Perceptions and Realities of Text-to-Image Generation

Jonas Oppenlaender
jonas.x1.oppenlander@jyu.fi
University of Jyväskylä
Jyväskylä, Finland

Johanna Silvennoinen
johanna.silvennoinen@jyu.fi
University of Jyväskylä
Jyväskylä, Finland

Ville Paananen
ville.paananen@oulu.fi
University of Oulu
Oulu, Finland

Aku Visuri
aku.visuri@oulu.fi
University of Oulu
Oulu, Finland

ABSTRACT

Generative artificial intelligence (AI) is a widely popular technology that will have a profound impact on society and individuals. Less than a decade ago, it was thought that creative work would be among the last to be automated – yet today, we see AI encroaching on many creative domains. In this paper, we present the findings of a survey study on people's perceptions of text-to-image generation. We touch on participants' technical understanding of the emerging technology, their fears and concerns, and thoughts about risks and dangers of text-to-image generation to the individual and society. We find that while participants were aware of the risks and dangers associated with the technology, only few participants considered the technology to be a personal risk. The risks for others were more easy to recognize for participants. Artists were particularly seen at risk. Interestingly, participants who had tried the technology rated its future importance lower than those who had not tried it. This result shows that many people are still oblivious of the potential personal risks of generative artificial intelligence and the impending societal changes associated with this technology.

CCS CONCEPTS

• **Computing methodologies** → *Artificial intelligence*; • **Human-centered computing** → *Human computer interaction (HCI)*; • **General and reference** → *Empirical studies.*

KEYWORDS

generative AI, text-to-image generation

ACM Reference Format:
Jonas Oppenlaender, Johanna Silvennoinen, Ville Paananen, and Aku Visuri. 2023. Perceptions and Realities of Text-to-Image Generation. In *26th International Academic Mindtrek Conference (Mindtrek '23), October 03–06, 2023, Tampere, Finland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3616961.3616978

Mindtrek '23, October 03–06, 2023, Tampere, Finland

ACM ISBN 979-8-4007-0874-9/23/10.
https://doi.org/10.1145/3616961.3616978

1 INTRODUCTION

Progress in generative artificial intelligence (AI) has exploded in recent years. Generative AI refers to a set of technologies that can synthesize text, images, or other media in response to written prompts as input. This technology has the potential to revolutionize various industries and greatly impact society, particularly in the creative domain. Less than a decade ago, the general consensus was that knowledge work and creative work would be among the last to be automated [3, 11]. However, recent developments in generative AI have contradicted these initial predictions [10]. We increasingly see generative AI being applied in highly creative domains, such as art [24], design [25], and research [23]. One particularly intriguing domain is text-to-image generation, as evident in the popularity of generative systems that can synthesize images from short descriptive text prompts. Such systems include Midjourney[1], Stable Diffusion [30], and DALL-E 2 [29]. Within a short period of time, Midjourney has become the largest Discord community [7], attracting millions of users. StableDiffusion has gained popularity in the open source community since it can be flexibly adapted and personalized to different subject-specific contexts, using fine-tuning on domain-specific images [12, 33]. Outputs from state-of-the-art diffusion models, such as the above, are often indistinguishable from images created by humans [17, 21, 22].

Generative AI has been trained on digital media collected from the Web without prior consent. Many artists and photographers fear for their livelihood as the uptake of generative AI for commercial use is growing [35, 36]. Some call this development "AI's Jurassic Park moment" [18] – an adapt-or-die moment that could potentially result in massive job loss across many sectors. Generative AI can also be an existential risk for organizations and individuals whose business model relies on human effort that can now be automated. For instance, book covers can now easily be illustrated with text-to-image systems, without the need to hire or contract a designer or illustrator for this task. Question answering websites, such as Stackoverflow, are at risk of becoming obsolete due to people's shifting habits of seeking answers for difficult questions from generative AI. Stock photography services are also heavily impacted by generative AI [1]. Even the Tech giant Google is affected as many people shift their search preferences to querying language models instead of tediously sifting through spammy search results [6, 14]. The capabilities of the state-of-the-art text-to-image systems put many organizations and creative professions under pressure.

