Artists' Strategies to Circumvent Representational Biases and Censorship in Text-to-Image Generation

Mijntje de Groote
University of Amsterdam
mijntjemadelief@gmail.com

Piera Riccio
University of Amsterdam
p.riccio@uva.nl

Abstract

As generative Text-to-Image (T2I) models become prevalent in creative workflows, users increasingly encounter systemic representational biases. This study investigates the lived experiences of artists interacting with these systems, focusing on how they navigate and circumvent these biases when attempting to depict complex human identities. We employ a qualitative, user-centered methodology with 10 artists. The sessions included using Stable Diffusion 3, integrating pre-task questionnaires, a Think-Aloud protocol, and post-task interviews. The findings highlight prominent representational challenges in the considered tool, including persistent stereotyping, commercialized perfection, hyper-sexualization and a white and Western default. To overcome these limitations, various mitigation strategies were employed by the interviewed artists, such as using "negative constraints", linguistic workarounds, self-censorship, prompt specification and extensive trial and error. Notably, the study identifies the emergence of "secondary biases", where the mitigation of one stereotype triggers a new one. Ultimately, this research demonstrates that circumventing representational biases demands significant added labor from the user, highlighting how the biased assumptions of models act as barriers to, rather than facilitators of, human-AI co-creativity.

Introduction

The use of generative AI has grown significantly in recent years, as shown by the widespread adoption of commercial text-generative tools, such as ChatGPT, which around 10 percent of the world's adult population uses (Chatterji et al. 2025). Text-to-Image (T2I) generation models like DALL-E (Ramesh et al. 2021) or Stable Diffusion (Rombach et al. 2022) are also used to generate a large number of images daily from simple textual prompts. These tools can be easily accessed by a wide audience, allowing the generation of highly realistic or imaginative outputs (Epstein et al. 2023). This growth is mirrored by an increase in academic research, analyzing the societal impacts of these new models (Luccioni et al. 2023; Naik and Nushi 2023; Riccio, Curto, and Oliver 2024).

From a human–AI co-creativity perspective, AI systems are often conceptualized not merely as passive tools (Rezwana and Maher 2023; Fang et al. 2025), but as collaborators along a spectrum of autonomy and engagement (Singh et al. 2025). These observations highlight that AI can influence outcomes in ways that require humans to adapt, negotiate, or rethink their creative intentions (Taylor et al. 2025). The use of T2I generation tools indeed does not come without challenges, including the amplification of representational biases, the automatic generation of fake news or abusive content, coupled with concrete threats to privacy, to data protection, to intellectual property rights, and to the environment (Solaiman et al. 2023), as well as to visual culture as a whole (Jääskeläinen and Åsberg 2024). Among the varied challenges on which users might have to rethink their creative intentions, we focus specifically on representational biases, analyzing strategies that artists[1] employ to navigate and circumvent them, providing insights on this nuanced discourse that quantitative audits might miss (Turchi et al. 2023).

In particular, we consider Don Ihde's post-phenomenological philosophy of technology (Ihde 1990) to interpret these interactions conceptually. Ihde distinguishes between *embodiment* relations, where a technology functions as an extension of human capabilities, and *alterity* relations, where a technology is experienced as a quasi-"other" with its own agency and perspective. T2I generation can be understood through both lenses (Riccio 2025): in an *embodiment* relation, the model acts as a tool that amplifies the artist's creative intent; in an *alterity* relation, the model can surprise, resist, or shape the artist's intentions, reflecting its internalized biases. In this case, the T2I model is not framed as a neutral instrument, but as a system that can "disobey" (Ihde 1990) and whose behavior must be negotiated.

Following this idea, we investigate how artists circumvent censorship and representational biases of T2I models when attempting to represent complex or marginalized human identities. Through a user-study including 10 artist participants, we examine the representational challenges they identify in model outputs and the mitigation strategies they develop in their prompts to overcome these limitations.

Our study shows that artists identify representational biases in T2I outputs, including stereotyping, hyper-sexualization, and Western or white-centric defaults. Rather than passively accepting these outputs, artists engage in iter-

[1]Note that, in this paper, we use the terms "creative users" and "artists" interchangeably.

ative prompting, linguistic workarounds, and trial-and-error exploration, turning the process into a form of negotiation. At the same time, attempts to mitigate these biases often activate secondary forms of stereotyping, suggesting that representational biases in generative models are intersectional rather than isolated phenomena. By centering human experience, our work complements system-focused audits, revealing how biases are navigated in practice during creative use.

Related Work

A growing body of research documents representational biases in T2I models, showing systematic distortions or omissions in how people are depicted based on attributes such as gender (Mandal, Leavy, and Little 2023), skin tone (Cho, Zala, and Bansal 2023), sexual orientation (Wang et al. 2023), or religion (de Almeida and Rafael 2024). For example, gender-neutral prompts often produce outputs in which high-status roles (*e.g.,* CEO) are predominantly associated with men or lighter-skinned individuals (Luccioni et al. 2023), or reinforce masculinity as a generation default (Wu, Nakashima, and Garcia 2024). These biases reflect imbalances in training data and can shape cultural assumptions at scale (Benjamin 2023). Early audit studies focused on predefined sensitive attributes, while more recent approaches, such as OpenBias (D'Incà et al. 2024) and StOP (Dehdashtian, Sreekumar, and Boddeti 2025), use open-ended methods to uncover latent associations and reveal the technical origins of representational distortions.

