ORIGINAL RESEARCH

Situating the social issues of image generation models in the model life cycle: a sociotechnical approach

Amelia Katirai[1] · Noa Garcia[2] · Kazuki Ide[1,3] · Yuta Nakashima[2] · Atsuo Kishimoto[1,2,3]

Abstract

The race to develop image generation models is intensifying, with a rapid increase in the number of text-to-image models available. This is coupled with growing public awareness of these technologies. Though other generative AI models—notably, large language models—have received recent critical attention for the social and other non-technical issues they raise, there has been relatively little comparable examination of image generation models. This paper reports on a novel, comprehensive categorization of the social issues associated with image generation models. At the intersection of machine learning and the social sciences, we report the results of a survey of the literature, identifying seven issue clusters arising from image generation models: data issues, intellectual property, bias, privacy, and the impacts on the informational, cultural, and natural environments. We situate these social issues in the model life cycle, to aid in considering where potential issues arise, and mitigation may be needed. We then compare these issue clusters with what has been reported for large language models. Ultimately, we argue that the risks posed by image generation models are comparable in severity to the risks posed by large language models, and that the social impact of image generation models must be urgently considered, even as their development and deployment are pushed forward.

1 Introduction

Rapid progress in image generation models has caused major disruption, both to the field of computer science, and to society at large. Within the span of just a few months in 2022, multiple text-to-image models, including DALL-E 2 [1] April 2022), Dall-E Mini [2], Midjourney (July 2022), and Stable Diffusion [3] August 2022) were released, with the ability to produce high-quality photo-realistic and artistic images from a text prompt. These developments were rapidly followed by the release of text-to-video models including CogVideo [4], Make-A-Video [5], or Imagen Video [6]. The high quality of the visual output generated, easy-to-use interfaces, and powerful commercial incentives have pushed these models out of research labs and into the awareness of the public. However, the speed of these advancements has created a vacuum, as there remains relatively little understanding or awareness of the social implications of these shifts. Where there is broader awareness of social issues, the attention is often unbalanced, with a tendency, for example, to prioritize issues related to intellectual property, while other issues remain insufficiently understood.

Although interest in the use of computers to generate art was present throughout the 1900s, traced back as far as to the 1950s, the technology ultimately reached a breakthrough in 2015 through Google's development of DeepDream, a neural network trained on over 10,000 paintings [7]. Earlier image generators were based on a type of machine learning referred to as generative adversarial models [8], but as will be described below, current text-to-image models leverage CLIP [9] embeddings to jointly capture language

✉ Amelia Katirai
a.katirai.elsi@osaka-u.ac.jp

1 Research Center on Ethical, Legal and Social Issues, Osaka University, Osaka, Japan

2 Institute for Datability Science, Osaka University, Osaka, Japan

3 Center for Infectious Disease Education and Research, Osaka University, Osaka, Japan

and visual semantics, drawing on CLIP's general-purpose image-text model [7]. Both CLIP and the image generation models based on it owe their high performance to a training process based on massive collections of data crawled and scraped from the Internet, which allows the models to classify images without necessarily requiring human labeling [7]. At the same time, the use of such uncurated data and its effects on the final model's output raises both immediate and longer-term social and ethical concerns, including risks of bias, threats to intellectual property and privacy, and broader impacts on the social environment.

To date there has been relatively little consideration of the social issues raised by image generation models, with the exception of some recent groundbreaking work in the field [10, 11]. In this paper, we provide a comprehensive categorization of the risks posed by the mass-scale adoption of these models. The concerns around image generation models share similarities with other types of generative models such as LLMs; however, the particularities and implications of the visual modality deserve to be paid special attention. Thus, as with the ongoing scholarly debate about large language models (LLMs) [12], our aim is to raise awareness about the potential negative societal impact of image generation models.

Joyce et al. [13, 14] have noted a need for increased consideration of the sociotechnical nature of machine learning models, attending to both the technical aspects of the models, and their social aspects, drivers, and implications. Conceptualizing this as the intersection of two axes—a technical axis and a social axis—we find that the relatively limited work to date on image generation models has been situated along either one of these axes. We report here on the results of a novel interdisciplinary collaboration at the *intersection* of these two axes, between experts in machine learning and those in the social sciences. This collaboration was undertaken between researchers at two research hubs at Osaka University. The Institute for Datability Science is an Institute focused on cutting-edge computer science research, while the social scientists at the Research Center on Ethical, Legal and Social Issues conduct research on the social and ethical implications of emerging applications of technology, and most recently, AI. Taking advantage of the research hubs' physical proximity and their shared focus on AI—albeit from distinct approaches—arrangements were made for a collaboration which, by combining strengths from each group of researchers, could provide a more comprehensive overview of the issues posed by image generators. The researchers from each hub met regularly to discuss methods and approaches appropriate for this review, and to consult on findings, with presentations from researchers introducing relevant research conducted on each side. The final determination of the issue clusters presented here was made through a collaborative process, refining the clusters, and then reflecting these in a shared manuscript, to which all authors contributed. Here, drawing on a review of the recent literature, as well as this interdisciplinary expertise, we first detail the components in the life cycle of image generation models, from how the training data is acquired to the long-term use of the generated output. Then, we offer a novel taxonomy for the potential social issues of image generation models. By intersecting the technical and social axes, we identify where in the model life cycle of the technology each issue arises or is reinforced. Indeed, as Huang et al. [15] have argued, it is imperative that ethical issues in AI be understood in the context of the AI lifecycle. Thus, we ground the longer-term—and at times more abstract—issues of these models within the models themselves, while also providing a guide both to the model life cycle, and to its associated issues, for researchers working along either of these axes.

It is noteworthy that, although we use "social" here as an overarching category, we refer broadly here to non-technical matters, including issues which have been framed elsewhere as ethical ones, as well as those from the humanities. There is growing concern over the social and ethical implications of AI, with attention to a broad range of themes ranging from emergent issues of environmental harm or the reproduction of social inequality, to longer-standing concerns over issues of control for advanced AI systems [16–20]. Huang et al. [15] propose a taxonomy of ethical issues in AI, arguing that these issues arise at multiple levels, including the individual, society, and the environment. Discussions over the ethics of AI have also expanded to include those informed by religious [21] or cultural values [22, 23]. However, the rise of generative models also present particular "risks and challenges," which extend beyond what had been previously seen in relation to AI [24].

To this end, along the technical axis, we identify four main phases in the model life cycle: data acquisition, model design, deployment, and the use of generated images; along the social axis, we identify seven issue clusters: data issues, intellectual property, bias, privacy, and the impacts on the informational, cultural, and natural environments, respectively. Given the strong recent attention to LLMs, we use LLMs as a point of comparison, comparing the risks identified for image generation models with those reported for LLMs. We find that, despite the relative lack of attention to the risks of image generation models, their risks are comparable in severity. Issues of intellectual property, bias, and the impact on the information and cultural environments are especially of concern, given the unique features of images as a medium. As a result of this sociotechnical examination, we argue that attention to these profound risks is urgently needed by researchers along both axes, and at their intersection. We close with preliminary recommendations
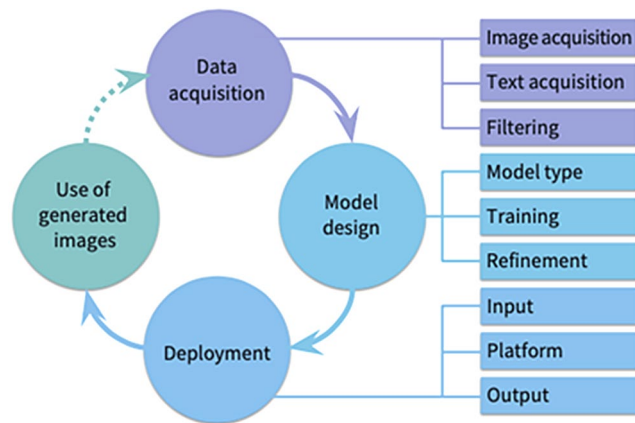
Fig. 1 The model life cycle

to minimize the negative impact of text-to-image models in society. Throughout, the paper contributes to the early body of literature on this theme by providing a summary of the key issues surrounding text-to-image models, while at the same time breaking ground by situating these social issues within the model-life cycle of text-to-image models.

2 The model life cycle

Text-to-image models generate images based on an input text. Here, we examine the life cycle of such models. The first models creating images from text were proposed back

in 2016 [25] by encoding an input sentence through a long short-term memory network (LSTM) that conditioned an image decoder based on a convolutional neural network (CNN). Datasets used for training such networks were rather small, limiting the model's capabilities and, as a result, producing notably blurry images. Later, with DALL-E [26], a breakthrough in image quality was achieved by leveraging CLIP [9] embeddings. As a method for encoding images and texts, CLIP learns the semantic correspondence between the visual appearance of objects in images and the words in text by comparing image and text data samples in a collection of about 400 million image-text pairs from the Internet. Current text-to-image models build on top of the semantic knowledge of these pre-trained CLIP embeddings to replicate the appearance and composition of the elements mentioned in the input text and generate high-quality images through copious amounts of training data.

