains image fixed effects. Positive coefficients therefore mean higher odds of being above a given response threshold, not proof of actual attitude change. Odds ratios are interpreted as higher or lower odds of a higher category of the 1-5 rating, not as odds of actual attitude change. Binary endorsement, defined as ratings of 4 or 5, is used for descriptive image-level rates and robustness checks because it is easy to interpret as agreement with the perspective-change item. No model treats the label A1-A8 as the realized display order.

For interpretability, we translated ordinal results into model-implied binary endorsement contrasts:

$$\hat{p}_m(a) = \sum_{r=4}^{5} \widehat{\Pr}\{Y_{ij} = r \mid A_{m,ij} = a, \mathbf{X}_{ij}\},$$
$$\Delta_m^{(4+)} = \hat{p}_m(1) - \hat{p}_m(0), \qquad a \in \{0, 1\}. \quad (2)$$

Here, $\mathbf{X}_{ij}$ denotes the observed image mix and the other co-reported affect flags. The contrast $\Delta_m^{(4+)}$ is a descriptive translation of the ordinal model into the probability of rating 4 or 5; it is not a causal effect estimate.

We used several diagnostics and sensitivity checks to assess the stability of the ordinal results and the interpretation of same-block affect measures. First, we used ordinal GEE as a clustered population-averaged check. Second, we used cutpoint-specific binary logits for Rating > 1, > 2, > 3, and > 4 as a proportional-odds diagnostic, following the logic of proportional-odds assessment and generalized ordinal checks (Brant, 1990; Liu and Koirala, 2012). Third, we added a participant response-tendency sensitivity using each respondent's leave-one-image-out mean rating. Fourth, we summarized predicted probabilities for ratings of 4 or 5 under observed image and affect distributions. Fifth, we examined affect co-occurrence, phi coefficients, variance-inflation factors, and reduced positive- versus negative-affect count models. Sixth, because no formal attention checks or time logs were part of the survey design, we ran an indirect engagement sensitivity excluding events with no open-text response and events with fewer than five open-text words. These checks support interpretation and robustness assessment; they do not change the descriptive, non-causal scope of the study. Because affect flags and ratings are same-block self-reports, we interpret all affect estimates under the common-method caution noted in same-source survey research (Chang et al., 2010).

4. Results

The results proceed in four steps. We first summarize image-level endorsement rates and rating distributions, then report the primary ordinal model, diagnostic checks, and human-coded interpretation. This order separates descriptive response patterns, model-based associations, and the open-text evidence used to contextualize how participants interpreted the images.

4.1. Descriptive Image-Level Endorsement

The analytic dataset contained 172 retained responses and 1,376 participant-image events. All events had a nonmissing 1–5 perspective-change rating. Ratings of 4 or 5 are used as a descriptive endorsement summary. By that criterion, 322 of 1,376 events (23.4%) endorsed perspective change. A3, the glass-ceiling image, had the highest endorsement rate (58/172, 33.7%; 95% Wilson CI 27.1%–41.1%),

Table 1. Human-coding sample, reliability, and analytic use.

Domain	Code family or code	Coded units	Reliability or design feature	Analytic use
Text sample	Nonempty image-linked response fields	240	Stratified by image and field: 30 units per image; 10 per text field per image	Human interpretation layer
Coder masking	Deidentified text units	240	Coders did not see images, participant IDs, ratings, labels, or study purpose	Limits confirmation bias in coding
Text themes	13 binary theme codes	240	κ range .162–.815; agreement .775–.996	Reliability-tiered use: 1 strong, 4 caution, 8 descriptive
Text themes	Gender-equity reflection	240	$\kappa = .815$; agreement=.942	Main interpretive support
Text themes	Racial/cultural, disability/ access, technology/ AI, and personal- experience codes	240 each	$\kappa = .600$–$.665$	Contextual support with caveat
Image features	14 visible feature codes	8 images	κ range .000–.750; agreement .000–.875	Stimulus-set description
Image features	Institutional context and access/disability salience	8 each	$\kappa = .750$; agreement=.875	Context only; cautious due to $n = 8$ images

followed by A4, the cultural-unity image (56/172, 32.6%; 95% CI 26.0%–39.9%). A1, the classroom image, had the lowest endorsement rate (18/172, 10.5%; 95% CI 6.7%–15.9%). Figure 1 shows both the binary endorsement rates and the full 1–5 rating distributions, so the descriptive pattern is not reduced to the 4–5 threshold alone. Table A.3 shows the eight stimulus images, allowing the image-level summaries to be read against the visible artifacts. Because question-order shuffling was enabled but respondent-level realized order was not exported, these cross-image differences are interpreted descriptively instead of causal effects.