Recent work has further demonstrated how these representational biases intersect with broader social inequalities. Studies show that T2I systems frequently default to men in professional contexts (Gorska and Jemielniak 2023) and rely on rigid, binary views of gender (Collyer-Hoar et al. 2025). Furthermore, AI imagery often exhibits presentational biases, such as depicting women in submissive poses (Sun et al. 2024), while heavily favoring western visual defaults (de Almeida and Rafael 2024) and stereotyped portrayals of race (Jääskeläinen and Åsberg 2024).

Beyond these more traditional forms of bias, a recent audit highlighted "invisibility biases", where certain human identities are effectively erased from AI-generated outputs (Riccio, Curto, and Oliver 2024). For instance, prompts depicting stigmatized experiences such as breastfeeding or anorexia are frequently blocked by commercial T2I platforms, even when non-harmful. These findings show that content moderation and safety mechanisms can unintentionally suppress legitimate representations, reinforcing taboos or normative assumptions and creating gaps in the AI-generated cultural landscape. Much of the literature on T2I moderation, however, has emphasized the importance of technical safeguards, investigating red-teaming practices (Ganguli et al. 2022), human-in-the-loop oversight (Kirk et al. 2023), dataset curation, retraining strategies (Gandikota et al. 2023; 2024; Schramowski et al. 2023), and the vulnerabilities of safety filters (Rando et al. 2022). While these studies are fundamental to understand the limitations and the potential risks of these systems, they approach moderation from a security perspective and do not consider the boundary between content moderation and censorship in cultural and creative production (Riccio, Hofmann, and Oliver 2024).

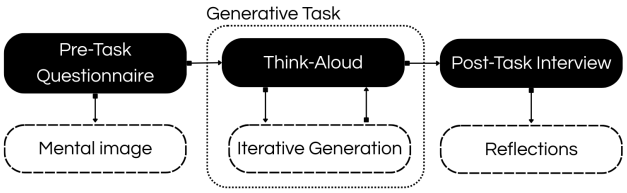

Figure 1: Summary of the methodological approach of this study, highlighting pre-task interviews in which participants create a mental image of what they want to represent; a generative task where the Think-Aloud protocol is adopted and post-task interviews to gather reflections on the experience.

To the best of our knowledge, most of the existing work on the topic of representational biases in T2I generation remains system-centered, analyzing outputs without considering human responses. An exception is Taylor et al. (Taylor et al. 2025), who examine how queer artists recognize normative and anti-sexual values embedded in T2I systems. Building on this approach, our study adopts a broader perspective: rather than focusing on a single community, we investigate how artists more generally navigate representational biases and content moderation. We hence complement existing literature by centering on the human experience, which purely quantitative or system-focused studies cannot reveal.

Method

Our study employs a qualitative, user-centered design to investigate how artists navigate and circumvent representational biases in T2I generative models. The study is centered around the task of interacting with a T2I system to generate representations of human identities that are likely to be stereotyped or censored, providing an opportunity to observe how they adapt their prompts to guide the model. In particular, the study also includes a pre-task questionnaire and a post-task interview, as summarized in Figure 1.

Before the generative task, participants are asked to formulate their own scenarios based on their personal and professional experiences, creating a "mental image" that they want to explore in the generation process. This user-driven approach is adapted from established algorithm auditing methodologies where participants are tasked with generating novel examples to identify problematic outputs or instances that may yield discriminatory statements (Fan et al. 2025). This method allows for the discovery of more nuances in the behavior of the studied system.

Participants

This study includes 10 participants, all artists or creative practitioners, as detailed in Table 1, who were at least 18 years old and able to communicate in English or Dutch, which is the native language of the author conducting the interviews. Participants represented a range of creative fields,

self-identified genders (6 women and 4 men), and seven nationalities spanning Northern and Southern Europe, Central, East, and Southeast Asia, Oceania, and Latin America. This diversity of cultural and professional backgrounds provides varied perspectives.

Data Collection

The study consists of one-on-one sessions with each participant, conducted in person or via an online call by the first author. The study underwent the ethics review procedure for Bachelor's thesis research at the University of Amsterdam and all participants have signed their written consent to take part in this study after being informed about the subject of the research. The pre-task questionnaire was administered via Qualtrics. The T2I model selected for this study was Stable Diffusion 3 (SD3), accessed via a free and user-friendly third-party platform[2]. This platform operates with a non-iterative generation mechanism. When a user modifies the text prompt, the system generates a new and distinct image result without building upon the visual components of the preceding output. This means that the participants of the study are unable to gradually sculpt an image through a sequence of minor adjustments and they have to develop adaptive and strategic language with each new attempt. This is a specific design choice of our study, to clearly reveal the workarounds chosen by the users to negotiate their creative ideas with the system.