[1] https://www.midjourney.com

The arrival of generative AI to creative domains raises a plethora of questions about the transformation of the creative industry, human creative practices, and the future of work. But while news about generative AI and its potential impact on the workforce is spreading, we should not forget that many people are still oblivious to the powers of state-of-the-art generative AI. Examining human perceptions of the change brought forth by generative AI sheds light on how this novel phenomenon will affect society. Further, how the role (including risks and possibilities) of generative AI is conceptualized affects the ways it is included in creative practices. Value-laden reporting in media and literature may decisively influence the adoption and regulation of AI (for good or worse) [8].

Against this backdrop, we examine perceptions of text-to-image generation technology, as a popular type of generative AI, among different groups of individuals, including artists, inexperienced users, and (self-reportedly) experienced users of text-to-image generation. The data was collected at the Researchers' Night, a public event at which researchers present their research to the public. Data was collected via an online survey focusing on people's understanding of the text-to-image generation as an emerging technology, its potential uses, and the dangers of the technology for the individual and society.

2 RELATED WORK

The relationship between AI and art has been explored extensively in recent years, in an effort to understand how the perception and attitudes of humans may be influenced by images generated by AI. In this section, we review several seminal studies in this area.

Hitsuwari et al. examined the aesthetic evaluation of AI-generated haiku poems, distinguishing between those created with human intervention and those made solely by AI [15]. Their study suggested that the most aesthetically pleasing haiku were those made in collaboration between humans and AI, implying a certain synergy that enriches creative output. This study also raised questions about the underestimation of AI art, suggesting a phenomenon of 'algorithm aversion.'

Millet et al. identified an anthropocentric bias in art appreciation, positing that recent AI advances in the art domain have challenged traditional human-centric perspectives on creativity [20]. Their experiments involving over 1,700 participants revealed a pervasive bias against AI-created art, which was seen as less creative and induced less awe, hinting at a persistent human bias towards creativity as an exclusively human trait. A similar bias was identified by Ragot et al. in their large-scale study involving 565 participants [28]. The researchers found that art perceived as human-made was evaluated significantly more favorably than that perceived as AI-made. This highlights the potential existence of a negative perception bias towards AI and a potential preference bias towards human-made creations.

In a study focused on the younger generation, Ting et al. explored the perception and acceptance of AI art [34]. The results showed a high level of acceptance, yet more than half of the respondents could not correctly identify the emotions expressed in AI art. This suggests that while AI is becoming more accepted, there are still gaps in its ability to elicit emotional resonance and comprehension. In the context of AI-generated images specifically, Lu et al. investigated whether these images could deceive human observers [17]. Their study found that humans could not significantly distinguish between real and AI-generated images, indicating the sophistication of current AI image generation. However, they also pointed out certain defects in AI-generated images that could potentially serve as cues for discerning authenticity. Finally, Pataranutaporn et al. outlined the potential positive uses of AI-generated media, especially for supporting learning and wellbeing [27]. They called for the inclusion of traceability measures to maintain trust in generated media, reminding us of the ethical implications of this evolving technology.

The retrospective by Rostamzadeh et al. serves as an important reminder of the continued need for ethical vigilance in the burgeoning field of AI-enabled creativity [32]. As AI's role in our creative endeavours continues to evolve, so must our understanding of the ethical boundaries within which it operates. Examining the ethical implications of computer vision in creative applications is important, since this area intersects with everyday life as technology advances. Potential implications encompass issues of privacy, bias, access, representation, and ownership, among others.

The above studies underscore the complexity of perceptions towards AI-generated art, including biases, acceptance, and the need for ethical considerations. Note, however, the great progress that has been realized in recent months. Many studies on the perception of AI art must now be considered outdated, given the strong progress of the field. Our paper provides a novel empirical perspective on this related work.

3 OUR APPROACH

3.1 Method

We gathered data from visitors of the Researchers' Night 2022 event at the University of Jyväskylä. This event is part of the European Researchers' Night[2] and intended for researchers to showcase their research to the general public. We invited visitors to complete an online survey. The online survey was chosen over other methods (e.g., in-person interviews) for several reasons. The Researchers Night is a well-visited public event. At times, the event can feel chaotic and there is a high noise level. To be mindful of people's limited time at the event, and to avoid confounding factors interfering with the data collection, participants were given fliers inviting them to complete the survey in the comfort of their home. Participation was incentivized with a raffle for three Amazon vouchers, each worth 30 EUR.