4.2. Primary Ordinal Model

The primary inferential analysis used the full 1–5 perspective-change rating. The ordered-logit model included image fixed effects and co-reported affect flags, with participant-clustered robust standard errors. Odds ratios are interpreted as higher or lower odds of a higher 1–5 perspective-change rating, not as evidence of durable belief change. Table 2 reports the primary ordered-logit estimates and the ordinal GEE clustered check. Note that, in the table, outcome is the full 1-5 perspective-change rating. Models include image fixed effects. Affect flags were selected in the same image block as the outcome; estimates are associations with higher or lower 1-5 ratings, not causal effects.

Figure 2 visualizes the primary ordered-logit estimates. In the primary model, Inspired was positive (OR=1.90, 95% CI [1.23, 2.92], $p = .004$), while Uncomfortable was negative (OR=0.61, 95% CI [0.38, 0.98], $p = .043$). Hopeful was positive in point estimate but not clearly detected with participant-clustered robust standard errors (OR=1.43, 95% CI [0.91, 2.24], $p = .123$). Angry was not clearly detected and is not used as an evaluation claim (OR=0.67, 95% CI [0.32, 1.41], $p = .289$). The ordinal GEE check also supported Inspired as positive and found Hopeful positive, while Uncomfortable weakened and Angry was near null.

Table 2. Primary ordinal model and ordinal clustered check for co-reported affect signals

Model	Affect flag	OR [95% CI]	p
Primary ord. logit	Inspired	1.90 [1.23, 2.92]	.004
Primary ord. logit	Hopeful	1.43 [0.91, 2.24]	.123
Primary ord. logit	Uncomfortable	0.61 [0.38, 0.98]	.043
Primary ord. logit	Angry	0.67 [0.32, 1.41]	.289
Ordinal GEE check	Inspired	1.46 [1.19, 1.80]	$<.001$
Ordinal GEE check	Hopeful	1.36 [1.08, 1.70]	.008
Ordinal GEE check	Uncomfortable	0.87 [0.67, 1.11]	.260
Ordinal GEE check	Angry	1.00 [0.71, 1.43]	.980

4.3. Diagnostics, Predicted Probabilities, and Same-Block Checks

The model diagnostics support a bounded reading of the affect results. Cutpoint-specific binary logits, used as a proportional-odds diagnostic, kept Inspired positive across all four thresholds and kept Uncomfortable negative across all four thresholds. A participant response-tendency sensitivity also preserved the main Inspired signal. Predicted probabilities make the effect size easier to read: holding the observed image mix and other affect flags at their observed distributions, the predicted probability of a 4-5 endorsement rating was 18.5% when Inspired was absent and 29.7% when Inspired was present. The corresponding predicted probability was 26.1% when Uncomfortable was absent and 17.9% when Uncomfortable was present.

The multi-select affect flags did not show problematic collinearity. The largest variance-inflation factor was 2.02. Inspired and Hopeful were nearly independent (phi=0.02), while Inspired and