Pre-Task Questionnaire The initial phase is designed to capture the participants' mental image and expectations before interacting with the T2I model (Ko et al. 2023). This mental image serves as the baseline against which the models output can be compared (Barve et al. 2025). The questions included in this phase are listed in Table 2.

Think-Aloud While interviews provide a powerful way for reflection, they remain susceptible to recall bias, where participants may forget their immediate, gut-level reactions or subsequently rationalize their behavior. To capture the in-the-moment processes during the task, the Think-Aloud protocol is adopted, suggesting participants to verbalize their thoughts while actively using the T2I model (Nielsen, Clemmensen, and Yssing 2002). This method has been effectively deployed in Human-Computer Interaction (HCI) research to study interactions with AI (Chen et al. 2023) and it requires participants to speak out about their thoughts, feelings and reasoning while they perform the task. Live detailed notes are taken of their thoughts and frustrations, while no audio is recorded. Participants decide when to end the task based on their satisfaction with the visual output or when they feel they have exhausted all their strategies. While they are free to stop at any time, a maximum limit of ten generated images is set to ensure consistency among the participants, and to match the capacity of the utilized interface.

Post-Task Interview The final phase consists of a post-task interview conducted immediately after the creation tasks. This phase provides participants the opportunity to reflect on the generated outcomes and their overall user experience.

[2]https://stablediffusion3.net/en/app, Last Access: 07.03.2026

(a) *Prompt: "A realistic Japanese girl in a kimono".*

(b) *Prompt: "A girl in Japanese traditional ethnic costume".*

Figure 2: Comparison showing the stereotype (left) and the successful linguistic workaround (right) used by Participant P01.

Data analysis

Detailed logs of the prompts and generated images are saved, allowing for later categorization into representational problems and mitigation strategies. The representational problems refer to the moments where the AI produced stereotyped or censored outputs. The mitigation strategies describe the creative workarounds the artists utilize.

Results

In this section we present the representational issues identified by the participants in our study and their mitigation strategies. A brief summary of the representations attempted by the participants is provided in Table 1.

Representational Issues

Stereotyping A primary issue identified by the participants is related to stereotyping and visual homogenization, where the model associates specific keywords to certain visual styles. Due to this, generated images are reduced to a uniform standard. For instance, P01, while trying to represent a Japanese girl wearing a kimono, observed that the word "kimono" triggered the images to be generated in anime style, even though they specifically asked for the image to be realistic, as shown in Figure 2a. P02, on the other hand, trying to prompt the model to generate a person with mental health issues, always obtained the same generic representation, despite attempting multiple times. The representation was, in this case, defaulting to a lonely and sad person, ignoring the complexity and possible invisibility of mental health issues (Riccio, Curto, and Oliver 2024).

Commercialized Perfection and Hyper-sexualization When P04 utilized neutral professional prompts such as "AI researcher", they noticed that the model consistently produced images of conventionally "attractive, polished young

Table 1: Participant Demographics and Field of Creative Practice.

Participant	Gender	Nationality	Field of Creative Practice	Attempted representation
P00	Male	Pakistani	Film	A man in the desert with a camel
P01	Female	Singaporean	Multidisciplinary arts	A Japanese girl wearing a kimono
P02	Female	Dutch	Museum	A person with mental health issues
P03	Male	Italian	Film	A southern Italian landscape
P04	Female	Italian	Gaming industry	A female AI research expert
P05	Female	New Zealander	Visual art	A Maori woman
P06	Female	Chinese	Creative writing	Chinese men placing a small baiju cup in front of a tombstone
P07	Female	Italian	Design	A cafe owner
P08	Male	Dutch	Creative technology	People kissing
P09	Male	Argentinian	Art	A gay man

Table 2: Questions of the Pre-Task Questionnaire.

#	Question
1	Could you briefly describe your artistic practice and your typical creative tasks?
2	On a 7-point Likert scale (1 = not familiar, 7 = extremely familiar), how familiar are you with generative AI?
3	Please describe a scenario, concept, or identity, perhaps from your own cultural background or creative practice, that you expect a generative AI model would either misrepresent, stereotype, or censor.
4	Before we generate this, what is the 'ideal' image you have in your head? Can you describe what you hope to see?
5	What do you expect the model will produce instead? What specific stereotype or form of censorship do you anticipate?

Prompt (a): "A female AI research expert"

Prompt (b): "A neutral looking Japanese female AI research expert focused at her job, typing at her laptop in a corporate tech open space, wearing casual clothes, working environment, office, messy desk, sunny day."

Figure 3: P04: Specificity used to bypass white default.

white woman, who appeared to be posing for an advertisement" (Figure 3). The participant felt that the model "prioritized beauty over reality". Similarly, P07 tried to represent a cafe owner, but they noted that the model tended to generate images that looked promotional, "with a lot of smiling and pretty faces", rather than reflecting the reality of cafe owners, who are "often tired, not smiling" and might not look as the person depicted in Figure 4.