The questionnaire consisted of 21 items on people's perception of text-to-image generation, emphasizing people's awareness of the risks and dangers associated with this technology (see Appendix A). The questionnaire started with three open-ended items (Appendix A.1) focusing on people's technical understanding of text-to-image generation, as well as potential future applications and the personal and societal dangers of this technology. The responses to the three open-ended items were analyzed using in vivo coding [5]. To this end, the first author read and then iteratively coded

[2]https://marie-sklodowska-curie-actions.ec.europa.eu/event/2022-european-researchers-night

all responses. Multiple codes were assigned, if needed, and iteratively improved and merged by frequently visualizing the codes in histogram charts. Due to the manageable amount of data and the coding being straight-forward, the coding did not require multiple raters and an analysis of inter-rater reliability [19].

In the second section of the survey questionnaire (see Appendix A.2), participants were presented with a scenario of a person submitting an AI-generated artwork to an art fair. This scenario was based on a real event [13]. The third questionnaire section (see Appendix A.3) inquired about participants' experience with text-to-image generation, followed by the importance of text-to-image generation for participants' current and future professional work (Appendix A.4). The questionnaire concluded with demographic questions (Appendix A.5).

Quantitative data were analyzed using an independent two-sample t-test. The significance level was set at $\alpha = 0.05$, and all tests were two-tailed. Effect sizes are reported with Cohen's d.

3.2 Participant Demographics

The online survey was completed by 35 participants (P1–P35, aged 19 to 50, $M = 33.7$ years, $SD = 9.3$ years). Participants had diverse educational backgrounds, the most common being computer science, literature, and information systems. Fourteen participants held a Bachelor's degree, 10 held a Master of Science, four a Master of Arts, three a doctoral degree, and one completed no academic degree. Twenty-four participants (69%) were students.

A third of the participants ($n = 12$; 34.3%) had used text-to-image generation before. The most popular system used by these participants was DALL-E Mini/Craiyon (7 participants), followed by DALL-E 2 (5 participants), Dream/Wombo (3 participants), and Stable Diffusion (2 participants). Participants estimated they had written an average of 20 prompts ($Max = 80$ prompts, $SD = 22$ prompts). Participants were, therefore, rather inexperienced with the emerging technology. Participants who had tried text-to-image generation were younger than those who had not tried the technology ($p < 0.05$, $d = 0.2$). Ten participants (29%) considered themselves as being artists. The art created by this group of participants includes paintings, drawings, writing, and digital forms of art. Less common art forms included clothing, music, handicrafts, and food art.

4 FINDINGS

The following section describes the findings of our survey and presents results on how the general populace understands the technology behind text-to-image generation, potential application areas, and the perceived importance of this technology. The section continues to present ethical challenges and some of the criticism and concerns towards text-to-image generation.

4.1 Understanding of Text-to-image Generation Technology

When asked how the text-to-image system works internally, the majority of the participants ($n = 21$; 60%) did not have a strong understanding of how text-to-image generation works. Many of these participants simply stated that the system "generates" images in response to keywords. In the remainder of this section, we focus on the participants who shared their theories of how text-to-image technology works in more detail.

Participants most often related the technology to image retrieval from a **database**. P1 (31y), for instance, likened the technology to *"some kind of huge photo library, each picture has been coded with a word that it describes it the best. Maybe some other words connected to it. Then it combines the words and finds the best fitting alternatives."* **Search engine retrieval** was a strong theme, mentioned by 13 participants (37%). Participant P10 (25y), for instance, thought *"they use Google or other search engines and combine some of the best results in some way."* One-the-fly retrieval from Google or some other repository or database was mentioned often among these participants. P3 (25y), for instance, mentioned that it *"fetches image from the repository and merges two or more pictures, pre defined ideas and develops a new one,"* and P16 (24y) thought *"when the system receives a text prompt, it goes through a large set of images that correspond to the particular prompt (kind of like Google Images I think), analyses them, and creates a new images based on these existing images."* P25 (19y) stated that *"it searches the web for all material containing and/or even mildly resembling the prompt(s) given. Then the system analyses all of the gathered material, combines elements from several (if not all) of them to generate the final image."* P7 (27y) intermixed two opposing theories on the inner working of the technology, stating that *"either the engine searches your input words from the internet and uses the images found as a reference to compile a completely new image OR the engine has been fed image data from the internet and it uses what it has learned to compile the new image."*