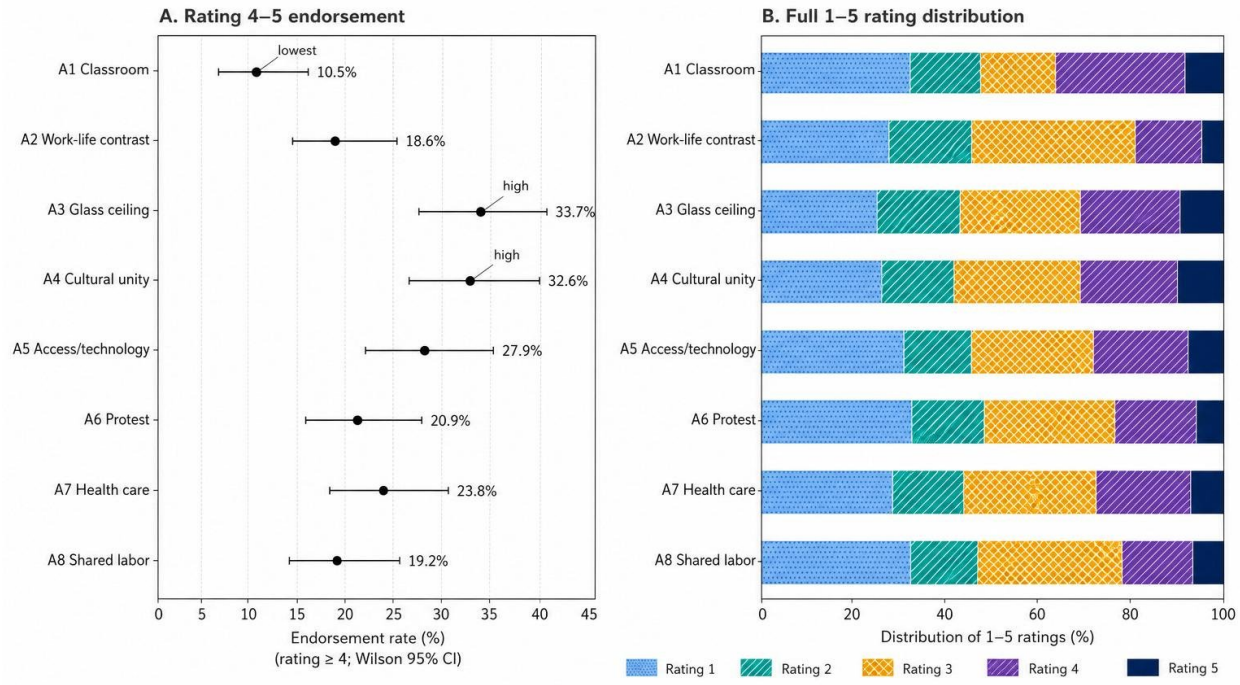


Figure 1. Per-image self-reported perspective-change endorsement and full rating distributions. Endorsement is a rating of 4 or 5 on the repeated 1-5 item. Question-order shuffling was enabled in Google Forms, but respondent-level realized order was not exported; cross-image differences are descriptive and not causal.

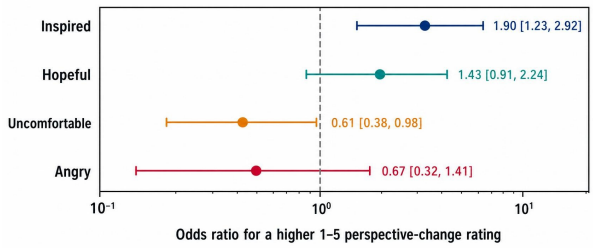


Figure 2. Primary ordinal estimates for co-reported affect signals. Odds ratios reflect higher 1–5 perspective-change ratings; OR=1 marks no association. Models use participant-clustered robust standard errors and image fixed effects.

Uncomfortable were negatively related (phi=-0.53). A reduced model using positive and negative affect counts gave the same broad pattern: positive affect count was associated with higher 1-5 ratings, while negative affect count was associated with lower 1-5 ratings. As an indirect engagement sensitivity, we refit the primary model after excluding events with no open-text response and after excluding events with fewer than five open-text words. The Inspired and Uncomfortable estimates were stable across these filters. These checks assess stability under alternative coding and engagement filters while preserving the study's descriptive, non-causal scope. Table A.4 summarizes these diagnostics.

4.4. Human-Coded Interpretation of Image Responses

The human-coded layer used a stratified 240-unit sample from the nonempty image-linked text fields. The sample included 30 coded units per image and was balanced across the three text fields and endorsement status. Coders were blind to the study purpose and hypotheses, did not see the stimulus images, and were not shown participant identifiers, numeric ratings, or binary endorsement labels. Table 3 is the main mixed-methods bridge: it links descriptive endorsement rates to coded text-theme signals and paraphrased examples. A3 had the highest endorsement rate and the strongest coded gender-equity signal in the sample. A4 had a high endorsement rate and a racial/cultural equity signal. A5 and A7 were linked to disability/access themes, while A8 linked strongly to gender-role interpretation even though its endorsement rate was lower than A3 and A4. Note, in the table, endorsement means a rating of 4 or 5 on the self-reported perspective-change item. Text themes come from a stratified 240-unit coded sample, with

30 coded units per image. Low-reliability codes are retained only as descriptive context.