Parallel to this, a few participants encountered a layer of hyper-sexualization. P05 tried to depict a Maori woman and noticed that the word "Maori" led to some images with naked shoulders and outfits that barely covered the body, as shown in Figure 5. P08 prompted the system to generate different people kissing and observed that the model defaulted to nudity and "magazine-style" bodies for gay men (Figure 6a), in direct contrast to the results of "straight people kissing", which P08 observed were more "traditional and clothed". This comparison is shown in Figure 6.

Figure 4: Posed cafe owner (P07)

Figure 5: Hyper-sexualized Maori woman (P05)

P09 also tried to generate a depiction of "a gay man" and confirmed a similar finding as P09: they observed that the model defaulted to nakedness and muscularity as main defining characteristics, as shown in Figure 7a. When prompting for non-muscular and diverse body types, the model found a way to increase the nudity of the images (Figure 7). This suggests the association between queer identity and hyper-sexualization is strong enough to override direct user commands.

(a) *Prompt: "Two gay males kissing".*

(b) *Prompt: "Two straight people kissing"*

Figure 6: P08 comparison: Hyper-sexualization of LGBTQ+ affection versus heteronormative affection.

White and Western Default The experience of P04 when trying to represent a female AI researcher suggested that the model has a strong bias to white people and western standards for professional roles. Despite using neutral or constraining prompts, the model consistently generated images of white Caucasian women with glasses for the role of an "AI research expert". P06, who utilized Chinese prompts to generate specific Chinese characters, mentioned how the model disregarded the context and generated the hair of the Chinese men in a westernized style (Figure 8). P07 observed that the neutral term "cafe owner" consistently produced images of a white person. Similarly, P09 noted a similar bias toward whiteness, when prompting for a "gay man".

Secondary Bias during mitigation A significant finding of this research was that when participants tried to circumvent the biases of the model, they sometimes encountered a second layer of stereotyping. P04 tried to steer away from the identified white default observed in previous generations, by retaining their previous prompt structure, but specifying "Japanese" in the prompt:

"A . . . female AI research expert, very focused typing at her laptop in a corporate open space, wearing casual clothes, working environment, office, messy desk"

Interestingly, the model shifted the environmental context of the scene. Unlike the previous bright office settings generated with prompt in which no nationality was specified, the model now generated an image where it appeared to be dark outside (Figure 9b). P04 correlated this to a secondary stereotype of the Japanese overworking culture.

Similarly, P07, probed the system to generate the same scene of a cafe owner but specifying different characteristics of the depicted subject, starting from the following prompt:

" . . . woman in her thirties serving an espresso behind the counter of an Italian 19th century café in Turin. She's the owner of the bar. She's wearing glasses, baseball hat and a hoodie. She looks quite tired from her long shift."

and varying the prompt by specifying "blonde", "Latin American" and "Gambian" as characteristics of the person to depict. The images of "blonde" white women were often depicted as smiley, polished and posing even though the prompt specified for them to be tired. Images of the "Gambian" and "Latin American" women were depicted unsmiling and more focused on the work, shown in Figure 10. P07 suggested that the model shifted from a glamorized bias for white people, to a working-class or labor-focused bias for marginalized identities.

(a) *Prompt: "A gay man. Twenty five years old. City life-style."*

(b) *Prompt: "A gay man. Consider the whole world. Twenty five / thirty years old. Consider other body types."*

(c) *Prompt: "A gay man not from europe. Consider different etnicities[a] and different body types according to world population."*

(d) *Prompt: "A gay man fully clothed without a healthy body. From ecuador. Over thirty-five years old."*

[a]This is a typo made by the user.

Figure 7: P09 Iteration process where the model persistently generated hyper-sexualized or nude figures despite specific prompts for "non-muscular" and diverse body types.

Failure of Identity and Culture The model demonstrated a lack in cultural specificity regarding human identities. For example, P01 mentioned how the girl they generated in Figure 2b looked actually Chinese or Vietnamese instead of Japanese. This indicates that the model lacks the precision to represent specific identities without blending them into a generic result. P06 struggled to get the model to represent "small Chinese baijiu cups", which they wanted to be placed as offerings in front of a tombstone (Figure 8).

Accidental cultural offense P00 attempted to represent a respectful representation of a religious Islamic figure. Since they feared the generation of a caricature, they avoided the explicit use of the term "prophet" and relied on a more descriptive prompt:

Figure 8: Final image obtained by P06 while attempting to represent Chinese men placing baiju cups in front of tombstones.

(a) *Prompt: "A female AI research expert . . . "*

(b) *Prompt: "A Japanese female AI research expert . . . "*

Figure 9: Initial bias and secondary stereotyping encountered by Participant P04 during mitigation.

"a sillohoutte made by the sun of a human figure that is standing on top of a dune. He is holding the regiens of a camel and is looking up. There should be a single person and a camel. The person should not be on top of the camel. He person should have the reins of the camel in his hand."

However, they noticed a glitch in the images, where the head of a camel was generated in the hands of a person (Figure 11). P00 noted that while this may look like a technical error of the model, this image could be interpreted as offensive to religious people, as it resembles a sensitive historical moment (the battle of Karbala).