The theme of **mixing or combining** existing images was raised by several participants ($n = 11$; 31.4%). Connecting to the theme of image retrieval from a database, P4 (31y; artist) thought that *"[it] fetches images from the repository and merges two or more pictures, predefined ideas and develops a new one,"* and P35 (33y; artist) mentioned that *"it tries to find corresponding pictures for the words in the text and then combines them to create a final picture."* The text-to-image system was thought to merge and fuse images, even though participants could not explain the inner workings in more detail. P11 (42y), for instance, thought *"it somehow can merge, fuse parts of the images to construct an image based on the keywords in the input sentence."* A different mental image was held by P29 (44y) who thought that the generative technology would replace parts of images: *"it depends on a huge database of labelled images with descriptions of the items shown in them and relations between them (the vase is on the table). Then it's a matter of replacing items. If the user writes: "a cat is on the table", and ML algorithm will replace what it knows as "the vase" with a foto of a cat in the place of the vase. The bigger and more diverse the dataset, the better the results."*

A minority of participants ($n = 8$; 22.9%) had some understanding of how the technology works. P23 (35y), for instance, wrote that *"The AI has learned to produce pictures while the prompts function as parameters for the algorithm. The AI probably has a large quantity of existing pictures as a learning material that has been coupled with keywords. Through feedback the AI has gradually become better at producing pictures that match the prompts."* However, in their description of the technology, only four participants explicitly distinguished between training and inference time. The 'training time' is when the AI learns from vast datasets, while the 'inference time' is when the AI applies this learned information to generate

new outputs. Distinguishing between training and inference is an important step in understanding how generative AI works. Most participants did not make this important distinction.

4.2 Potential Application Areas of Text-to-image Generation

Creative domains were most common when it comes to participants' thoughts about potential applications for text-to-image generation (see Figure 1). Creating **digital art** was the strongest application area seen by participants. Participants thought the technology was well suited for creating digital artworks, illustrations, logos, and other visual media. Besides directly creating artworks with the technology, one participant also mentioned that generative AI could *"be a good tool for artists to have multiple different references while making their own artwork"* (P15; 21y; artist). P3 (25y) acknowledged commercial use of text-to-image generation and mentioned the synergy of this technology with other technologies, such as non-fungible tokens (NFT), for selling and buying digital artworks. While some participants thought of text-to-image generation as a powerful co-creative tool in the toolbox of artists (e.g., P10, P15, and P23), many participants thought the technology was a potential replacement for artists and designers altogether. P11 (42y) imagined *"a system where you give it the requirements of a design, you press a button, and you get a multitude of designs to choose from."* This system grants non-experts the creative capabilities that were once exclusive to professional designers. P9 (30y) acknowledged that text-to-image technology, therefore, *"lowers the barrier to creating images."* Text-to-image generation systems *"may provide a method for quickly creating the needed pictures when required for any given project and decrease the effort and skills needed for their creation"* (P9; 30y). Generative design systems may provide users with an end-to-end way of producing artworks, without having to turn to artists and designers as middlemen in the creation process. P14 (44y; artist) pondered that *"it will remove the human in the production of illustrations for all sorts of purposes."*

Further application areas mentioned by participants included **brainstorming and ideation**, in application areas such as new product development (NPD) and design. Text-to-image generation could help to visualize ideas and designs, *"to see what it could look like"* (P2; 24y). P27 (39y) acknowledged that *"these systems could lessen the need to create visual material from scratch."* Generative AI provides a means to synthesize *"the optimal photo or image for some [specific] purpose"* (P19; 50y). In general, text-to-image generation *"makes it easier and faster to create pictures for ads, maybe animated TV shows, etc."* (P2; 24y). Text-to-image generation was seen as a fast and cheap alternative to *"manipulating photos or doing digital art"* (P7; 27y). Text-to-image generation *"decreases the costs of creating pictures traditionally (aka. with a camera, studio set-ups, the cost of artists' work, etc.)"* (P9; 30y). Text-to-image generation considerably lowers the price for illustrations and *"producing pictures for advertising products"* (P5; 29y). Therefore, one large application area was seen in **advertising and marketing**, to generate *"cheap images for ads and illustrations. It might replace stock photography websites and companies"* (P4; 31y; artist), and would be suitable for "game concept art or marketing art" (P15; 21y; artist). Participants further mentioned that the technology could be used in journalism and media, to generate images for magazine articles and websites, and a broad range of visual illustrative media, such as *"illustrations for cards, childrens' books or almost anything"* (P33; 45y; artist).