Regardless of the obtained image quality, we identify four phases involved in all text-to-image models' life cycle, as shown in Fig. 1, which are (1) data acquisition, (2) model design, (3) deployment, and (4) use of generated images; each phase is further divided into the sub-processes defined below.

2.1 Data acquisition

As with any machine learning application, text-to-image models need data samples to learn the internal

Table 1 Datasets used in text-to-image models

Dataset	Year	Image acquisition		Text acquisition			Filtering (automatic unless otherwise stated)	Data consent/ License
		Number	Method	Number	Method	Language		
SBU Captions [27]	2011	1 M	Flickr	1 M	Flickr	English	Text quality	No
Flickr30K [28]	2014	0.032 M	Flickr	0.159 M	Crowdsourcing	English	Text quality	
MSCOCO [29]	2015	0.33 M	Flickr	1.5 M	Crowdsourcing	English	Image content	No
Visual Genome [30]	2017	0.11 M	Flickr	5.4 M	Crowdsourcing	English	Manual text quality	
GCC3M [31]	2018	3.3 M	Web crawler	3.3 M	Web crawler	English	Image and text format Offensive content Text-image matching	No
Nocaps [32]	2019	0.015 M	Flickr	0.166 M	Crowdsourcing	English	None	CC BY 2.0
TextCaps [33]	2020	0.028 M	Flickr	0.142 M	Crowdsourcing	English	Image content Manual text quality	CC BY 2.0
GCC12M [34]	2021	12.4 M	Web crawler	12.4 M	Web crawler	English	Image and text format Offensive content Text-image matching	No
WIT [35]	2021	11.5 M	Wikipedia	37.5 M	Wikipedia	108 lang.	Image format Text quality Offensive content	Research license
RedCaps [36]	2021	12 M	Reddit	12 M	Reddit	English	Offensive content	Public form to request removal
LAION-400M [37]	2021	400 M	Web crawler	400 M	Web crawler	English	Image and text format Text-image matching	No
LAION-5B [38]	2022	5850 M	Web crawler	5850 M	Web crawler	+100 lang.	Image and text format Text-image matching	No

representations of the model and conduct the task of interest. The data acquisition phase consists of gathering and preprocessing those data samples. Text-to-image models require data to be in the format of image-text pairs—for each image there should be an accompanying text with its description. Data samples are used for training, fine-tuning, or evaluating text-to-image models. An overview of data sets used in this phase of the model life cycle is provided in Table 1. Within the data acquisition phase, we identify three sub-phases:

1) **Image acquisition**: The sub-phase for obtaining images. Early image-text datasets were mostly domain-specific and relatively small, with images carefully picked from the web or captured by the dataset authors (e.g., Oxford-102 Flower [39], Caltech-UCSD Bird [40]). In medium-scale and general-domain collections, such as MS-COCO [41], images were obtained by keyword search and filtered during the annotation process. Lately, text-to-image models are trained on web-scale collections of images crawled from millions of websites and with little-to-no human intervention, such as LAION-5B [38].

2) **Text acquisition**: The sub-phase for obtaining text describing the images. In small- and medium-scale image datasets, text can be written by human annotators [42]. Human annotations are not feasible in web-scale collections, so the text is commonly obtained together with the images from the alt-text HTML attribute in the original websites [31].

3) **Filtering**: The sub-phase for discarding data samples that are not appropriate for the task. Filtering can be based on multiple criteria, including data domain (e.g. only keeping images from flowers [39]), image quality (e.g., removing images with dimensions smaller than 400 pixels [31]), image or text content (e.g., deleting samples labelled as "not safe for work" or NSFW [36]), license (e.g., only images with research permissive license [35] etc.) In small- and medium-scale datasets, there can be a human verification step to check the quality of the text [33]. In web scale datasets, however, automatic filters are applied [31, 33]0.2.

2.2 Model design

Different architectures, training strategies and parameters can be used to develop a text-to-image generation model. We refer to the process of choosing these components as the model design phase.

1) **Determination of model type**: The choice of the core technology used in the text-to-image model. Early models (e.g. [42]) were based on generative adversarial networks (GANs) and built on top of recurrent neural networks (RNN)-based text encoders and convolutional neural networks (CNN)-based image decoders [8]. In most of the newest models, text is encoded with a CLIP [9] pre-trained Transformers [43] CLIP embeddings, and images are decoded either with autoregressive [44, 45] and diffusion [46–48] techniques.

2) **Training process**: The process for automatically adjusting (or learning) a model's parameters according to a target task. There are several factors involved in the training process:

   a. **Optimization objective**: Text-to-image models are commonly optimized to increase image quality by setting the target task to make, given an input text, the generated image be as similar as possible to the image originally paired with said text (hence, the necessity of paired image-text samples in the data acquisition process). The target task used during training is a design choice by itself, and by optimizing models based solely on image quality, other variables such as efficiency, cost, or fairness tend to be overlooked.

   b. **Training duration**: Models learn to adjust their internal parameters by iterating over training data and fitting the target task for each of the training samples. In large models, the training process with thousands of millions of text-image pairs can take several months with large computational costs. Information about this process is rarely available, and only DALL-E mini and Stable Diffusion have disclosed this information. According to the model card[1], Stable Diffusion v1.1 was trained for 150,000 GPU hours and emitted 11,250 kg CO2 eq.

   c. **Hyperparameter search**: In addition, models usually have multiple hyperparameters that need to be set manually and are not learnable through training, e.g., learning rate, batch size, number of training steps, etc. It is common to train multiple models with different hyperparameters to find the optimum configuration, increasing computation time and cost. This hyperparameter search is usually not reported in the computational cost of the models.

---

[1] https://github.com/CompVis/stable-diffusion/blob/main/Stable_Diffusion_v1_Model_Card.md.

3) **Refinement**: Although not yet reported in text-to-images models, LLMs have been improved by the use of reinforcement learning (RL) techniques with annotations on whether the generated data is good [49]. Fine-tuning has also been introduced for particular text-to-image models, such as Stable Diffusion [50].

2.3 Deployment

Once a model is trained, the next phase is to share it and make it available to users. We refer to this phase as the deployment phase.

1) **Input**: The input to the model consists of a text, also known as a prompt. The input text is fed into the model to generate the output.
2) **Platform**: Where the model is hosted: private server (e.g., DALL-E 2), third-party server (e.g., DALL-E mini), or the user's own machine (e.g., CogView, Stable Diffusion).
3) **Output**: The output consists of the generated images. For an identical input text, different models can generate different images according to each model's style. The use of such output is listed as a separate entry of the model life cycle because of their importance.

2.4 Use of generated images

The final phase is the use of the generated images, which may be utilized by users in a range of settings. However, as the generated images can be published on the Internet, they can also potentially be collected in the data acquisition

phase to train future models.[2] We see two key avenues through which this can occur.

1) **By users**: As the generated images can be published back on the Internet, they can potentially be collected in the data acquisition part for training future models.
2) **By the model developer**: Improvement of the model by annotating the generations. Large language models have been reported to use reinforcement learning for this task and could be applied to text-to-image models in the future [49].

Thus, despite the diversity in the range of text-to-image models available, they share commonalities in the steps across their life cycle. A summary of the major text-to-image models is provided in Table 2.

3 Social issues in image generation models

Following on this map of the model life cycle, we report on a review of the literature on the social issues arising from image generation models. Expanding on existing approaches to image generation models by Zhang et al. [51] and Ko et al. [52] and with acknowledgement of recent work on image generation models [10, 53] and reference to recent work on LLMs [12, 54], we present seven issue clusters: data issues, intellectual property, bias, privacy, and the impacts on the informational, cultural, and natural environments.

The review undertaken here was based on a flexible review method. A flexible approach was prioritized over a structured or systematic approach to accommodate the rapid developments in the field. The review drew primarily on hand-searches and used a snowballing approach to expand from core papers in the field to other related works. This

Table 2 General text-to-image models. There are two main families of models: autoregressive (AR) or diffusion (Diff.). Access: X neither the code nor the model has been released; ✓ the model can be freely accessed; ✎ the model is only accessible under registration

Model	Date	Institution	Type	Params(B)	Images (M)	Access
DALL-E	2021/02	OpenAI	AR	12.0	250	X
CogView	2021/06	Tsinghua Uni.	AR	4.0	30	✓
GLIDE	2021/12	OpenAI	Diff.	5.0	250	X
GLIDE (filtered)	2021/12	OpenAI	Diff.	0.3	137	✓
CogView 2	2022/03	Tsinghua Uni.	AR	24.0	30	✓
DALL-E mini	2022/04	Craiyon	AR	0.4	15	✓
DALL-E 2	2022/05	OpenAI	Diff.	6.5	650	✎
Imagen	2022/06	Google	Diff.	3.0	860	X
Parti	2022/06	Google	AR	20.0	9	X
Midjourney	2022/07	Midjourney	-	-	-	✎
Stable Diffusion (v1.1)	2022/08	StabilityAI	Diff.	1.5	2,000	✓
Firefly	2023/03	Adobe	-	-	-	✎

---

[2] This is why we have connected "Generated images" and "Data acquisition" with a dotted arrow in Fig. 1.

was supplemented with independent searches of the academic literature and grey literature, and reference sharing within the interdisciplinary team to identify relevant works from various fields. Throughout, we prioritized most-recent works. The review here is non-exhaustive but intended to convey key themes from the literature. Thus, the literature search was ended once saturation was perceived to be reached given the scope of the paper.

