Typology of risks of generative text-to-image models

Citation for published version:
Bird, C, Ungless, E & Kasirzadeh, A 2023, Typology of risks of generative text-to-image models. in F Rossi, S Das, J Davis, K Firth-Butterfield & A John (eds), *AIES '23: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society.* Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery (ACM), New York, pp. 396–410, Sixth AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society, Montréal, Quebec, Canada, 8/08/23. https://doi.org/10.1145/3600211.3604722

Digital Object Identifier (DOI):
10.1145/3600211.3604722

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
AIES '23

Typology of Risks of Generative Text-to-Image Models

CHARLOTTE BIRD*, School of Informatics
University of Edinburgh, Scotland
EDDIE L. UNGLESS*, School of Informatics
University of Edinburgh, Scotland
ATOOSA KASIRZADEH, Alan Turing Institute
University of Edinburgh, Scotland

This paper investigates the direct risks and harms associated with modern text-to-image generative models, such as DALL-E and Midjourney, through a comprehensive literature review. While these models offer unprecedented capabilities for generating images, their development and use introduce new types of risk that require careful consideration. Our review reveals significant knowledge gaps concerning the understanding and treatment of these risks despite some already being addressed. We offer a taxonomy of risks across six key stakeholder groups, inclusive of unexplored issues, and suggest future research directions. We identify 22 distinct risk types, spanning issues from data bias to malicious use. The investigation presented here is intended to enhance the ongoing discourse on responsible model development and deployment. By highlighting previously overlooked risks and gaps, it aims to shape subsequent research and governance initiatives, guiding them toward the responsible, secure, and ethically conscious evolution of text-to-image models.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; *Text input*; • **Applied computing** → *Media arts*; • **Social and professional topics** → User characteristics.

Additional Key Words and Phrases: Generative AI, Generative models, Text-to-Image models, Responsible AI, AI ethics, AI safety, AI governance, AI risks

Forthcoming in Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2023)

1 INTRODUCTION

In recent years, significant progress has been made in developing large language models and related multi-modal generative models, such as text-to-image models. We will collectively refer to these models as "generative models."[1] Generative models process and combine information from various modalities, including visual, textual and auditory data. The range of applications for generative models spans multiple fields. In entertainment, they can generate realistic-looking images or movie characters [44, 151]. In advertising, these models can be employed to create personalized ad content [26, 42]. They can aid scientific research by simulating complex systems or hypothesizing about empirical phenomena [3, 12, 18]. In education, they can facilitate personalized learning, catering to unique needs and learning pace of each student [7, 159].

While introducing exciting opportunities, generative models also pose risks. These risks have attracted significant scrutiny from the AI ethics and safety community. The social and ethical risks of large language models, along with the text-to-text technologies they support, have been intensely discussed within the literature [13, 168]. For instance, it is widely acknowledged that existing language technologies can potentially cause harm by producing inappropriate, discriminatory, or harmful content [45, 47, 63, 167, i.a.], or that the alignment of language technologies with beneficial human values is far from a straight forward task [6, 51, 85]. This paper extends this line of inquiry from language models to text-to-image generative models, examining potential risks and harms resulting from their development and use. To identify and illuminate these

[*] Equal contribution
[1] These models are also known by some researchers as foundation models [24].

Risk	Stakeholders	Harm	Anticipated	Observed
Discrimination and Exclusion				
Cultural and racial bias	Users, Affected	Representational harm	[79, 105]	[15, 37, 177]
Gender & sexuality bias	Users, Affected	Representational harm	[79, 105, 112]	[15, 37, 157]
Class bias	Users, Affected	Representational harm	-	[15]
Disability bias	Users, Affected	Representational harm	-	[15]
Loss of work for creatives	Sources, Users	Financial loss	[64, 112, 119]	[54]
Religious bias, ageism	Users, Affected	Representational harm	-	-
Dialect bias	Users	Allocative harm, repr. harm	-	-
Pre-release moderation	Developers, Affected	Psychological harm	-	-
Job replacement	Affected, Regulators	Financial loss, Emotional harm	[54]	[41, 117]
Harmful Misuse				
Sexual images	Subjects, Users, Affected	Repr. harm, emot. harm, fin. loss	[79, 105]	[177]
Sexualising images of children	Subjects, Users, Affected, Regulators	Emotional harm	-	[174]
Violent or taboo content	Developers, Users, Affected	Emotional harm, incite violence	[28, 79, 105, 149]	-
Privacy infringement	Sources, Subjects, Regulators	Privacy loss	-	[35]
Copyright infringement	Users, Sources, Regulators	Financial loss	-	[35, 147, 163]
Cybersecurity Threats	Sources, Subjects, Regulators	Repr. harm, security loss	-	-
Misinformation and Disinformation				
Likeness reproduction	Subjects, Users, Affected	Repr. harm, emotional harm	[82, 106, 107, 115, 134]	[147]
Misleading harmful content	Users, Affected	Repr. harm, emotional harm	[14, 27, 29, 57, 61, 81, 113, 121, 153]	[115, 172]
Fraud and scams	Users, Affected	Emotional harm, financial loss	[14, 89, 103, 121, 162]	-
Detection and classification bias	Developers, Subjects, Users, Affected	Allocative harm	-	[100, 110, 125, 128, 175]
Polarisation	Users, Affected	Repr. harm, incite violence	[4, 14, 29, 39, 65]	[100]
Miscommunication	Developers, Users, Affected	Allocative harm, loss of trust	[27, 79, 177]	-
Soco-political Instability	Users, Affected, Regulators	Loss of trust, incite violence	[5, 10, 30, 39, 158, 171, 171?]	[100]