Mitigation Strategies

Participants adopted labor-intensive strategies to obtain representations they perceived as more satisfactory and less biased. We discuss these techniques as mitigation strategies, including cases in which they failed to overcome system limitations. By revealing the extensive and often invisible labor required to circumvent these biases, we argue that assumptions embedded in AI systems function as barriers to co-creativity.

Negative constraints To manage the boundaries on sensitive or unwanted content, artists utilized negative constraints. They specifically prompted what not to show. This technique was applied either within the text prompt itself

(a) Blonde (b) Gambian (c) Latin American

Figure 10: Comparison of different identifiers used by Participant P07 to test representational shifts.

Figure 11: Example of a technical glitch resulting in involuntary symbolic offense (P00).

or through the dedicated negative prompting field provided by the platform. P00 specifically told the model that the face should not be visible. P02 used a negative constraint, to make sure the image of a person with mental health issues would not portray any wounded imagery. P04 specifically prompted "No glasses", "No white person", "No pretty face". However, these negative constraints still failed to remove these features from the generation. P07 instructed the system to avoid depicting a cafe worker smiling, noting that the model's default outputs were unrealistically staged and cheerful. P09 struggled with the fact that the gay man generated by the model was constantly muscular and white. Therefore they tried to constrain these features. Table 2 details the specific constraints applied by participants

Linguistic Workaround P01 noticed the reason why the generated images were anime style, was because of the word "kimono". To circumvent this problem, they started using a different description to get the same end result. By avoiding the trigger word, P01 successfully bypassed the bias of the model and achieved a realistic representation (Figure 2b).

Self-censorship P00 avoided sensitive religious or cultural scenarios, because they were afraid the model would make them look like cartoons or be disrespectful. Therefore, P00 pivoted to a safer topic. This suggests that the unexpected behaviors of the model limits the creative process of the artist even before generating an image.

Table 3: Specific Negative Constraints Applied by Participants.

Participant	Specific Constraint
P00	"Face shouldn't be visible", "The person should not be on top of the camel"
P02	"No wounds", "Not outside"
P04	"No pretty face", "No white person", "No glasses"
P07	"Not smiling", "No apron"
P09	"Does not go to the gym", "Not from Europe", "Not White", "Non muscular", "Without a healthy body"

Prompt specification P03, who tried to depict Southern Italian landscapes, mentioned that they noticed in prior experience that such landscapes are often stereotyped and depicted as Tuscany (a central Italian region). Anticipating this, they utilized a highly descriptive prompt, which successfully avoided the stereotype, contrasting with a simpler prompt that triggered it (Figure 12). P05, P06 and P08 also found that being more descriptive was necessary to achieve more accurate representations. In addition, to try to work around the white default, both P04 and P07 added different specific identifiers to their prompts. Also P09 had a similar experience, where the model defaulted to a white, muscular gay man. While trying to depict a more diverse man, P09 had to become heavily descriptive in their prompts to overcome the bare-skin and muscular defaults (Figure 13).

Prompt (a): "A young boy from south Italy walks around his small town with his friends."

Prompt (b): "A south Italy landscape, summer, late afternoon. A small italian town's countryside. The sun is about to set. Wind farms are all around the place. Few houses are seen in the back. At the horizon, the mountains stand still. It's very humid and hot."

Figure 12: P03: Prompt specificity for regional accuracy.

Trial and error Almost all artists used the method of trial and error. They would prompt a sentence, see the result, and then add words or delete words. P09 exemplified this struggle by attempting to force the model to generate an image of a clothed gay man. After the initial image was generated, P09 sought a more diverse representation. However, these attempts resulted in shirtless, muscular images. P09 first used the prompt "A gay man . . . walking down the street", assuming that an outdoor setting would imply clothes. When this failed, P09 became more direct and said " A gay man fully clothed . . . ". When this failed again, they eventually added to the prompt that the man was in the mountains, which finally resulted in a non-sexualized result (Figure 13). This interaction highlights that the T2I process is not just about typing a prompt, but a long process of trial and error.

Prompt (a): "A gay man. Twenty five / thirty years old. City life."

Prompt (b): "A feminine gay man fully clothed in winter from Japan, in the mountains. He does not have a muscular body. Over thirty-five years old."

Figure 13: P09: Specificity to overcome hyper-sexualization.

Discussion

This study highlights how representational biases shape the co-creative interaction between artists and T2I models. Because participants were explicitly asked to explore identities and situations that might be affected by stereotyping or censorship, their interactions with the system often revolved around identifying and circumventing these biases. In this section, we highlight broader implications of our findings, while also discussing the limitations of our approach.