Image-feature coding is used only as context. Among image-feature codes, institutional context and access/disability salience met the cautious-use threshold. Other feature labels remain useful for describing the bounded stimulus set, while feature-effect claims are left to future designs with larger stimulus samples. The eight images are therefore treated as a bounded case set of DALL-E 3 equality and equity imagery, with the visible stimulus set shown in Table A.3.

5. Discussion

The discussion translates the empirical results into a contribution for research on GenAI social systems: inclusive output should be evaluated at the point of human reception, not only at the point of generation. The section explains what the study shows, why GenAI provenance matters for equality and equity imagery, and how the evidence record can support next-stage studies with logged order, control images, validated scales, and delayed follow-up.

5.1. Main Finding

This paper gives a bounded account of how participants responded to eight DALL-E 3 equity images in a platform-shuffled survey where the exported data did not retain respondent-level display sequences. The main result is a disciplined reception pattern: descriptive image-level endorsement rates varied, and same-block co-reported affect signals were associated with the full 1–5 perspective-change rating. Inspired was the most consistent positive co-reported affect signal. Uncomfortable was negative in the primary ordinal model but weaker in clustered checks. Hopeful was mixed, and Angry was model-dependent. This pattern is useful because it shows that generated equity imagery should not be treated as interchangeable social content.

5.2. Why GenAI Matters Here

The GenAI setting matters because the images are not merely visual stimuli; they are synthetic communication artifacts produced by a system that can generate inclusive imagery quickly, repeatedly, and at low cost. In settings where image-generation tools are used for education, training, public communication, or organizational messaging, equality and equity images may be produced as part of a broader effort to avoid exclusionary representation and foreground visible inclusion. That production goal is important, but it does not establish audience response.

Our findings show the value of pairing inclusive output with reception evidence. The same stimulus set produced varied endorsement rates, different affective response signals, and distinct human-coded interpretations. This means that generated equality and equity imagery should be evaluated with viewer response data, uncertainty estimates, human-coded interpretation, and a clear record of how the stimuli were produced. The practical value of the paper is therefore an evaluation pattern for testing generated social imagery before relying on it in human-facing settings, rather than a ranking of which image is best.

5.3. Evaluation Guidance for GenAI Social Imagery

The findings support evaluation guidance rather than content-design rules. First, report per-image response distributions and intervals rather than only aggregate rates. Second, use the full ordinal response item for inference when possible, and reserve binary endorsement rates for interpretation. Third, measure affect in the same block only as a co-reported response signal unless the design establishes temporal order. Fourth, use human coding when images concern equality, equity, identity, or access. Fifth, preserve the image files, model/tool, generation date, input record, participant-facing description, and screening criteria. Sixth, when order effects are part of the research question, log randomized display order.

This evaluation pattern helps separate artifact production, viewer response, and interpretation. The artifact record clarifies what was generated. The event-level data show how each artifact was received. The ordinal model preserves the response scale. The affect diagnostics reduce overreading of same-block affect. The coding layer keeps sensitive interpretation in human judgment rather than automated sentiment labels. Together, these elements separate descriptive patterns, model associations, and future-study hypotheses.

5.4. Implications for IS and Human-AI Interaction Research

The study also suggests a clearer role for IS research on GenAI content. Many GenAI studies focus on model capability, output quality, or system adoption. This paper shifts the unit of concern to the generated social artifact as used in a human setting. The artifact is not only a model output; it is a designed communication object with provenance, prompt history, visual content, platform delivery, and audience interpretation. That framing is central for IS work since organizations

Table 3. Mixed-methods joint display linking descriptive image response rates to human-coded text themes.

Image	Endorsement	Human-coded theme signals	Paraphrased interpretation example
A1	18/172 (10.5%); CI 6.7%-15.9%	Technology/AI 6/30 (caution); gender equity 4/30 (strong); other themes descriptive	Noted diversity in the classroom while also noticing uneven access to technology.
A2	32/172 (18.6%); CI 13.5%-25.1%	Structural barriers 17/30 (descriptive); social role 12/30 (descriptive); disability/access 3/30 (caution)	Linked the contrast to unequal access to resources, rest, and stability.
A3	58/172 (33.7%); CI 27.1%-41.1%	Gender equity 26/30 (strong); social role and structural barriers 13/30 each (descriptive)	Read the scene as women breaking barriers and seeking equal opportunity.
A4	56/172 (32.6%); CI 26.0%-39.9%	Racial/cultural equity 16/30 (caution); prior-view reinforcement 6/30 (descriptive)	Emphasized different cultures gathering together in a positive shared space.
A5	48/172 (27.9%); CI 21.7%-35.0%	Disability/access 17/30 (caution); technology/AI 15/30 (caution)	Connected technology with accommodations, inclusion, and support for disabled people.
A6	36/172 (20.9%); CI 15.5%-27.6%	Racial/cultural equity 8/30 (caution); institutional fairness 7/30 (descriptive)	Focused on racial equality, solidarity, and collective public action.
A7	41/172 (23.8%); CI 18.1%-30.7%	Disability/access 18/30 (caution); institutional fairness 19/30 (descriptive)	Linked the image to equal access and personalized health care.
A8	33/172 (19.2%); CI 14.0%-25.7%	Gender equity 12/30 (strong); social role 21/30 (descriptive)	Reflected on gender roles and shared domestic labor.