Table 1. Risk Typology. We provide detailed analysis in Section 4.

risks, we perform a comprehensive review of literature related to text-to-image (TTI) models. In particular, we conduct an initial search using 8 seed papers, supplementing with manual search (our search methodology is detailed in Appendix A). Collected papers are analysed for immediate risks, stakeholders, and empirical investigations.

Our systematic examination yields a typology of risks associated with state-of-the-art TTI models, such as DALL-E 2 [129]. Our findings are summarized in Table 1. Our typology and discussion analysis are limited to immediate risks, inspired by a taxonomy from Weidinger et al. [167]. Our typology is divided into three key categories: I. Discrimination and Exclusion; II. Harmful Misuse; III. Misinformation and Disinformation. We recognize that these categories are not mutually exclusive. However, defining distinct categories enables clearer understanding and supports the implementation of more robust mitigation strategies.

Our typology is further refined by identifying the stakeholders involved in the development and use of these systems. Inspired by the probing question from Blodgett et al. [21]: "How are social hierarchies, language ideologies, and NLP systems co-produced?", we interlace this concern into our research and typology formulation. This process helps us to illustrate how the technologies supported by TTI models can reinforce existing social hierarchies via stakeholder identification.

We adopt the stakeholder categories of developers, users, regulators and affected parties from Langer et al. [93]. We use "affected parties" referring to those influenced by the output of these models. We further extend the categorization by introducing "data sources" and "data subjects" – individuals or entities who generate and/or appear in the images used to train TTI models. Additionally, we ascribe the nature of potential harm, such as representational or allocative [9], to the identified stakeholders. We also touch upon risks of harm to environment [13, 112].

To organize the literature, we propose a practical distinction between two types of risks: "anticipated" and "observed." The former refers to risks that are primarily predicted by researchers due to their expertise and familiarity with the field. The latter, on the other hand, are risks that have been empirically investigated, providing insights into the potential magnitude of harm. This classification underscores the need for comprehensive empirical investigations into many of the identified risks. With this distinction in mind, we highlight several risks that, to our knowledge, have not yet been adequately discussed. We further contribute with an analysis of the challenges posed by proposed mitigation strategies (in 5) and an identification of open questions, supplemented by suggestions for policy change (in 6). Finally, we advocate for enhanced collaboration among researchers, system developers, and policymakers. Through our categorisation and discussion, our intention is to foster a better understanding of the potential futures – both positive and negative – of TTI models, and by extension, other generative models.

2 GENERATIVE TEXT-TO-IMAGE MODELS

A TTI model is a type of generative neural network designed to synthesise images based on textual prompts [131]. When given a prompt, the model generates an image that, in some sense, visually represents the information in the text. TTI systems typically leverage a combination of natural language processing (NLP) and computer vision techniques to produce images. The NLP component extracts relevant information such as objects, attributes, and relationships from the text, while the computer vision component generates an image based on this information.

Various generative architectures have shown promise in image synthesis tasks [59]. These include flow-based models [49], auto-regressive models [118] and variational autoencoders [90]. However, the advent of generative adversarial networks (GAN) [68] marked a significant acceleration in the capabilities of generative models.

A typical TTI GAN employs two types of deep neural networks – a generator and a discriminator. The generator synthesizes an image from a text input, while the discriminator evaluates the generated image, determining its authenticity. Through adversarial training, the generator refines its ability to create increasingly realistic images. The introduction of transformer architecture in 2017 spurred substantial progress in NLP [160], subsequently extending to vision tasks as evidenced by early versions of DALL-E. Additionally, CLIP [128], a model that learns visual concepts from natural language supervision, became pivotal in image generation tasks.