Opacity and Bias as Problematic Friction

Participants repeatedly identified representational biases in T2I outputs, including stereotyping, hyper-sexualization, commercialized aesthetics, and white or Western defaults. To navigate these issues, they employed strategies such as negative constraints, linguistic workarounds, increased prompt specificity, and extensive trial-and-error. Participant P04 described this as similar to "correcting a kid", emphasizing that the model required continuous guidance to align with the user's intentions (Qadri et al. 2024). In this sense, prompting becomes a co-creative dialogue in which human creativity, model limitations, and cultural assumptions intersect. A notable finding is that attempts to mitigate one bias sometimes produced secondary forms of stereotyping (Figures 9 and 4). These examples suggest that representational biases in T2I systems are intersectional (Sufian et al. 2025; Doh et al. 2026). Efforts to correct one bias may therefore

activate other latent stereotypes, highlighting the limitations of prompt-based mitigation strategies and the need for internal technical corrections (Solaiman et al. 2023).

This problem is exacerbated by the opacity of T2I models. The millions of parameters and complex statistical associations that generate outputs are largely inaccessible to users (Buhrmester, Münch, and Arens 2021; Wang et al. 2019), leaving them unable to fully anticipate how the model will interpret prompts. To make sense of the system, users develop informal sense-making strategies akin to "algorithmic imaginaries" or folk theories (Bucher 2012; DeVito 2021). In practice, this means that artists must experiment, iterate, and speculate extensively to achieve outputs aligned with their intentions, a labor-intensive process driven by the model's inherent opacity.

These biases and opaque mechanisms produce outputs that diverge from participants' intentions, creating what prior work frames as AI–human friction (DeSchryver, Henriksen, and Leahy 2025). However, differently from existing literature, that also highlights how users can take advantage of these systems to mitigate existing disparities, for example in gender representations (Shihadeh and Ackerman 2023), our findings indicate that this friction is not inherently productive. Instead, it often represents a constraint that requires significant labor to navigate and can result in misrepresentation, cultural offense, or reinforcement of secondary biases. While occasional outputs subverted expectations in constructive ways (such as P07 seeing a woman as the default cafe owner, or P08 not expecting to be able to represent gay men kissing), these moments were rare. By focusing on the human experience of these failures, our work emphasizes that bias in T2I systems is a structural problem: it cannot be fully addressed through prompt engineering alone, as also demonstrated by priori work (Fraser, Kiritchenko, and Nejadgholi 2023), and poses concrete risks for cultural harm and misrepresentation.

Self-Censorship and Algorithmic Power

In addition, our study revealed that artists' creative processes are influenced not only by the outputs of T2I models, but also by the anticipation of system constraints. For instance, P00 avoided explicitly referencing a sensitive religious figure in their prompt due to concerns that the model might produce disrespectful or caricatured representations, even without such instructions. Instead, they relied on indirect descriptions to reduce the risk of offensive results. This anticipatory behavior illustrates how self-censorship emerges in response to the perceived power of the algorithm, which can act unpredictably or "disobey" human intentions (Ihde 1990). The influence of such constraints extends *before* the generation process begins, shaping how artists conceptualize and approach their work with Generative AI.

This phenomenon can be understood through the lens of algorithmic power, a concept developed in the literature on social media and content moderation (Jiang et al. 2023; Petre, Duffy, and Hund 2019; Cotter 2023). Just as moderation algorithms regulate visibility and community compliance, the biases and safety filters embedded in T2I models can exercise power over creative outcomes, limiting what can be safely or successfully represented. Prior work highlights how users adapt to opaque algorithmic systems by predicting system behavior (Bucher 2012; DeVito 2021), yet these adaptations often require extra labor and can restrict expressive freedom as shown in our study. In artistic contexts, such self-censorship risks flattening the richness of cultural expression (Riccio et al. 2022; Taylor et al. 2025).

Taken together, our findings suggest that anticipatory self-censorship is not merely a personal creative choice, but a response to power asymmetries embedded in the design and behavior of AI systems (Riccio, Hofmann, and Oliver 2024).

Limitations

This study is not without limitations. First, although the data collection reached thematic saturation (Bernard, Wutich, and Ryan 2016), the study involved a relatively small sample of participants. While the qualitative approach provides in-depth insights into artists' interactions with T2I systems, the findings may not be fully generalizable to all creative users. Second, the experiments were conducted using Stable Diffusion 3 through a specific third-party platform. As a result, the observed behaviors may partly reflect the characteristics of this particular model and its implementation. Different generative models or platforms may exhibit distinct representational biases, moderation mechanisms, or safety filters (Riccio, Curto, and Oliver 2024).

Conclusion

This paper examined how artists navigate representational biases and censorship in T2I models when attempting to depict marginalized identities. Through a qualitative user study with 10 creative practitioners, we explored how artists identify, interpret, and attempt to circumvent biases embedded in generative systems. Our findings show that interactions with T2I models are shaped by a combination of representational biases and algorithmic opacity. Participants frequently encountered outputs characterized by white or Western defaults, commercialized aesthetics, and the hypersexualization or homogenization of marginalized identities. Attempts to correct these biases through prompt engineering often required extensive trial-and-error and, in some cases, triggered secondary forms of stereotyping, highlighting the intersectional nature of representational biases within the model's internal associations. These dynamics place a significant burden on artists, who must engage in labor-intensive strategies such as negative constraints, linguistic workarounds, and increasingly specific prompts in order to achieve meaningful representations. Moreover, the unpredictability of generative systems can shape creative processes even before image generation occurs, as artists may anticipate misrepresentation or censorship and adapt their prompts accordingly. In this sense, interaction with T2I models reflects broader dynamics of algorithmic power, where opaque systems influence not only outputs but also the conceptualization of creative work. Taken together, our findings suggest that bias in T2I systems should not be understood solely as a technical limitation or a source of creative friction, but as a structural issue with implications for

cultural representation and artistic agency. Designing more inclusive generative systems therefore requires moving beyond user-side prompt workarounds toward deeper technical interventions.