These issue clusters were extracted from the literature using an approach informed by thematic analysis [55]. Key areas of discussion in the literature were first identified and noted. These were then grouped into broader thematic categories which we refer to here as issue clusters. The issue clusters were reviewed and adjusted by the interdisciplinary team in light of their expertise. It is noteworthy that there are novel initiatives across multiple areas to bring progress on addressing or resolving some of the issues identified. However, issues were taken as the focus of this review, and such initiatives were treated as outside the scope of this exploratory analysis. Each issue cluster and noteworthy points raised in relation to it will be taken up in turn in the sub-sections that follow.

3.1 Data issues

The first issue cluster identified is data-related issues. Image generation models are reliant on massive amounts of "unconstrained data"—datasets gathered through means such as web scraping. These methods, though cost-efficient, involve "regarding people as `raw material' without the right to privacy, or the agency to consent or opt-out" [56]. Gebru et al. and others [56, 57] highlight the necessity of appropriate documentation for datasets used in machine learning, to ensure that harms can be avoided or remedied. Yet, the size of the datasets used in the development of generative models poses an inherent obstacle to such mitigation, particularly as they lack metadata which could be used as ground truth for their evaluation [56]. This is compounded as the creators of the models may be disincentivized from making public the details of their training data, due to the major investments involved in their creation [58]. Therefore, the identification of emergent issues related to training data often relies on the analysis of output. For example, Schramowski et al. [59] analyzed output generated using Stable Diffusion, finding "a significant amount of inappropriate content generated which, viewed directly, might be offensive, ignominious, insulting, threatening or might otherwise cause anxiety," highlighting problematic training data as a key issue. This points to inherent issues in this approach, where the size of the data set prohibits documentation. This has led some scholars to critique the way in which larger datasets are

"mythologized as beneficial" [56] for training models, proposing rather that smaller datasets and thorough documentation of their contents may help to offset some of the issues arising from them, as will be discussed in multiple sections below.

It is also significant that a lack of documentation can hamper efforts to filter out problematic data from the training sets. Yet, filtering itself can become a site for the reproduction of multiple types of social inequality. For example, Monea [60] argues that content moderation can lead to the creation of a "digital closet," through which LGBTQ + people are "so pornographied" through content moderation that their sexual expression is restricted. Ungless et al. [61] identify a tendency to "consistently" and problematically represent "certain non-cisgender identities. as less human, more stereotyped and more sexualized," including through the generation of pornographic images. Yet, their study also found that when NSFW images were blocked, this was "likely to contribute to the erasure of non-cis identities," rather than to "prevent the generation of harmful output." In a similar vein, based on their attempts at bias mitigation, Nichol et al. [62] "hypothesize" that the training data used for DALL-E 2 "may be biased toward presenting women in more sexualized contexts," but found that filtering data to remove this content led to underrepresentation of women in output in response to neutral prompts. Further, the content moderation process can itself reproduce social inequality as reported by Gray and Suri [63] for AI broadly, who highlight the precarious employment conditions in which content moderators engage in often traumatic work. In early 2023, Perrigo [64] drew attention to these issues in relation to generative models particularly, raising alarm over the poor working conditions of Kenyan content moderators for OpenAI's ChatGPT—issues which are also reflected in the data annotation process [56].

3.2 Intellectual property

Data issues are also closely linked with issues of intellectual property, which can arise both through the training data used for image generation models, and through the output generated by the systems, which may (intentionally or unintentionally) have similarities to copyrighted work [65]. It is notable here that understandings of "fair use" differ across national contexts: in the United States, for example, published works may be used for training models, while in the EU, artists are able to disallow the use of their work if it will ultimately be used commercially [66]. However, the removal of copyrighted data from a large-scale training dataset may be impractical [67], given the large amounts of data and lack of documentation as described above [57]. Moreover, it may be viewed as "undesirable" [65], as the

inclusion of copyrighted works can be seen to strengthen the resulting model by providing a broader range of sources to draw on. It is significant that, as reported by Dehouche and Dehouche [7], the development of text-to-image models was supported by services such as *ArtStation* which "encourage artists to include detailed labels describing their work in order to make it more accessible to persons with disabilities," thus providing ready text-image pairs, but raising complex issues in relation to intellectual property when these images are used to train image generation models.

The tendency of image generation models to memorize and regenerate images ("image regurgitation" [68]) can pose a particular problem when the models are developed from "stolen artwork from the internet [sic.] without permission". This has led to legal action, with a notable example being a lawsuit against Stable Diffusion by Getty Images, with output from Stable Diffusion containing a regenerated version of the Getty Images watermark, overlaid on "grotesque" [69] images. There is an alternative perspective, however, which sees the regurgitation of training images as a positive, as it allows for the generation of specific images which have cultural significance [70]. Though mitigations are, as will be discussed below, beyond the scope of this paper, it is noteworthy here that initiatives are underway to attempt to fight against violations of intellectual property. Although multiple companies purportedly offer the opportunity to opt-out of the use of data in AI training, reports suggest that many of these processes are window-dressing or and ineffectual [71]. As an example, "Nightshade" has been proposed as a possible mechanism for disincentivizing violation of "do not crawl" indicators and the use of copyright-protected images, by allowing for a model to be "poisoned," such that images cannot be generated as intended in response to particular prompts [72].

An additional issue is how intellectual property rights should be handled for the works generated by AI, or the prompts used. Though the United States Copyright Office [73] only recognizes "'the fruits of intellectual labor' that 'are founded in the creative powers of the mind,'" it is as yet unclear how much human involvement would be required to meet this standard, or whether a well-articulated prompt, for example, might be sufficient [74, 75]. These issues have come to the fore given a growing market for prompt trading, where prompts are sold for generating art in particular styles [76, 77]. Though some conclude that the balance of arguments lies against granting copyright to machine-generated works [78], others argue that "[n]o AI acts alone, completely divorced from the influence of humans" [79]. Thus, the rise of generative models brings into focus an urgent need to reconsider how copyright and other intellectual property rights are conceptualized [75].

3.3 Bias

A further issue in image generation models is bias [51, 67, 80]. As described above, the proprietary and closed nature of image generation models poses a barrier to the evaluation of the full extent of algorithmic biases present in generative models [80]. Though further work is urgently needed [80], preliminary work has begun to document how the output from image generation models reproduces and amplifies harmful stereotypes and biases [80–82]. "Idiosyncratic biases along certain dimensions" [83] are reported in each of three major image generation models (DALL-E 2, Midjourney, and Stable Diffusion), though there are major challenges in ensuring that research on biases keep pace with the development of new or updated models.

The presence of intersectional biases in the models has been found to be a result of a lack of representativeness in the training data gathered from the Internet [84]. They are also due to a Western and American-oriented bias in the models [85]; there is thus a need to examine how bias may be reinforced through the use of other languages [86]. Cho et al. [87, 88] found that the use of image-text pairs from the Internet embedded intersectional biases between gender and skin tone in image generation models, finding this tendency reflected in output from models such as Stable Diffusion v1.4, which were more likely to generate images of certain gender and skin tone combinations even when prompts were neutral. Moreover, Cheong et al. [81] found that DALL-E Mini was "liable to reproduce, exacerbate, and reinforce extant human social biases," with a tendency to produce images that reinforced racial and gender stereotypes. Luccioni et al. [80] evaluated major image generation models, including Stable Diffusion v.1.4 and DALL-E 2, finding that both perpetuated stereotypical associations between race or gender and occupation, and did not accurately reflect their real-world distribution of genders and races in particular occupations—a tendency which has also been reported for DALL-E Mini [81].

Bianchi et al. [82] show that bias—including heteronormative biases—appear in images of people not just through the gendered and racialized depictions of individual figures in these images, but also through the way in which background objects such as houses or cars are rendered, finding that these contextual factors reproduce problematic stereotypes about intersections between multiple factors of identity including socio-economic status, race, and gender. There is also the risk of further amplification of bias through prompt modifications, as Struppek et al. [85] found that the use of even a single homoglyph in a prompt led to the generation of harmful stereotypes about non-Western cultures [61].

Bias in image generation models is of particular concern given the potential downstream applications of their output, as well as their possible impact on culture, as will be described below. These can include, as noted by Ungless et al. [61], representational harms through problematic representations of groups or individuals, as well as allocational harms through the distribution of opportunities or other goods. Output from image generation models is reported to be used not only in, for example, the generation of stock images which may then be widely used in a range of settings, but also in highly consequential settings, such as the creation of sketches of suspects in crime based on eyewitness accounts [80]. There is also use in the medical domain, as Adams et al. [89] propose the use of DALL-E in generating images for education and training in radiology, even as their experiments with DALL-E 2 identified gender bias in the model.