Diffusion models [145], which define a Markov chain parameterized by deep neural networks to reverse noisy data and sample from a desired data distribution, have recently achieved state-of-the-art results in image synthesis [48, 76, 134, 148]. The success of these models has stimulated a rapid proliferation of popular and open-source diffusion models, which are the subject of many of the papers in this taxonomy.

3 STAKEHOLDERS AND POWER DYNAMICS

A comprehensive discussion of stakeholders, emphasizing their relative power, is crucial for understanding the associated risks. As various researchers have articulated, it is essential to underscore power inequities by considering what might be absent from a dataset [62, 102]. We build upon this observation, and various other insights on the relations between power structures and socio-technical algorithmic systems [21, 84, 86], structuring our analysis around the inclusion or exclusion of various groups in the development and deployment of these models. In Table 1 and Section 4, we pinpoint six categories of stakeholders most likely to be impacted by the risks we identify: system developers, data sources, data subjects, users, affected parties, and regulators.

3.1 System Developers

Developing state-of-the-art TTI systems requires vast compute and storage capabilities. Consequently, development is dominated by actors who have such access, such as companies in the Global North and China. These tend to be primarily concentrated within a small group of for-profit companies and well-funded academic institutions (e.g. OpenAI, Meta, Stability AI, Google, DeepMind, Midjourney). Companies like Hugging Face are making efforts towards open-access TTI systems. However, it still remains unclear how these models compare competitively with for-profit models.

This concentration of resources can lead to a lack of diverse perspectives in the data curation and model development teams, which can result in the exacerbation of specific biases in the training data [170]. As a result, source and output images that reflect only the hegemonic perspective might go unnoticed, as those curating the data or developing the models are often blinkered by their own experiences. For instance, Bianchi et al. [15] and Yu et al. [177] found models reflected Western culture in their output, for example Western dining, wedding and clothing practices; and "couples" and "families" were exclusively heterosexual.

3.2 Data Sources

Current data collection methodologies often deny content creators the opportunity to provide consent [64] or be acknowledged as "collaborators" [144]. Furthermore, the widespread issue of inadequate curation in large datasets contributes to a multitude of problems [19] .[2] It results in opaque attributions, makes output reasoning convoluted, and complicates efforts towards harm reduction [19].

Certain TTI systems have been shown to replicate images from their training data, which can be thought of as "Digital Forgery" [147]: artists may find that models trained on their images produce near identical copies. Further, popular datasets such as ImageNet, CelebA, COCO, and LAION have been criticized for issues related to attribution and consent [20, 64]. These concerns have even prompted legal actions by creators and stock image websites against companies that deploy such technologies [31, 32, 173].

3.3 Data Subjects

The concern that "data available online may not have been intended for such usage" is significant [35]. While much of the public discourse around TTI systems has concentrated on copyright issues regarding training datasets, we bring attention to the problem of image subjects' consent, including situations of conflicting consent [88, 92].

The matter of image reproduction must be contemplated within the scope of privacy [147]. This concern applies to instances such as the unauthorized use of celebrity images or pornographic depictions of sex workers. While the focus often centers on the harm incurred by exposure to explicit content, the potential negative impact on the subjects of these images should not be overlooked. Explicit content is prevalent in many datasets, and users frequently retrain models to generate specific explicit content. However, some subjects of these images, such as sex workers, are not adequately considered in these discussions (though c.f. Birhane and Prabhu [19]).

3.4 Users

Before discussing typical users, we highlight that access to TTI models can be exclusionary. Commercial models often preclude certain territories, and successful use of these systems requires fluency in the input language (matching the dialect of the training data), or access to an accurate translation tool. We delve deeper into these issues further in Section 6.

TTI systems can serve as powerful tools for professionals in fields such as design, advertising, and art [36, 109, 112, 141]. They represent fresh avenues of exploration for creative individuals [38, 119, 119, 135], and can offer accessible resources for a wider audience [177], even holding potential to "democratise" art [112, 119]. The fact that Stable Diffusion boasts ten million daily active users [56] testifies to the public's keen interest in leveraging TTI models for their personal entertainment.

[2] Inadequate curation can mean that the data may contain inaccuracies, bias, or irrelevant information, all of which can propagate into AI systems trained on such data, leading to unreliable or potentially harmful outcomes.

On the flip side, TTI systems can be used for malicious purposes. In the realm of misinformation and disinformation, players such as hyper-partisan media, authoritarian regimes, state disinformation actors, and cyber-criminals have been identified as potential malicious users [4, 5, 14]. “Information operations” [107] are broadly acknowledged as a malicious use case. Additionally, Paris and Donovan [121] have identified a subset of enthusiasts, both unskilled and skilled hobbyists, who create harmful content, a substantial portion of which is pornographic. This exploitative content often gains viral attention [2].