rarely deploy isolated models. They deploy workflows, policies, and messages that people must interpret.

A second implication is that GenAI evaluation should distinguish artifact intent from artifact reception. A generation process may aim for inclusive output, a screening process may verify non-photorealistic rendering and theme match, and a platform may present the image cleanly. None of those steps establishes how viewers will read the image. In this study, the same bounded stimulus set produced different endorsement rates and different coded interpretations. That pattern supports a sociotechnical view in which generated social content should be evaluated through both system-side documentation and viewer-side response data.

5.5. Human Coding and Sensitive Interpretation

The human-coded layer is useful because the text responses show that similar endorsement rates can reflect different interpretations. A3 was mainly read through gender-equity themes, A4 through racial/cultural unity, A5 and A7 through disability/access and institutional care, and A8 through gender roles and domestic labor. These coded patterns support interpretation while reserving feature-effect claims for larger image sets. Low-reliability image-feature codes are limited to descriptive context.

The coder process also matters. Coders did not see the stimulus images, participant identifiers, numeric ratings, or binary endorsement status, and they were blind to the study purpose and hypotheses. This blinding strengthens the coding layer by separating text interpretation from stimulus viewing and outcome ratings. Reporting reliability thresholds and limiting low-reliability codes to descriptive use makes the mixed-methods layer more auditable.

6. Scope Conditions and Future Work

This study is intentionally scoped as a bounded evaluation of immediate responses to eight AI-generated equality and equity images. Question-order shuffling was enabled, while respondent-level display sequences were not exported; image-level comparisons are therefore interpreted as descriptive response patterns. The stimulus set, repeated single-item outcome, same-block affect flags, course-anchored and referral sample, and stratified 240-unit text sample support the study's main goal: a transparent early-stage evaluation of inclusive GenAI imagery and human reception.

This scope also defines clear next-stage designs. Future studies can compare generated, human-made, neutral, and no-image conditions; log realized display order; use validated multi-item scales; add attention and timing measures; test larger image corpora; retain rejected generations as part of the stimulus record; add delayed follow-up; and pre-register confirmatory comparisons. These extensions would move from bounded reception evaluation toward stronger tests of durable attitude change, GenAI-specific effects, image-feature causality, affect timing, and population-level moderation.

7. Conclusion

This study began from a practical question: when AI systems produce visibly inclusive equality and equity imagery, how do viewers actually respond? The answer is not captured by generation-side intent, the output alone, or a single aggregate mean. Across 1,376 viewing events, participants endorsed perspective change at different rates across the eight stimuli, and their same-block affect reports helped clarify those response patterns. Inspired was the clearest positive affect signal. Uncomfortable pointed in the opposite

direction in the primary ordinal model, while clustered checks were more modest. The human-coded text layer showed that participants were not merely rating visual appeal; they interpreted the images through social themes such as gender equality, cultural inclusion, access, disability, technology, and labor.

The broader contribution is a disciplined way to test inclusive GenAI imagery before it is used in organizational, educational, or public communication settings. The approach keeps the stimulus record visible, treats ordinal self-report as ordinal data, reports image-level variation with uncertainty, adds human interpretation where sensitive themes are involved, and keeps the scope of inference explicit. The key lesson is practical and methodological: inclusive GenAI content should be evaluated as a sociotechnical artifact whose human reception is measured, not assumed.

Acknowledgments

[blind]

References

Albarracín, D. and Shavitt, S. (2018). Attitudes and attitude change. *Annual Review of Psychology*, 69(1):299–327.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., and Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1493–1504. ACM.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46(4):1171–1178.