As Fraser et al. [90] caution, the "worldview" of a particular model is not made visible to the user. Thus, "even when outputs appear biased, users' reactions are often to assume that the model made a one-off mistake and that its worldview truly aligns with their own." The authors suggest that "expert users" were aware of gender and racial biases in the models, but "novice users" often perceived the models as black boxed and thus were unaware of possible biased output, particularly as there are concerns that "the public's understanding of these models, from their development to their training data and their operational logics, is often limited" [91]. This suggests that users may also unwittingly amplify bias through their personal use or promotion of problematic images. Yet, as Hutchinson et al. [92] note, "[i]t is impossible to avoid ambiguous inputs," given the nature of the visual modality. This creates additional space for bias to be introduced into the generated images, particularly as "underspecification" [92] remains a key issue. Although, as Offert and Phan [93] find, there have been attempts on the part of providers of text-to-image models to covertly modify prompts inputted by users to achieve more diverse outputs, current models are overly dependent on users to input prompts calling for diversity, whereas, they would ideally utilize an "Ambiguity In, Diversity Out" approach [90, 92].

Finally, as Bianchi et al. [82] note, issues of bias are not easily mitigated, given that any understanding of an image is inherently subjective [17]. Seeking to control the "worldviews," including biases, reflected in output from a particular model "requires considerable engineering time and effort" [90]. Indeed, recent research has highlighted the way in which annotation can introduce bias into models [94], including through biases embedded in language itself [95], while even attempts to mitigate bias can in fact further reinforce it [96]. Offert and Phan [93] draw attention to the fundamental "whiteness" of the output produced by models

such as DALL-E 2, arguing that even more than bias, it is whiteness that is the "problem of foundation models" (see also [19, 57, 97, 98]). More broadly, Benjamin [19] and Joyce et al. [13] critique a tendency to seek out technical fixes to patch over issues of bias as a distraction from the urgent need to critically examine the systemic inequalities that produce bias on a societal scale. Thus, critical attention is needed to image generation models as potential site for the reproduction and amplification of biases.

3.4 Privacy

Given the opacity of training datasets, as delineated above, image generation models also raise privacy issues when training datasets include "privacy-sensitive information" [58] including facial images scraped from online sources. As discussed above, companies are disincentivized from making documentation on datasets—or the datasets themselves—publicly available for scrutiny, posing obstacles to addressing privacy concerns [58].

Diffusion models' image regurgitation tendencies run contrary to expectations that diffusion models would protect privacy [99]. For example, though the use of diffusion models to generate synthetic training data has been proposed to mitigate privacy concerns, the tendency of diffusion models to regenerate the original images would obviate the benefits of such an approach [99]. Moreover, privacy risks are heightened if captions for particular data samples in the training dataset are known, as users may prompt the generation of privacy-sensitive images, with membership leakage reported to be a "severe threat" in image generation [58, 99]. Struppek et al. [100] show that backdoor attacks on image generation models could potentially be used for the malicious creation of images of individuals, raising an urgent need for technical safeguards to protect against such risks [101].

These risks are heightened when the generated images are inherently sensitive, including facial images, which, as Stark [102] has argued, are uniquely "central to our personal identities and social lives." Generative models may be used to generate problematic or compromising images of individuals by utilizing this capacity to generate facial images. Similarly, Van Le et al. [101] draw attention to Textual Inversion and DreamBooth as two techniques commonly used to generate personalized output, but which are liable to be abused to generate harmful images. There have been reports of the presence of medical images of individuals in the LAION dataset, as well as of radiological images including CT, MRI, and ultrasound images in data for DALL-E 2 [89]. Ultimately, Zhang et al. [67] suggest that it is "nearly unfeasible to fully address" privacy concerns, though it is noteworthy that some efforts have been made to offset these

issues such as through OpenAI's efforts to deduplicate training data for DALL-E 2 to reduce images replicated in output [68, 103].

3.5 Impact on the informational environment

As has been reported for LLMs [104], the output of generative models can have a profound impact on the informational environment over the longer term, posing "severe threats to society" [105, 106]. Here, we define the informational environment broadly, including cases in which generated images are harmful or damaging. Thus, there is significant overlap between the issues identified here and those coming under the umbrella of privacy in the section above.

There are concerns particularly about the potential for highly realistic images and deepfakes generated through image generation models to harm the informational environment [105, 106]. This is in part as image generation establishes new norms through which "audio and video content cannot be taken at face value," thus threatening trust in the media, and exacerbating shifts towards post-truth societies [107]. There are early signs that these shifts may already be underway, as Oppenlaender et al. [108], in a small-scale, exploratory online survey of public perspectives on the use of image generation models, found that participants were already concerned about the "use of AI-generated imagery for opinion manipulation, fake news, and 'deep fakes [*sic*]'".

Furthermore, though their study focused on images synthesized by GANs, Nightingale and Farid [109] found that "synthetically generated faces are not just highly photorealistic, they are nearly indistinguishable from real faces and are judged more trustworthy," emphasizing that such artificial images carry "serious implications for individuals, societies, and democracies." They highlighted the risk that such images could be used for "nonconsensual intimate imagery, fraud, [and] disinformation campaigns" [109]. This includes harmful child sexual abuse material (CSAM), as noted by Bird et al. [53, 110] and by the Internet Watch Foundation [111] which recently found "20,254 AI-generated images… posted to one dark web CSAM forum in a one-month period." Though these risks are not entirely novel [112], they can be expected to be amplified through the ease of access to, the "democratization" [52] of, and the increasing public awareness of image generation systems, as "high-quality fakes now seem to come out from an assembly line." Similarly, Agarwal et al. [113] call for urgent attention to the risks of AI-synthesized imagery, which they argue can "pose a significant threat to our democracy, national security, and society," which is compounded by the likelihood that such synthesized images could be rapidly spread [107]. This occurs in part through the "liar's dividend," through

which "public figures caught in genuine recordings of misbehavior will find it easier to cast doubt on the evidence against them" [107]. Ricker et al. [106] "find that existing detectors, which claim to be universal, fail to effectively recognize images synthesized by different [diffusion models]," highlighting an urgent need for mitigation of these risks, including through mechanisms for improved detection (see also [112]), or mechanisms to make the provenance of AI-generated output clear, such as through the use of watermarks [74, 109].

The generation of more explicitly harmful images also poses a particular risk to the informational environment. Qu et al. [114] identify five kinds of unsafe images, including sexually explicit, violent, disturbing, hateful, and political images, finding through their study that just under 15% of all analyzed images generated through a text-to-image model could be classed into these categories, despite attempts to restrict the output of harmful content. In addition, Struppek et al. [100] showed that backdoor attacks on image generation models could be injected within two minutes and used to generate harmful images, with even "a single violent or explicit image" potentially causing harm to the viewer. This highlights the potential vulnerabilities that can be exploited to cause further harm to the informational environment, beginning at the level of the individual.

3.6 Impact on the cultural environment

Image generation models further pose risks to culture. Srinivasan and Uchino [115] identify three key cultural risks through the use of image generation models: first, as described above, the biases embedded in image generation models risk causing harm to the cultural environment; second, the models allow users to stereotype the styles of artists or movements, while failing to convey the intent of the original artists, thus threatening cultural heritage; and third, through the generation of inaccurate or unrepresentative images of historical events, they impede historical and cultural preservation. These issues are compounded by imbalances in the art styles drawn on in training data, such as an over-representation of Renaissance and other European art, or of modern art [115], with biases amplified by the Western contexts in which the models are developed [70].

The use of image generation models in art generation may limit creativity, by generating "predictable" [52] images, and impeding personalization. Yet, the longer-term impact of generative models on the cultural environment is often discussed with reference to photography; it transformed how art was produced and appreciated, suggesting that image generation models may, in a similar way, ultimately become a tool to enhance the work of artists [66]. These uses may include identifying reference images, speeding up

prototyping processes, reducing bias in art, and empowering artists to work with new software or with new media [7, 52]. Nonetheless, at present many artists are "perplexed and disoriented" [52] regarding how to implement AI tools in their art, and there is also a growing concern over the implications of the replacement of human artists by image generation models [116], as employment opportunities, particularly for freelance artists, may be lost if such models are utilized instead of humans [66]. This is problematic given that artists' works are used to train models, as described above, yet the profit from any such training remains with the companies providing the models, ultimately creating a situation in which companies are "profiting off the artistic endeavors of others, and possibly supplanting those human artisans" [66].

These impacts may extend to education in the area of the visual arts [7] and the potential impact on the next generation of artists is particularly of concern [117]. Oppenlaender et al. [108] found that participants in their study were concerned about the impact of generative models on culture, including concerns over a "decline in human creativity," suggesting that this is an issue which members of the public are already attuned to. However, Oppenlaender [118] has argued elsewhere that current cultural understandings of creativity are overly limited by a focus on product, and thus overlook the role of creativity which arises from the process of image generation using text-to-image models, which includes human involvement, thus complicating understandings of the implications of the generative models for the arts. A body of works considering the concept of "computational creativity" [119, 120] is also relevant here. For example, Jordanous [120] notes the existence of "[n]egative preconceptions about computational creativity [which] primarily exist because creativity is seen as something which is essentially human," though indicating that computers can, in fact, be creative.

In this vein it is also significant that Ragot et al. [121] found confirmation of a tendency to evaluate human-created artwork more highly than works created by generative models. They suggest that negative bias towards artwork generated by models may be due to intergroup bias, through which "AI may be anthropomorphized and thus considered as the outgroup," drawing on psychological research which suggests that individuals may have negative biases towards outgroups. Nonetheless, AI-generated art has been awarded prestigious prizes, including cases where the judges were unaware of the use of image generation models [117, 122].