3.5 Affected Parties

This section highlights both direct and indirect stakeholders who may be impacted by TTI systems.

Creatives. TTI systems can empower creatives by expanding their toolkit, but it is crucial to note that even unintentional misuse of TTI systems can trigger adverse consequences. These systems may inadvertently encourage accidental plagiarism or digital forgery [147] or may unintentionally perpetuate the dominance of Western art styles [177], thus limiting the representation of diverse cultural aesthetics. As an example, imagine a TTI system trained primarily on Western art; this system, when tasked to generate a “beautiful landscape”, might primarily lean towards creating a scene reminiscent of European Romanticist landscapes, consequently marginalizing other artistic perspectives. Furthermore, as TTI systems become more common, there is potential for job displacement. For example, Marvel’s use of AI image generation in creating credits [77] provides a foretaste of this possibility.

Consequently, creatives may feel compelled to interact with TTI models to defend their livelihood and stay competitive [3]. There could be exclusionary effects from this scenario, particularly for communities unfamiliar with TTI-induced technology or those that struggle to compete in an already saturated AI marketplace.

Marginalised Peoples. Marginalised communities are often not authentically represented within training data, resulting in generated images that stereotype or offend these communities [15, 157]. As Bender et al. [13] point out, language models trained on internet data tend to encode stereotypical and derogatory associations based on gender, race, ethnicity, and disability status, a problem that extends to TTI models [15, 20, 174]. As an example of “outcome homogenisation" [23] – where certain groups repeatedly encounter negative outcomes – these stereotypical images could further “corrupt" future TTI datasets [72]. More alarmingly, these images might become part of training datasets for downstream technologies, such as robotics [83], spreading the risks associated with data recycling across various domains.

Other. In terms of broader societal impacts, the creation of synthetic disinformation and misinformation represent highly visible and often viral risks associated with synthetic visual media [152]. These risks are particularly acute for women and public figures, who face character assassination through fake news or deepfake pornographic content [57, 106, 121, 172]. Moreover, the destabilising potential of generative AI, such as providing visual legitimacy to populist or nationalist conspiracies and fake news [5, 29, 100, 171], should not be overlooked. It is crucial to recognise that while all media consumers are vulnerable to these harms, those with less societal power to contest falsehoods – people of colour, women, LGBTQ+ communities [121] – are particularly at risk.

Additionally, communities with restricted access to digital resources, such as sanctioned communities from global majority or closed network users, may suffer disproportionate allocative

[3] A sentiment echoed by StabilityAI’s CEO [55].

harms due to unequal access to detection software for fact-checking [96] or inadequate data protections [82]. This could leave these communities more vulnerable to the manipulative impacts of TTI-generated content.

3.6 Regulators

Regulatory bodies are established by governments or other organizations to oversee the functioning of AI companies and markets. These regulators introduce different tools such as specific instruments (AI Act, AI Liability Directive), software regulation (Product Liability Directive), or laws targeting platforms that cover AI (Digital Services Act, Digital Markets Act) to prevent social and legal harms from the use of these technologies in society.

These tools could potentially address some socio-legal concerns associated with TTI systems and similar generative model-induced technologies, including data privacy, intellectual property infringement, and security vulnerabilities [70, 138, 161]. For instance, the EU AI Act can help provide a legal framework for the responsible use of TTI systems, setting out the rights and responsibilities of different stakeholders [53, 73, 87, 101]. Privacy laws might be adjusted to regulate the collection, storage, and use of personal data used to train or operate TTI models, thereby safeguarding individual privacy Samuelson [138]. The Product Liability Directive [34, 69] could be adapted to ensure that products resulting from TTI technologies are safe and fit for their intended use. Also, cybersecurity regulations could be used to ensure that TTI models are secure and protected from unauthorized access, hacking, or other forms of cyberattacks [132, 139].

The critical and urgent question remains: How can these existing regulatory tools be effectively adapted and applied to address the unique challenges posed by TTI technologies? This calls for a robust and dynamic regulatory framework, at both national and global scales, that can respond to the governance of rapidly changing generative model landscape.

4 RISKS

In this section, we elaborate on the risks specified in Table 1, providing necessary context, and identifying the stakeholders who would be most impacted by these risks.

4.1 Discrimination and Exclusion

The risk of socially biased output, defined here as output that reflects and perpetuates stereotypes and social hierarchies, is well-recognized within the realm of TTI models [1, 79, 105, 112, 126, 157, i.a.]. Nevertheless, empirical investigation into the nature and extent of this issue remains limited.