Chang, S.-J., van Witteloostuijn, A., and Eden, L. (2010). From the editors: Common method variance in international business research. *Journal of International Business Studies*, 41(2):178–184.

Crano, W. D. and Prišlin, R. (2006). Attitudes and persuasion. *Annual Review of Psychology*, 57(1):345–374.

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., and Ebel, P. (2019). The future of human-ai collaboration: A taxonomy of design knowledge for hybrid intelligence systems. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences.

Doss, C., Mondschein, J., and Shu, D. (2023). Deepfakes and scientific knowledge dissemination. *Scientific Reports*, 13(1).

Epstein, Z., Levine, S., Rand, D. G., and Rahwan, I. (2020). Who gets credit for ai-generated art? *iScience*, 23(9):101515.

Fredrickson, B. L. (2004). The broaden-and-build theory of positive emotions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1449):1367–1377.

Hancock, J. T. and Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3):149–152.

Johnson, R., Grove, A., and Clarke, A. (2017). Pillar integration process: A joint display technique to integrate data in mixed methods research. *Journal of Mixed Methods Research*, 13(3):301–320.

Lima, G., Zhunis, A., Manovich, L., and Cha, M. (2021). On the social-relational moral standing of ai: An empirical study using ai-generated art. *Frontiers in Robotics and AI*, 8:719944.

Liu, X. and Koirala, H. P. (2012). Ordinal regression analysis: Using generalized ordinal logistic regression models to estimate educational data. *Journal of Modern Applied Statistical Methods*, 11(1):242–254.

Oppenlaender, J. (2022). The creativity of text-to-image generation. In *Proceedings of the 25th International Academic Mindtrek Conference*, pages 192–202. ACM.

Reinhard, P., Li, M. M., and Peters, C. (2023). Introduction to the minitrack on applications of human-ai collaboration: Insights from theory and practice. In *Proceedings of the 56th Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences.

Schoenherr, J. R., Abbas, R., Michael, K., Rivas, P., and Anderson, T. D. (2023). Designing ai using a human-centered approach: Explainability and accuracy toward trustworthiness. *IEEE Transactions on Technology and Society*, 4(1):9–23.

Storey, V. C., Yue, W. T., and Zhao, J. L. (2025). Generative artificial intelligence: Evolving technology, growing societal impact, and opportunities for information systems research. *Information Systems Frontiers*, 27(5):2081–2102.

Zhang, Y., Jiang, L., Turk, G., and Yang, D. (2023). Auditing gender presentation differences in text-to-image models.

Appendix

Table A.1. Anonymous sample characteristics and missingness summary

Domain	Category or field	Count	Summary
Retained sample	Survey responses	172	100% retained
Event file	Participant-image events	1,376	172 participants $\times$ 8 images
Age	Valid numeric entries	170	Mean=23.6, SD=7.3; median=21, IQR=20-23
Age	Note-like age-field entries	2	Retained outside numeric age summaries
Gender	Man / woman	97 / 72	Other or multiselect suppressed because $n < 5$
Race/ethnicity	Asian or Asian American only / White or European only / other or multiracial	60 / 68 / 44	Broad anonymous buckets
Education	Some college / high school or GED / bachelor's / master's	88 / 34 / 26 / 21	Other or prefer-not suppressed because $n < 5$
Field	CS / Data Science / other	61 / 32 / 79	Broad field buckets
Collection year	2024 / 2025 / 2026	90 / 75 / 7	Direct recruitment institution masked for review
Outcome missingness	1-5 perspective-change rating	0 / 1,376	Complete for all image events
Affect missingness	Four co-reported affect flags	0 / 1,376	Derived for all image events
Open-text missingness	No nonempty image-linked text field	38 / 1,376	Used as an indirect engagement check