3.7 Impact on the natural environment

In the context of growing concern over the breach of planetary boundaries [18, 123, 124], there is an urgent need to consider the extractive processes behind generative models and their environmental impact [18, 125, 126]. Van Wynsberghe [16] has drawn attention to concern over the sustainability of AI as an emergent wave of ethical concern. Brevini [126] has called for urgent action on "quantifying and reducing the environmental costs and damages of the current acceleration" of AI development. There has to date been relatively limited attention to the environmental costs of generative models. However, concerning LLMs, for example, Bender et al. [54] notably highlight environmental costs as one of the primary concerns in pursuing their development. Despite this, Rillig et al. [127] report that, even two years after the initial call from Bender et al. "the debate so far largely missed the potential implications of current and future LLM tools for the environment," a trend which we find to be similarly true for text-to-image models. Yet, Luccioni et al. [128] indicate that "[t]asks involving images are more energy- and carbon-intensive compared to discriminative tasks," with image generation as one of the most intensive of these tasks—highlighting the urgency of these considerations.

In 2018, OpenAI [129] announced that "the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.4-month doubling time," leading to an increase of 300,000 times since 2012—a figure which would have been 7 times if aligned with Moore's Law's two-year doubling period. This highlights the growing environmental burden arising from the training of large generative models, and particularly through increased carbon emissions. Utilizing the Machine Learning Impact Calculator [130] and based on an analysis of hardware, use times, cloud providers, and region, the estimated emissions from the development of Stable Diffusion v1 were equivalent to 15,000 kg of carbon dioxide [3].

It is noteworthy, however, that emissions are just one aspect of the environmental burden from image generation models, and little data has been made available about other aspects [18, 128, 130, 131]. These aspects are varied and include the use of significant amounts of freshwater to cool data centers, as reported for LLMs, even at a time when shortages of freshwater are growing increasingly worrisome, and major tech companies have committed to reducing their "water footprints" [132]. They also include the costs inherent in the needed hardware and infrastructure across their life cycles, ranging from the extractive processes involved in metals and minerals used in the systems such as rare earth elements, through to their end-of-life as e-waste [18, 133, 134]. These processes cause major harm directly and indirectly to both human and environmental health [133]. As increasingly advanced generative models are designed and implemented, there are suggestions [135] that new devices to operate these models will also need to

be concurrently made available– further heightening these environmental costs, and leading to increased e-waste, with its attendant impacts on health [18, 133].

4 Discussion

Thus, we have brought together the technical and social perspectives on the life cycle of image-generation models. Here, we situate the social issues arising from image generation models within their life cycle. We also compare these issues in image generation models with those of LLMs to argue that there is an urgent need for consideration of the risks of image generation models at minimum commensurate with the focus that has been given to LLMs.

4.1 Social issues in the model life cycle

This section examines the intersection of the technical and social axes, situating the social issues addressed above within the model life cycle, thus we review each issue cluster identified in the results above, and assess their relevance within the life cycle. As will be clear through the discussion below, each issue cluster is relevant across the model life cycle, though its relevance may be heightened in particular phases.

Data issues appear across multiple sub-phases of the model life cycle. They are intrinsic in current data acquisition procedures for text-to-image models. They also appear through model design, particularly when the models are trained and later refined. These issues include the involvement of human in model refinement under exploitative conditions. These issues are then inevitably reflected in the output from the models. For example, the lack of documentation of the content of datasets and the challenges in pre-emptively identifying and tackling risks means that the generated output contains problematic content. As discussed above, if these generated images are scraped and used in further training, these issues recur cyclically in the next generation of models.

Intellectual property issues likewise appear similarly across multiple sub-phases of the model life cycle. Issues arise both through the data which is used for the models, which may include works in violation of intellectual property rights, but can also appear in output, as when works generated by a text-to-image model are overly similar to existing works. Although as yet unreported, there may also be platform-based issues if platform owners skim and later utilize inputted prompts. These issues also extend into the deployment phase, as debates have arisen regarding the handling of intellectual property in prompt generation and input.

In relation to bias, the training data used often reflects structural and embedded social biases present in the sources from which the data is scraped. Then, these biases are embedded in training, and can also be reinforced through the refinement process particularly if humans are involved in this process and their own personal biases are reflected. Bias can also become pertinent in the input process, as Offert and Phan [93] identified a tendency in DALL-E 2 for prompts to be modified unbeknownst to the user in order to generate more diverse images. Once these images are outputted, bias risks being further reproduced through the use of these problematic images.

Privacy-sensitive images may be used in data acquisition and training, heightening the relevance of privacy concerns in this phase. Additional issues arise in the deployment phase if users utilize the text-to-image models to attempt to generate privacy-sensitive images, such as explicit images or images depicting specific individuals. Moreover, as with intellectual property issues above, there is a risk that platform owners skim prompts inputted by users. The spread of these generated images, and their potential re-use in further models pose additional privacy risks.

Issues in the informational environment can arise through data acquisition as there is no way to assess the accuracy or factual nature of the image-text pairs acquired. These issues then come to the fore in deployment, where the models can be used to intentionally or inadvertently generate images which misrepresent reality. Without watermarking or other indications through which to clarify the provenance of these images, there are key concerns about their detrimental impact on the informational environment at large once they are used and are in circulation.

As with intellectual property issues, cultural issues occur when culturally important images are scraped for use by the model. The issues similarly re-emerge in deployment and use of the images, such as when the generated images misrepresent culturally important events or works. They also pose broader cultural issues when the platforms are utilized to replace the work of human artists.

And finally, issues related to the natural environment are unique in this analysis as they emerge across all phases of the model life cycle. Each step in the model life cycle implicitly involves an environmental cost, as each piece of hardware and all infrastructure that is used in the development and deployment of text-to-image models create burdens and pressures on the natural environment. Though there remains a dearth of relevant research in this area, these are nonetheless critical considerations.

As is clear through the discussion above, the social issues of image generation models appear across the model life cycle. For this reason, it is essential that pre-emptive consideration of these issues be embedded throughout all

phases of the model life cycle. In Table 3 below, we propose a matrix which could be used in risk assessments of present and future image generation models at the intersection of the social and technical axes.

The table reflects the two axes of consideration in this paper—the issue clusters which arise, and the phases of the lifecycle at which they may arise. The matrix is designed to be used in one of two ways—one simpler, and one more advanced. First, it may be used to indicate with a binary yes/no whether or not a particular issue cluster may be relevant at a certain phase of the lifecycle. To this, an additional category can be added to indicate where issues are anticipated, even if not already reported. This allows for possible risks to be identified without any omissions.

As a more advanced application, the matrix may be used as a part of a risk assessment process, drawing on an adapted version of ISO/IEC guidelines for impact assessments [136]. This would involve setting out criteria for assessing the severity of a risk on a four-point scale, followed by criteria for assessing the likelihood of a risk on a separate four-point scale. A risk assessment score would be generated by multiplying the scores for severity and likelihood, and a cut-off value set, beyond which risk is deemed to be unacceptable.

Whichever method is used, the aim is to facilitate critical discussion in the development and evaluation process for image generation models, not only about where risks may arise, but also about where risks may be incompletely understood. Cells of the matrix which prove challenging to fill would suggest that there is insufficient information about a particular risk and caution is needed. For example, as many developers of image generation models do not provide complete information at present about their models, the matrix may be difficult to fill for particular models, indicating that models with low transparency should be used with great caution, if at all. Moreover, as discussed above, in the case of image generators generally, risks from each issue cluster are present across the lifecycle. This would likewise indicate that special caution is needed, and mitigations required for individual issue clusters.

4.2 Comparison with LLMs

Recent research such as by Bender et al. [54] and Weidinger et al. [12] have highlighted the risks associated with LLMs. Yet, despite their growing utilization and societal influence, relatively little comparable evaluation of image generation models has been conducted, though there are first movements in the field [10, 53]. Here, we compare the issues arising from LLMs, as reported in prior research, with the issues we have identified through this study, considering each of the seven issue clusters along the social axis, finding levels of risk across multiple issue clusters that meet or exceed those of LLMs described above.

Beginning with data issues, just as with LLMs, image generation models rely on massive, undocumented datasets, which means that both models and their datasets are black-boxed. This poses obstacles to the identification of possible sources of harm and to their remediation. Furthermore, though issues around attempts at content moderation have yet to receive as much attention as for LLMs [64], the documented trauma of content moderators [63] for other types of image-based content suggest that these issues can be expected to be similarly present for image generation models.

Intellectual property issues arise with both LLMs and image generation models, yet the spate of recent lawsuits around image generation models—which have yet to be matched by a comparable level of legal activity around LLMs—suggest that issues are heightened in this area. Indeed, though LLMs may potentially produce identical text to a copyrighted work, similarities on the level of style rather than exact replication may be harder to detect and more contestable when the medium is text; however, the similarities between images—particularly given diffusion models' tendency to regenerate training data—are more

Table 3 Proposed risk assessment matrix

Phase	Data	IP	Bias	Privacy	Information	Cultural	Natural
Data acquisition							
Image							
Text							
Filtering							
Model design							
Model type							
Training							
Refinement							
Deployment							
Input							
Platform							
Output							

readily identifiable, foregrounding intellectual property issues.