Author Contributions

Mijntje de Groote conducted the primary research and wrote the first draft of the manuscript. **Piera Riccio** provided ongoing guidance and supervision for the research project and actively participated in the writing, reviewing, and editing of the text.

References

Barve, S.; Mao, A.; Shi, J. M.; Juneja, P.; and Saha, K. 2025. Can we debias social stereotypes in ai-generated images? examining text-to-image outputs and user perceptions. *arXiv preprint arXiv:2505.20692*.

Benjamin, R. 2023. Race after technology. In *Social Theory Re-Wired*. UK: Routledge. 405–415.

Bernard, H. R.; Wutich, A.; and Ryan, G. W. 2016. *Analyzing qualitative data: Systematic approaches*. SAGE publications.

Bucher, T. 2012. Want to be on the top? algorithmic power and the threat of invisibility on facebook. *New Media & Society* 14(7):1164–1180.

Buhrmester, V.; Münch, D.; and Arens, M. 2021. Analysis of explainers of black box deep neural networks for computer vision: A survey. *Machine Learning and Knowledge Extraction* 3(4):966–989.

Chatterji, A.; Cunningham, T.; Deming, D. J.; Hitzig, Z.; Ong, C.; Shan, C. Y.; and Wadman, K. 2025. How people use chatgpt. Technical report, National Bureau of Economic Research.

Chen, V.; Liao, Q. V.; Wortman Vaughan, J.; and Bansal, G. 2023. Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7(CSCW2):1–32.

Cho, J.; Zala, A.; and Bansal, M. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3043–3054.

Collyer-Hoar, G.; Rubegni, E.; Tomczyk, B.; Baines, A.; and Gruia, L. 2025. " suits as masculine and flowers as feminine": Investigating gender expression in ai-generated imagery. In *Proceedings of the 2025 ACM designing interactive systems conference*, 915–928.

Cotter, K. 2023. "shadowbanning is not a thing": black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society* 26(6):1226–1243.

de Almeida, F., and Rafael, S. 2024. Bias by default.: Neocolonial visual vocabularies in ai image generating design practices. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. New York, NY, USA: Association for Computing Machinery.

Dehdashtian, S.; Sreekumar, G.; and Boddeti, V. N. 2025. Oasis uncovers: High-quality t2i models, same old stereotypes. *arXiv preprint arXiv:2501.00962*.

DeSchryver, M.; Henriksen, D.; and Leahy, S. 2025. From friction to synergy: The complex interplay of human creativity and ai. *Possibility Studies & Society* 27538699251350483.

DeVito, M. A. 2021. Adaptive folk theorization as a path to algorithmic literacy on changing platforms. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–38.

D'Incà, M.; Peruzzo, E.; Mancini, M.; Xu, D.; Goel, V.; Xu, X.; Wang, Z.; Shi, H.; and Sebe, N. 2024. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12225–12235.

Doh, M.; Gulati, A.; Canali, C.; and Oliver, N. 2026. Aesthetics as structural harm: Algorithmic lookism across text-to-image generation and classification. *arXiv preprint arXiv:2601.11651*.

Epstein, Z.; Hertzmann, A.; of Human Creativity, I.; Akten, M.; Farid, H.; Fjeld, J.; Frank, M. R.; Groh, M.; Herman, L.; Leach, N.; et al. 2023. Art and the science of generative ai. *Science* 380(6650):1110–1111.

Fan, X.; Xiao, Q.; Zhou, X.; Pei, J.; Sap, M.; Lu, Z.; and Shen, H. 2025. User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in ai companions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19.

Fang, C.; Zhu, Y.; Fang, L.; Long, Y.; Lin, H.; Cong, Y.; and Wang, S. J. 2025. Generative ai-enhanced human-ai collaborative conceptual design: A systematic literature review. *Design Studies* 97:101300.

Fraser, K. C.; Kiritchenko, S.; and Nejadgholi, I. 2023. Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems. In *ICCC*, 288–292.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2426–2436.

Gandikota, R.; Orgad, H.; Belinkov, Y.; Materzyńska, J.; and Bau, D. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5111–5120.

Ganguli, D.; Lovitt, L.; Kernion, J.; Askell, A.; Bai, Y.; Kadavath, S.; Mann, B.; Perez, E.; Schiefer, N.; Ndousse, K.; et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Gorska, A. M., and Jemielniak, D. 2023. The invisible women: uncovering gender bias in ai-generated images of professionals. *Feminist Media Studies* 23(8):4370–4375.

Ihde, D. 1990. *Technology and the lifeworld: From garden to earth*, volume 560. Indiana University Press.