Another strong application area was **fun and entertainment**. Text-to-image generation makes a fun pass-time, according to participants. The entertainment industry was seen as an application area, for instance for making animations and games *"only with a script"* (P8; 25y). Some participants likened the fun derived from text-to-image generation to meme creation (e.g., P18; 36y). P20 (42y) pondered about social applications, and thought it *"would be fun to 'play' with it with other people and create a social-pictures, or something like that."* P24 (37y) mentioned that *"the systems are great fun and humorous,"* but had concerns that *"if the systems get 'better' in the future, it ruins that fun."*

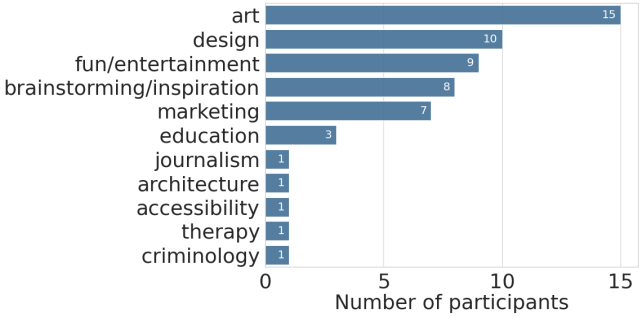

Figure 1: Participants' thoughts on potential application areas of text-to-image generation.

In the remainder of this section, we highlight some less commonly mentioned application areas, including **education, therapy, journalism, criminology, and accessibility**. P31 (23y; artist) recognized the broad potential of text-to-image generation, and stated that *"the potential of these systems is infinite. They could be used in schools to aid teaching, in therapy, to speed up design processes such as games, etc."* As for applications in education, P24 (37y) thought that the technology was useful because *"pictures are in many cases a more effective way to describe things than words."* Text-to-image generation was seen both as a tool to aid teaching in schools (P31; 23y; artist) as well as to *"inspiring kids, giving ideas"* (P1; 31y). In the educational context, text-to-image generation could be applied to illustrate educational materials. One participant mentioned that text-to-image generation could be useful in criminology *"to reconstruct crime scenes in some cases"* (P33; 45y; artist). P14 (44y; artist) also alluded to a forensic use by mentioning that *"it can be used to detect connections between images. It can link images e.g. an image of a person with some rash can be linked to some disease."* Last, participants mentioned applications for accessibility, as a kind of universal tool that could help people with accessibility needs: *"It could be used as a kind of visual dictionary or translator; you say a word and the machine draws it"* (P34; 25y; artist).

4.3 Perceived Importance of Text-to-Image Technology

Most participants responded that text-to-image generation did not hold any importance in their personal and professional lives, but

acknowledged that it could play an increasingly important role in the future (see Figure 2). Interestingly, those who had tried it before found text-to-image generation not as important for their professional future, as opposed to those who had not tried it before. This difference was significant ($p < 0.05$, $d = 0.53$) and not found among self-declared artists.

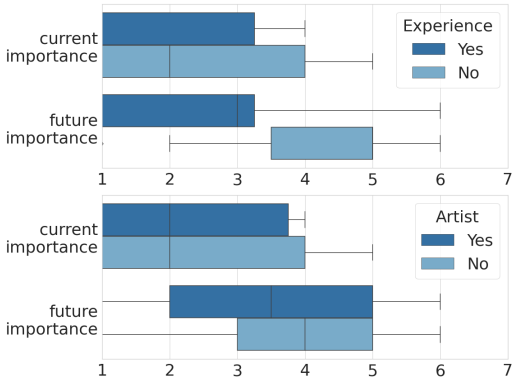

Figure 2: Boxplot comparison of current and future professional importance of image generation for participants who did and did not try text-to-image generation before (left) and self-identified artists and non-artists (right), on a Likert scale from 1 – Not At All Important to 7 – Extremely Important.