Bias is an issue which arises with both LLMs and image generation models. Yet, as discussed above, particularly with reference to work by Bianchi et al. [82], image generation models include contextual information beyond what is in a more straightforward medium such as text; thus, bias can be reproduced not only in the foregrounded object in an image, but also in the background context, and it can be difficult for users to provide sufficient specificity in prompts to avoid this reproduction of bias. Here, then, the potential for bias to manifest in output is amplified in comparison with LLMs.

Privacy issues arise similarly in LLMs and image generation models, as both may draw on personal information about individuals which has been posted online. In relation to LLMs, there are risks, for example, that individuals input sensitive information into LLMs, and that this data be leaked or otherwise reflected in output towards other users, as was reported for ChatGPT [137]. However, privacy risks are heightened in image generation models given that they risk the regurgitation of training data, including facial images, which are inherently sensitive. Of particular concern in relation to privacy are reports of sensitive medical images present in the datasets for image generation models.

The impact on the informational environment is particularly severe in the case of image generation models. Though it is well-documented that LLMs may be used for disinformation campaigns and to flood the informational environment with inaccurate or misleading information (e.g. [104]), images evoke visceral responses in their viewers [138], and are often even more likely to be taken for granted as true than are text (see [107]). In comparison with text, it is a relatively recent development that images can be faked, and thus the cultural sensitivity towards the need to question visual evidence is as yet underdeveloped, further heightening the risks in this area. Indeed, images are "remarkably slippery things, laden with multiple potential meanings, irresolvable questions, and contradictions" [17].

Similarly, the impact on the cultural environment is also notable in image generation models, as reported by [115]. The impact of image generation systems extends beyond just the future of art and imagery, to also encompass their impact on how the past is understood, if image generation models are used to generate images of historical events. Moreover, they risk bringing about fundamental changes to how art and creativity are produced and understood, and to the livelihoods of artists.

The impact of LLMs and image generation models on the natural environment hold similarities in that both are insufficiently understood, as there remains insufficient documentation of their impact across their life cycles, including both software and hardware. Given the relatively larger amount of data which is drawn on for LLMs in comparison with image generation models, there are indications that the impact on the natural environment may be more pronounced in the case of LLMs. Nonetheless, further research is urgently needed to better understand these impacts.

Thus, across the majority of these issue clusters, image generation models pose risks comparable to those of LLMs.

4.3 Recommendations

Although mitigations are outside the scope of this exploratory analysis, a number of recommendations arose from this interdisciplinary work and are included here. These are grouped into three categories: mitigations by design, mitigations post-hoc, and general recommendations. Mitigations by design are those which should be considered during the development phase of the models and involve pre-emptive consideration of the issues arising from each issue cluster. Mitigations post-hoc aim to mitigate already existing issues in the models. Finally, the general recommendations are applicable across the phases of the model life cycle.

4.3.1 Mitigations by design

a. To be aware of data imbalances and to increase diversity in datasets used for training generative models, to offset issues related to bias and the impact on the cultural environment. It is noteworthy that, although as discussed above, imbalances in text-to-image models' datasets have been reported [139, 140], there are insufficient resources to effectively address them at present.
b. To assess model performance not only in terms of accuracy or image quality, but also in terms of factors directly related to the issue clusters, such as bias or environmental impact.
c. To create standardized benchmarks to this end.

4.3.2 Mitigations post-hoc

a. To put increased effort into technical mitigation strategies such as bias mitigation known as machine unlearning [141]. However, techniques that target a single issue may not be optimal. For example, whereas machine unlearning can help with copyright and privacy issues, it may worsen the representation of minorities [142].
b. To make datasets and models publicly available to allow for external audits by individuals and institutions, to facilitate identifying and addressing emergent issues.

4.3.3 General recommendations

a. Mitigations should be undertaken with awareness of their possible impact on other issue clusters. As we have shown here, each phase of the model life cycle can involve multiple issues, with overlapping effects.

b. Given the limitations to a "soft" [143], researcher-driven approach to ethical technology development, we call for further regulation on image generation models that is grounded in currently observed and anticipated risks along the seven issue clusters identified. Although there are attempts within the global community to begin to regulate AI such as through the European Union's AI Act and the G7 Hiroshima AI Process, there has been notable unevenness in the consideration of these issue clusters, with significant attention to data and intellectual property clusters, with less concern over, for example, the impact on the natural environment. It is essential that future, comprehensive regulation address all seven of the issue clusters.

4.4 Final remarks

We have synthesized the literature on the social issues posed by image generation models, identifying seven key issue clusters: data issues, intellectual property, bias, privacy, and impacts on the informational, cultural, and natural environments. We have then intersected these considerations along the social axis with those along the technical axis, highlighting where these risks may arise along the model life cycle. Furthermore, through comparison with the literature on LLMs, this study has shown that the risks associated with image generation models at least match, if not exceed, the risks posed by LLMs, though image generation systems have received significantly less critical attention to date.

Given the rapid pace of development of generative models, the limited time frame and scope is an inherent limitation of this study, as well as the focus on English-language literature. This work should be updated and expanded to capture further developments in the field. Moreover, if the development and deployment of image generation models are to be pursued at the current frenetic pace, this work indicates that there is an urgent need to consider how these issues and risks can be mitigated. An in-depth exploration of existing mitigations for each of the identified issue clusters was beyond the scope of this study, but future work must examine this. To this end, we highlight here the merits of sociotechnical considerations grounded in interdisciplinary collaboration, drawing together the on-the-ground perspectives of technical experts with the insights of experts in the social sciences.

Author contributions **Amelia Katirai**: Conceptualization, Investigation, Methodology, Writing — original draft, Writing — review & editing; **Noa Garcia**: Conceptualization, Data curation, Investigation, Visualization, Writing– original draft, Writing– review & editing; **Kazuki Ide**: Conceptualization, Writing– review & editing, Investigation, Supervision; **Yuta Nakashima**: Conceptualization, Writing– original draft, Writing– review & editing; **Atsuo Kishimoto**: Conceptualization, Writing– review & editing, Supervision.

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. (2022). ArXiv Prepr ArXiv220406125

2. Dayma, B., Patil, S., Cuenca, P., Saifullah, K., Abraham, T., Le Khac, P., et al.: Dall· e mini. HuggingFacecom Httpshuggingface Cospacesdalleminidalle-Mini Accessed Sep 29. 2021; (2022)

3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–95. (2022)

4. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J., CogVideo: Large-scale pretraining for text-to-video generation via transformers. In: ICLR (2023)

5. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., et al.: Make-A-Video: Text-to-video generation without text-video data. In: ICLR (2022)

6. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., et al.: Imagen video: High definition video generation with diffusion models. ArXiv Prepr ArXiv221002303; (2022)

7. Dehouche, N., Dehouche, K.: What's in a text-to-image prompt? The potential of stable diffusion in visual arts education. Heliyon; (2023)

8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al.: Generative adversarial networks. Commun. ACM. **63**(11), 139–144 (2020)

9. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)

10. Bendel, O.: Image synthesis from an ethical perspective. AI Soc.;1–10. (2023)

11. Bird, C., Ungless, E., Kasirzadeh, A.: Typology of risks of generative text-to-image models. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. pp. 396–410. (2023)

12. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.S., Mellor, J., et al.: Taxonomy of Risks posed by Language Models. In: 2022 ACM Conference on Fairness, Accountability, and Transparency [Internet]. Seoul Republic of Korea: ACM; 2022 [cited 2023 Mar 20]. pp. 214–29. https://doi.org/10.1145/3531146.3533088

13. Joyce, K., Smith-Doerr, L., Alegria, S., Bell, S., Cruz, T., Hoffman, S.G., et al.: Toward a sociology of Artificial Intelligence: A call for Research on inequalities and Structural Change. Socius Sociol. Res. Dyn. World. **7**, 237802312199958 (2021)

14. Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L.A., Mateos-Garcia, J., et al.: Sociotechnical Safety Evaluation of Generative AI Systems [Internet]. arXiv; [cited 2024 Mar 27]. (2023). http://arxiv.org/abs/2310.11986

15. Huang, C., Zhang, Z., Mao, B., Yao, X.: An overview of artificial intelligence ethics. IEEE Trans. Artif. Intell. **4**(4), 799–819 (2022)

16. van Wynsberghe, A., Sustainable, A.I.: AI for sustainability and the sustainability of AI. AI Ethics. **1**(3), 213–218 (2021)

17. Crawford, K., Paglen, T., Excavating, A.I.: The politics of images in machine learning training sets. AI Soc. **36**(4), 1105–1116 (2021)

18. Crawford, K.: Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press, New Haven (2021)

19. Benjamin, R.: Race after Technology: Abolitionist Tools for the new Jim code. Oxford University Press (2020)

20. Gebru, T.: Race and Gender. In: Dubber MD, Pasquale F, Das S, editors. The Oxford Handbook of Ethics of AI [Internet]. 2020 [cited 2022 Nov 4]. pp. 252–69. https://academic.oup.com/edited-volume/34287/chapter-abstract/290662826?redirectedFrom=fulltext