Jääskeläinen, P., and Åsberg, C. 2024. What's the look of "negative gender" and "max ethnicity" in ai-generated images? a critical visual analysis of the intersectional politics of portrayal. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. New York, NY, USA: Association for Computing Machinery.

Jiang, J. A.; Nie, P.; Brubaker, J. R.; and Fiesler, C. 2023. A trade-off-centered framework of content moderation. *ACM Transactions on Computer-Human Interaction* 30(1):1–34.

Kirk, H.; Bean, A.; Vidgen, B.; Rottger, P.; and Hale, S. 2023. The Past, Present and Better Future of Feedback Learning in Large Language Models for Subjective Human Preferences and Values. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Ko, H.-K.; Park, G.; Jeon, H.; Jo, J.; Kim, J.; and Seo, J. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th international conference on intelligent user interfaces*, 919–933.

Luccioni, S.; Akiki, C.; Mitchell, M.; and Jernite, Y. 2023. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems* 36:56338–56351.

Mandal, A.; Leavy, S.; and Little, S. 2023. Multimodal composite association score: Measuring gender bias in generative multimodal models. *arXiv preprint arXiv:2304.13855*.

Naik, R., and Nushi, B. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 786–808. New York, NY, USA: Association for Computing Machinery.

Nielsen, J.; Clemmensen, T.; and Yssing, C. 2002. Getting access to what goes on in people's heads? reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction*, 101–110.

Petre, C.; Duffy, B. E.; and Hund, E. 2019. "gaming the system": Platform paternalism and the politics of algorithmic visibility. *Social Media+ Society* 5(4):2056305119879995.

Qadri, R.; Mirowski, P.; Gabriellan, A.; Mehr, F.; Gupta, H.; Karimi, P.; and Denton, R. 2024. Dialogue with the machine and dialogue with the art world: evaluating generative ai for culturally-situated creativity. *arXiv preprint arXiv:2412.14077*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rando, J.; Paleka, D.; Lindner, D.; Heim, L.; and Tramèr, F. 2022. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*.

Rezwana, J., and Maher, M. L. 2023. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction* 30(5):1–28.

Riccio, P.; Oliver, J. L.; Escolano, F.; and Oliver, N. 2022. Algorithmic censorship of art: A proposed research agenda. In *ICCC*, 359–363.

Riccio, P.; Curto, G.; and Oliver, N. 2024. Exploring the boundaries of content moderation in text-to-image generation. In *European Conference on Computer Vision*, 161–178. Springer.

Riccio, P.; Hofmann, T.; and Oliver, N. 2024. Exposed or erased: algorithmic censorship of nudity in art. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–17.

Riccio, P. 2025. *Human aesthetics under the representational power of Artificial Intelligence*. Ph.D. Dissertation, Universitat d'Alacant/Universidad de Alicante.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22522–22531.

Shihadeh, J., and Ackerman, M. 2023. Shattering bias: A path to bridging the gender divide with creative machines. In *ICCC*, 278–282.

Singh, S.; Hindriks, K.; Heylen, D.; and Baraka, K. 2025. A systematic review of human-ai co-creativity. *arXiv preprint arXiv:2506.21333*.

Solaiman, I.; Talat, Z.; Agnew, W.; Ahmad, L.; Baker, D.; Blodgett, S. L.; Chen, C.; Daumé III, H.; Dodge, J.; Duan, I.; et al. 2023. Evaluating the social impact of generative ai systems in systems and society. *arXiv preprint arXiv:2306.05949*.

Sufian, A.; Distante, C.; Leo, M.; and Salam, H. 2025. T2ibias: Uncovering societal bias encoded in the latent space of text-to-image generative models. In *Interdisciplinary Workshop on Responsible AI for Value Creation*, 57–71. Springer.

Sun, L.; Wei, M.; Sun, Y.; Suh, Y. J.; Shen, L.; and Yang, S. 2024. Smiling women pitching down: auditing representational and presentational gender biases in image-generative ai. *Journal of Computer-Mediated Communication* 29(1):zmad045.

Taylor, J.; Mire, J.; Spektor, F.; DeVrio, A.; Sap, M.; Zhu, H.; and Fox, S. E. 2025. Un-straightening generative ai: How queer artists surface and challenge model normativity. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 951–963.

Turchi, T.; Carta, S.; Ambrosini, L.; and Malizia, A. 2023. Human-ai co-creation: evaluating the impact of large-scale text-to-image generative models on the creative process. In *International symposium on end user development*, 35–51. Springer.

Wang, D.; Yang, Q.; Abdul, A.; and Lim, B. Y. 2019. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–15. Glasgow, UK: ACM.

Wang, J.; Liu, X. G.; Di, Z.; Liu, Y.; and Wang, X. E. 2023. T2iat: Measuring valence and stereotypical biases in text-to-image generation. *arXiv preprint arXiv:2306.00905*.

Wu, Y.; Nakashima, Y.; and Garcia, N. 2024. Stable diffusion exposed: Gender bias from prompt to image. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*, volume 7, 1648–1659.