4.4 Ethics of Disclosing AI Generation

About half of the participants ($n = 19$; 54.3%) were of the opinion that it should be disclosed when something was created with AI. Ten participants (28.6%) had no strong opinion about this, and six participants (17.1%) thought that AI-generated images do not need to be labeled as such. When presented with the scenario of an AI-generated artwork being submitted to an art fair, participants thought that it was unethical to submit without disclosing that the image was created with AI (see Figure 3). Interestingly, not labeling a submission to an artwork contest as "created by AI" was seen just as unethical as submitting an artwork created with a stolen prompt.

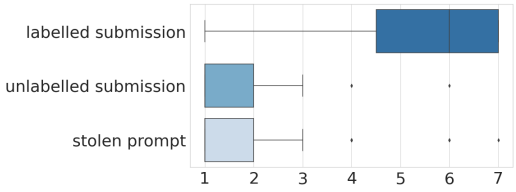

Figure 3: Rating of the ethics of submitting a text-to-image artwork to an art fair on a Likert scale from 1 – Not Ethical At All to 7 – Highly Ethical.

4.5 Criticism and Concerns about Text-to-image Generation

Many participants did not see a risk or danger for themselves. P24 (37y), for instance, mentioned *"I couldn't think of anything that could be dangerous for myself."* But while the majority of participants did not think that text-to-image poses a personal danger to themselves, participants still voiced many concerns about the effect of this emerging technology on society as a whole (see Figure 4). It was a common theme to not see dangers personally, but note them for society. P17 (23y) remarked, for instance, *"I don't see much danger to myself. I find most internet-based activity affecting negatively the society."*

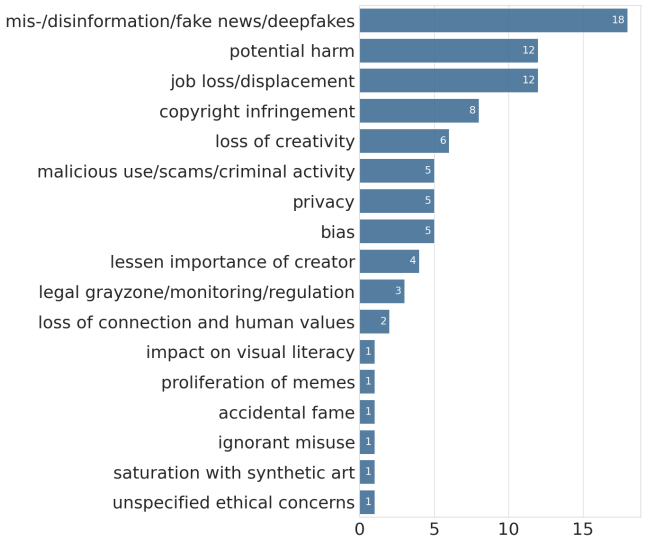

Figure 4: Participants' thoughts on risks and dangers of text-to-image generation.

The use of AI-generated imagery for opinion manipulation, fake news, and "deep fakes" was leading cause for concern (see Figure 4). Many participants warned that synthetic images could be spread naively (**misinformation**) or for malicious purposes (**disinformation**). Text-to-image generation could be used for "creating false re-creations or look-a-like versions that can cause harm" (P19; 50y). In the hands of authoritarian leaders, the technology was seen as especially dangerous. Text-to-image generation *"could increase the amount of disinformation circulating on social media. Certain deepfake-systems have already caused havoc for example in the politic fields, so an AI system like this, if powerful enough, poses a huge threat"* (P25; 19y). The realism of images synthesized by generative systems was seen as problematic in the context of disinformation and fake news. P6 noted that *"it makes so authentic-looking pictures, that it might be difficult to differentiate generated fake pictures from genuine ones. These pictures might be used for propaganda for example in fake news."* According to P7 (27y), *"more and more fake images will start circulating"* which will cause an *"ethical problem: Is it okay to make AI generated pictures of other people? Is there a difference between making these images of private people vs. public figures? It becomes increasingly harder to distinguish what media is factual and what AI generated."* This would make it *"harder to know in the future what information is reliable - a photo isn't as reliable proof as it used to be"* (P11; 42y). P12 (48y) worried about this development, noting *"that more and more things can be*