21. Anderson, J., Laracy, F.J.R., Marlowe, T.A., Multi-Disciplinary: Anal. Catholic Social Teach. Implications Eng. Technol.;**18**(6). (2020)

22. McStay, A., Emotional, A.I., Ethics, Spice, J.: Contributing community, wholeness, sincerity, and heart. Philos. Technol. **34**(4), 1781–1802 (2021)

23. Coeckelbergh, M.: AI Ethics. The MIT Press, Cambridge (2020)

24. KIROVA, V.D., Ku, C., Laracy, J., Marlowe, T.: The Ethics of Artificial Intelligence in the era of generative AI. J. Syst. Cybern Inf. **21**(4), 42–50 (2023)

25. Mansimov, E., Parisotto, E., Ba, J.L., Salakhutdinov, R.: Generating images from captions with attention. In: ICML. (2016)

26. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., et al.: Zero-shot text-to-image Generation, pp. 8821–8831. In: ICML. PMLR (2021)

27. Ordonez, V., Kulkarni, G., Berg, T.: Im2text: Describing images using 1 million captioned photographs. Adv. Neural Inf. Process. Syst.;**24**. (2011)

28. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Trans. Assoc. Comput. Linguist. **2**, 67–78 (2014)

29. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., et al.: Microsoft COCO captions: Data collection and evaluation server. (2015). ArXiv Prepr ArXiv150400325

30. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**, 32–73 (2017)

31. Sharma, P., Ding, N., Goodman, S., Soricut, R., Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In: Proceedings of ACL. (2018)

32. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., et al.: Nocaps: Novel object captioning at scale. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8948–57. (2019)

33. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: Computer Vision–ECCV.: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer; 2020. pp. 742–58. (2020)

34. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–68. (2021)

35. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M., Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2443–9. (2021)

36. Desai, K., Kaul, G., Aysola, Z., Johnson, J., RedCaps: Web-curated image-text data created by the people, for the people. ArXiv Prepr ArXiv211111431; (2021)

37. Schuhmann, C., Kaczmarczyk, R., Komatsuzaki, A., Katta, A., Vencu, R., Beaumont, R., et al.: NeurIPS Workshop Datacentric AI. Jülich Supercomputing Center (2021). LAION-400 M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs

38. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., et al.: LAION-5B: An open large-scale dataset for training next generation image-text models. ArXiv Prepr ArXiv221008402; (2022)

39. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics & Image Processing. IEEE; pp. 722–9. (2008)

40. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. California Institute of Technology (2011). Report No.: CNS-TR-2011-001

41. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al.: Microsoft COCO: Common objects in context. In: ECCV. Springer; pp. 740–55. (2014)

42. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative Adversarial text to Image Synthesis, pp. 1060–1069. In: ICML. PMLR (2016)

43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. In: NeurIPS (2017)

44. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. ICLR; (2014)

45. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al.: Generative adversarial nets. In: NeurIPS (2014)

46. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML p. 2256–2265. (2015)

47. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: NeurIPS (2019)

48. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS p. 6840–6851. (2020)

49. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., et al.: ImageReward: Learning and evaluating human preferences for text-to-image generation. (2023). ArXiv Prepr ArXiv230405977

50. Hugging Face: Stable Diffusion text-to-image fine-tuning [Internet]. [cited 2024 May 27]. https://huggingface.co/docs/diffusers/v0.13.0/en/training/text2image

51. Zhang, C., Zhang, C., Zheng, S., Qiao, Y., Li, C., Zhang, M., et al.: A complete survey on generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 all you need? (2023). ArXiv Prepr ArXiv230311717

52. Ko, H.K., Park, G., Jeon, H., Jo, J., Kim, J., Seo, J.: Large-scale text-to-image generation models for visual artists' creative works. In: IUI. pp. 919–33. (2023)

53. Bird, C., Ungless, E.L., Kasirzadeh, A.: Typology of Risks of Generative Text-to-Image Models. (2023)

54. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 列. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency [Internet]. Virtual Event Canada: ACM; 2021 [cited 2023 Mar 31]. pp. 610–23. https://doi.org/10.1145/3442188.3445922

55. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qual. Res. Psychol. **3**(2), 77–101 (2006)

56. Paullada, A., Raji, I.D., Bender, E.M., Denton, E., Hanna, A.: Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns. **2**(11), 100336 (2021)

57. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., et al.: Datasheets for datasets. Commun. ACM. **64**(12), 86–92 (2021)

58. Wu, Y., Yu, N., Li, Z., Backes, M., Zhang, Y.: Membership Inference Attacks Against Text-to-image Generation Models. ArXiv Prepr ArXiv221000968.; (2022)

59. Schramowski, P., Brack, M., Deiseroth, B., Kersting, K.: Safe latent diffusion: Mitigating Inappropriate Degeneration in Diffusion models. (2023). ArXiv Prepr ArXiv221105105

60. Monea, A.: The Digital Closet. MIT Press (2022)

61. Ungless, E.L., Ross, B., Lauscher, A.: Stereotypes and Smut: The (Mis) representation of Non-cisgender Identities by Text-to-Image Models. ArXiv Prepr ArXiv230517072.; (2023)

62. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., et al.: GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, pp. 16784–16804. In: ICML. PMLR (2022)

63. Gray, M.L., Suri, S.: Ghost work: How to stop Silicon Valley from Building a new Global Underclass. Eamon Dolan Books (2019)

64. Perrigo, B.: Exclusive: OpenAI used Kenyan workers on Less Than 2 per hour to make ChatGPT less toxic. Vice (2023)

65. Vyas, N., Kakade, S., Barak, B.: Provable copyright protection for generative models. ArXiv Prepr ArXiv230210870.; (2023)

66. Ghosh, A., Fossas, G.: Can There be Art Without an Artist? In: NeurIPS Workshop: Human Evaluation of Generative Models. (2022)

67. Zhang, E., Wang, K., Xu, X., Wang, Z., Shi, H.: Forget-Me-Not: Learning to forget in text-to-image Diffusion models. (2023). ArXiv Prepr ArXiv230317591

68. Nichol, A.: DALL-E 2 pre-training mitigations. OpenAI Blog (2022)

69. Vincent, J.: Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement. The Verge. The Verge. (2023)

70. Cetinic, E.: The Myth of Culturally Agnostic AI Models. ArXiv Prepr ArXiv221115271.; (2022)

71. Weatherbed, J.: How to keep your art out of AI generators. The Verge [Internet]. Feb 8 [cited 2024 May 27]; (2024). https://www.theverge.com/24063327/ai-art-protect-images-copyright-generators

72. Shan, S., Ding, W., Passananti, J., Wu, S., Zheng, H., Zhao, B.Y., Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. In: 2024 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society; pp. 212–212. (2024)

73. United States Copyright Office: Copyrightable Authorship: What Can Be Registered. (2021)

74. Chesterman, S.: AI-generated content is taking over the world. But who owns it? NUS Law Work Pap No 2023002.; (2023)

75. Fenwick, M., Jurcys, P.: Originality and the future of copyright in an age of generative AI. Available SSRN 4354449.; (2023)

76. Guadamuz, A.: Do androids dream of electric copyright? Comparative analysis of originality in artificial intelligence generated works. Intellect. Prop. Q.; (2017)

77. Shen, X., Qu, Y., Backes, M., Zhang, Y.: Prompt Stealing Attacks Against Text-to-Image Generation Models. ArXiv Prepr ArXiv230209923.; (2023)

78. Gervais, D.J.: The machine as author. Iowa Rev.;105:2053. (2019)

79. Epstein, Z., Levine, S., Rand, D.G., Rahwan, I.: Who gets credit for AI-Generated art? iScience.;**23**(9):101515. (2020)

80. Luccioni, A.S., Akiki, C., Mitchell, M., Jernite, Y.: Stable Bias: Analyzing Societal representations in Diffusion models. (2023). ArXiv Prepr ArXiv230311408

81. Cheong, M., Abedin, E., Ferreira, M., Reimann, R.W., Chalson, S., Robinson, P., et al.: Investigating gender and racial biases in DALL-E Mini images. PNAS; (2023)

82. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., et al.: Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. (2022). ArXiv Prepr ArXiv221103759

83. Fraser, K.C., Kiritchenko, S., Nejadgholi, I.: A Friendly Face: Do text-to-Image systems Rely on stereotypes when the Input is Under-Specified? (2023). ArXiv Prepr ArXiv230207159

84. Steed, R., Caliskan, A.: Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency [Internet]. Virtual Event Canada: ACM; 2021 [cited 2023 Apr 7]. pp. 701–13. https://doi.org/10.1145/3442188.3445932

85. Struppek, L., Hintersdorf, D., Kersting, K.: The biased artist: Exploiting cultural biases via homoglyphs in text-guided image generation models. ArXiv Prepr ArXiv220908891.; (2022)

86. Seshadri, P., Singh, S., Elazar, Y.: The Bias amplification Paradox in text-to-image generation. (2023). ArXiv Prepr ArXiv230800755

87. Cho, J., Zala, A., Bansal, M.: DALL-Eval: Probing the reasoning skills and social biases of text-to-image generative transformers. ArXiv Prepr ArXiv220204053.; (2022)