Table A.2. Stimulus provenance summary

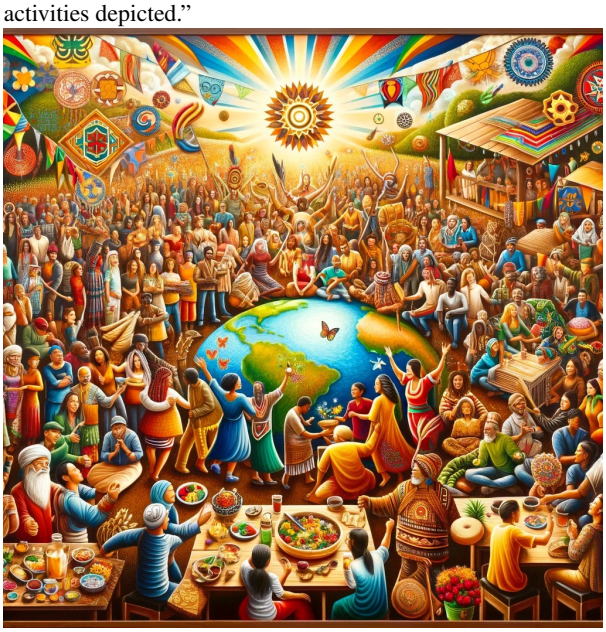
Item	Record used in revision
Stimuli	Eight AI-generated equality and equity images, A1-A8
Model/tool	DALL-E 3
Generation date	April 8, 2024
Input record	"generate an artistic image of [participant-facing description]"
Screening	Artistic/non-photorealistic rendering, no apparent real-person identity, prompt-theme match
Provenance scope	Deployed stimuli, model/date, input record, screening criteria; regeneration counts outside retained record
Participant-facing text	questionnaire wording used as authoritative description

Table A.3. Stimuli details. Images generated with DALLE-3 on April 8, 2024.

A1: "This artwork depicts a diverse group of children learning together in a brightly lit classroom. Notice the range of emotions and engagement across the students."

A2: "This piece presents contrasting scenes of individuals from various socio-economic backgrounds, focusing on their work-life balance. Observe the differences in environments and activities depicted."

A3: "This artwork captures a pivotal moment for a woman in a corporate setting, symbolizing the breakthrough of the glass ceiling. Pay attention to the expressions and symbolic elements that hint at both challenges and achievements."



A4: "Here, we see a vibrant gathering of people from various cultures, coming together in celebration and unity. Notice the variety of cultural symbols and the expressions of joy and cooperation."

Table A.3. Stimuli details (continued). Images generated with DALLE-3 on April 8, 2024.

A5: "This piece imagines a world where technology enhances independence and empowerment for individuals with disabilities. Look for the innovative solutions and expressions of freedom and capability."

A6: "This artwork focuses on a peaceful protest for racial equality, highlighting the solidarity among participants. Notice the diversity within the crowd and the symbols of peace and equality."

A7: "In this scene, individuals of different ages, races, and backgrounds receive personalized healthcare. Observe the approaches and the sense of care and attention given to each person."

A8: "This artwork presents a modern household where domestic responsibilities are shared among partners, challenging traditional gender roles. Focus on the dynamics and interactions that reflect partnership and equality."

Table A.4. Diagnostics, interpretability checks, and engagement sensitivity for the ordinal results.

Check	What was tested	Key result	Analytic interpretation
Cutpoint-specific logits	Four threshold models for Rating > 1, > 2, > 3, and > 4.	Inspired OR range: 1.72-2.04; Uncomfortable OR range: 0.29-0.64.	Inspired was positive at all cutpoints. Uncomfortable was negative at all cutpoints, although the size varied.
Participant response-tendency sensitivity	Added each respondent's leave-one-image-out mean rating.	Inspired OR=1.52 [1.15, 2.01]; Uncomfortable OR=0.77 [0.54, 1.08].	The Inspired result is not only a general tendency by some respondents to rate all images higher.
Predicted endorsement probability	Translated ordinal estimates into predicted probability of ratings 4-5 under observed image and affect distributions.	Inspired: 18.5% to 29.7% (+11.2 points). Uncomfortable: 26.1% to 17.9% (-8.2 points).	The odds-ratio pattern is visible in more interpretable probability terms.
Affect co-occurrence and collinearity	Checked multi-select affect overlap with phi coefficients and VIFs.	Maximum affect VIF=2.02; Inspired+Hopeful phi=0.02; Inspired+Uncomfortable phi=-0.53.	Affect flags were not too collinear for the main associative read.
Reduced affect-count model	Replaced four flags with positive and negative affect counts.	Positive count OR=1.64 [1.12, 2.38]; negative count OR=0.62 [0.42, 0.91].	The same broad positive-versus-negative affect pattern appears under a simpler coding.
Open-text engagement sensitivity	Refit the primary model after excluding events with no open text and events with fewer than five open-text words.	Inspired OR range: 1.78-1.82; Uncomfortable OR range: 0.56-0.59.	Results were similar under indirect engagement filters. This is not an attention check.