88. Bansal, H., Yin, D., Monajatipoor, M., Chang, K.W.: How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? ArXiv Prepr ArXiv221015230.; (2022)

89. Adams, L.C., Busch, F., Truhn, D., Makowski, M.R., Aerts, H.J., Bressem, K.K.: What does DALL-E 2 Know about Radiology? J. Med. Internet Res. **25**, e43110 (2023)

90. Fraser, K.C., Kiritchenko, S., Nejadgholi, I.: Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems. ICCV Accept; (2023)

91. Munn, L., Magee, L., Arora, V., Unmaking, A.I., Imagemaking: A methodological toolkit for critical investigation. (2023). ArXiv Prepr ArXiv230709753

92. Hutchinson, B., Baldridge, J., Prabhakaran, V.: Underspecification in Scene Description-to-Depiction Tasks. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) [Internet]. Online only: Association for Computational Linguistics; pp. 1172–84. (2022). https://aclanthology.org/2022.aacl-main.86

93. Offert, F., Phan, T.A., Sign That, Spells: DALL-E 2, Invisual images and the racial politics of Feature Space. (2022). ArXiv Prepr ArXiv221106323

94. Andrews, J.T., Zhao, D., Thong, W., Modas, A., Papakyriakopoulos, O., Nagpal, S., et al.: Ethical considerations for collecting human-centric image datasets. ArXiv Prepr ArXiv230203629; (2023)

95. Wolfe, R., Caliskan, A.: American == white in multimodal language-and-image AI. In: AIES. pp. 800–12. (2022)

96. Hirota, Y., Nakashima, Y., Garcia, N.: Quantifying societal bias amplification in image captioning. In: CVPR p. 13450–13459. (2022)

97. Katz, Y.: Artificial Whiteness: Politics and Ideology in Artificial Intelligence. Columbia University, New York (2020)

98. Cave, S., Dihal, K.: The whiteness of AI. Philos. Technol. **33**(4), 685–703 (2020)

99. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., et al.: Extracting training data from diffusion models. ArXiv Prepr ArXiv230113188; (2023)

100. Struppek, L., Hintersdorf, D., Kersting, K.: Rickrolling the artist: Injecting Invisible backdoors into text-guided image generation models. (2022). ArXiv Prepr ArXiv221102408

101. Van Le, T., Phung, H., Nguyen, T.H., Dao, Q., Tran, N., Tran, A.: Anti-DreamBooth: Protecting users from personalized text-to-image synthesis. ArXiv Prepr ArXiv230315433; (2023)

102. Stark, L.: Facial recognition is the plutonium of AI. XRDS Crossroads ACM Mag Stud. **25**(3), 50–55 (2019)

103. Chen, C., Fu, J., Lyu, L.: A pathway towards responsible AI generated content. (2023). ArXiv Prepr ArXiv230301325

104. OpenAI: GPT-4 System Card. (2023)

105. Sha, Z., Li, Z., Yu, N., Zhang, Y.: DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. (2023). ArXiv Prepr ArXiv221006998

106. Ricker, J., Damm, S., Holz, T., Fischer, A.: Towards the detection of Diffusion Model deepfakes. (2023). ArXiv Prepr ArXiv221014571

107. Chesney, R., Citron, D.: Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. Foreign Aff. **98**, 147 (2019)

108. Oppenlaender, J., Visuri, A., Paananen, V., Linder, R., Silvennoinen, J.: Text-to-image generation: Perceptions and realities. (2023). ArXiv Prepr ArXiv230313530

109. Nightingale, S.J., Farid, H.: AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proc. Natl. Acad. Sci. **119**(8), e2120481119 (2022)

110. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes [Internet]. arXiv; 2021 [cited 2023 Apr 10]. http://arxiv.org/abs/2110.01963

111. Internet Watch Foundation. How AI is being abused to create child sexual abuse imagery [Internet]: (2023). https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/

112. Verdoliva, L.: Media forensics and deepfakes: An overview. IEEE J. Sel. Top. Signal. Process. **14**(5), 910–932 (2020)

113. Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting World Leaders Against Deep Fakes. In: CVPR Workshops. p. 38. (2019)

114. Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., Zhang, Y.: Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. ArXiv Prepr ArXiv230513873; (2023)

115. Srinivasan, R., Uchino, K.: Biases in generative art: A causal look from the lens of art history. In: FAccT p. 41–51. (2021)

116. Shan, S., Cryan, J., Wenger, E., Zheng, H., Hanocka, R., Zhao, B.Y.: GLAZE: Protecting artists from style mimicry by text-to-image models. ArXiv Prepr ArXiv230204222; (2023)

117. Vimpari, V., Kultima, A., Hämäläinen, P., Guckelsberger, C.: An adapt-or-die type of Situation: Perception, Adoption, and Use of Text-To-Image-generation AI by Game Industry professionals. (2023). ArXiv Prepr ArXiv230212601

118. Oppenlaender, J.: The Creativity of Text-to-Image Generation. In: Proceedings of the 25th International Academic Mindtrek Conference [Internet]. Tampere Finland: ACM; 2022 [cited 2023 Apr 26]. pp. 192–202. https://doi.org/10.1145/3569219.3569352

119. Colton, S., Charnley, J., Pease, A.: Computational creativity theory: The FACE and IDEA descriptive models. In: Proc. Int. Conf. Comp. Creativity (2011)

120. Jordanous, A.: Evaluating computational creativity: a standardised procedure for evaluating creative systems and its application [PhD Thesis]. University of Sussex; (2013)

121. Ragot, M., Martin, N., Cojean, S.: AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence? In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems [Internet]. Honolulu HI USA: ACM; 2020 [cited 2023 Apr 26]. pp. 1–10. https://doi.org/10.1145/3334480.3382892

122. Seymour, T.: The camera never lies? Creator of AI image rejects prestigious photo award. The Art Newspaper [Internet]. 2023 [cited 2023 May 4]. Available from: The camera never lies? Creator of AI image rejects prestigious photo award

123. Wallace-Wells, D.: The Uninhabitable Earth. Tim Duggan Books, New York (2019)

124. Richardson, K., Steffen, W., Lucht, W., Bendtsen, J., Cornell, S.E., Donges, J.F., et al.: Earth beyond six of nine planetary boundaries. Sci. Adv. **9**(37), eadh2458 (2023)

125. Dhar, P.: The carbon impact of artificial intelligence. Nat. Mach. Intell. **2**(8), 423–425 (2020)

126. Brevini, B.: Is AI good for the Planet? Polity, Cambridge (2021)

127. Rillig, M.C., Ågerstrand, M., Bi, M., Gould, K.A., Sauerland, U.: Risks and benefits of large language models for the environment. Environ. Sci. Technol. **57**(9), 3464–3466 (2023)

128. Luccioni, A.S., Jernite, Y., Strubell, E.: Power hungry processing: Watts driving the cost of ai deployment? ArXiv Prepr ArXiv231116863.; (2023)

129. Amodei, D., Hernandez, D.: AI and compute. OpenAI Blog. (2018)

130. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the Carbon Emissions of Machine Learning. In: NeurIPS Workshop: Tackling Climate Change with Machine Learning. (2019)

131. Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O., Green, A.I.: (2019). ArXiv Prepr ArXiv190710597

132. Li, P., Yang, J., Islam, M.A., Ren, S., Making, A.I.: Less thirsty: Uncovering and addressing the Secret Water Footprint of AI models. (2023). ArXiv Prepr ArXiv230403271

133. Parvez, S.M., Jahan, F., Brune, M.N., Gorman, J.F., Rahman, M.J., Carpenter, D., et al.: Health consequences of exposure to e-waste: An updated systematic review. Lancet Planet. Health. **5**(12), e905–e920 (2021)

134. Pitron, G.: The Rare Metals War: The dark side of Clean Energy and Digital Technologies. Scribe (2020)

135. The Economist: So long iPhone, Generative AI needs a new device. The Economist [Internet]. https://www.economist.com › business › 2023/10/05

136. ISO/IEC 29134:2023 Information technology — Security techniques — Guidelines for privacy impact assessment [Internet]: (2023). https://www.iso.org/standard/86012.html

137. OpenAI: March 20 ChatGPT outage: Here's what happened. (2023)

138. Ewbank, M.P., Barnard, P.J., Croucher, C.J., Ramponi, C., Calder, A.J.: The amygdala response to images with impact. Soc. Cogn. Affect. Neurosci. **4**(2), 127–133 (2009)

139. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: Misogyny, pornography, and malignant stereotypes. ArXiv Prepr ArXiv211001963; (2021)

140. Garcia, N., Hirota, Y., Wu, Y., Nakashima, Y.: Uncurated Image-Text Datasets: Shedding Light on Demographic Bias. (2023)

141. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C.A., Jia, H., Travers, A., Zhang, B., et al.: Machine unlearning. In: 2021 IEEE Symposium on Security and Privacy (SP). IEEE; pp. 141–59. (2021)

142. Koch, K., Soll, M.: No Matter How You Slice It: Machine Unlearning with SISA Comes at the Expense of Minority Classes. In: 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). pp. 622–37. (2023)

143. van Maanen, G.A.I., Ethics: Ethics washing, and the need to Politicize Data Ethics. Digit. Soc. **1**(2), 9 (2022)