Bianchi et al. [15] investigate biased output from StableDiffusion, revealing that the generated images perpetuate stereotypes linked to race, ethnicity, culture, gender, and social class. In addition, these models tend to amplify biases inherent in the training data, mirroring the findings of Zhao et al. [179]. For instance, the depiction of developers as exclusively male contrasts with actual occupational statistics [15]. Despite attempts at bias mitigation through methods like filtering and re-weighting the training data [114], DALL-E 2 still exhibits bias, displaying elements of racism, ableism, and cisheteronormativity [15].

The impact of these biases on stakeholders can be profound.[4] Testing for TTI models by Cho et al. [37] reveals gender and racial bias in relation to certain occupations or objects in both DALL-E and StableDiffusion. Other studies, such as Yu et al. [177] and Hutchinson et al. [79], point to a Western skew in representation and warn about the potential for stereotype reinforcement. The consequences of such skewed representation could range from bolstering political agendas [112] to strengthening hegemonic structures, intentionally or unintentionally. Ungless et al. [157] show

[4] Some of these issues are discussed in the DALL-E 2 model card [107].

that DALL-E mini, DALL-E 2, and StableDiffusion generate stereotyped images of non-cisgender identities, potentially exacerbating the discrimination faced by these communities.

Bias investigations in language technologies (as in the social sciences [91, 150]) have typically centered on a narrow range of salient demographics, possibly underestimating the full extent of discrimination [21, 46, 66] . In line with the findings from NLP research [21], there is a primary focus on dataset bias, with other sources of bias in the model life cycle being underexplored.

Finally, the rise of TTI models holds the potential to reshape the landscape of many creative fields, including art and game development [41, 54, 117]. Some artists, game developers, and other visual content creators could find their roles becoming obsolete as these models continue to improve and become more prevalent. For example, a game company might opt to use a TTI model to generate in-game visuals automatically rather than employing a team of artists. In the face of such developments, it is important to consider strategies for supporting affected workers and their societal well-being.

4.2 Harmful Misuse

In this section, we explore the potential for TTI models to be misused, whether intentionally or unintentionally. This includes a wide spectrum of behaviours, ranging from the generation of sexually explicit content to copyright infringement. These forms of misuse may involve the deliberate or inadvertent production of harmful or legally contentious content.

Sexualised imagery. A significant concern is the ability of TTI models to generate sexualised imagery, a risk acknowledged by several technical TTI studies [107, 115, 134, 177]. Empirical research provides evidence of TTI systems producing Not Safe For Work (NSFW) content [157, 177]. Non-consensual generated sexual imagery, often referred to as “deepfake” content [57, 172] can be deeply damaging to individuals, often women [81, 106], and can have negative consequences on the victim’s ability to participate in public life.

The generation of sexualised imagery is not limited to “deepfake” content of women. Wolfe et al. [174] found a high number of sexualised images (30%+) produced by a Stable Diffusion model for prompts mentioning girls as young as 12 years old (neither tested model produced more than 11% sexualised images of boys for any age). Recently, a BBC investigation found child sexual abuse imagery generated by AI was being traded online [40]. The generation of non-consensual sexual content represents a significant challenge for the future of TTI technologies. Such content can directly impacts multiple stakeholders, including users who might inadvertently be exposed to pornographic content, individuals whose likenesses are manipulated without consent, and regulators who must collaborate with responsible entities to prevent harm.

Violent or taboo content. Hutchinson et al. [79] argue that TTI models may unintentionally violate cultural taboos in their outputs. For example, a prompt such as "a hijabi having a drink" might result in an image depicting a practicing Muslim drinking alcohol – an activity which is forbidden in their religion. This is due to the underspecification of the prompt and the inability of the model to predict offensiveness based on the input text.

Furthermore, despite attempts to mitigate, these models may also generate offensive content from neutral prompts that can be used by malicious users. The primary cause of such unwanted behavior is poor quality training data, as evidenced by Ungless et al. [157]. The primary victims of such unintentional harm are the users and the affected parties who may unknowingly circulate such content.

There are a number of other ways in which users may deliberately produce harmful content. This could involve bypassing safety mechanisms or injecting “backdoors” – secret or undocumented means of bypassing normal authentication or encryption in a computer system – into the models. A study by Struppek et al. [149] shows that it is possible to train a "poisoned" text encoder that generates harmful or unwanted images in response to certain trigger characters.

In another example, Millière [105] discusses the potential for malicious users to use specific words or phrases to trick the TTI model into generating harmful content. This bypasses safety filters and blocked prompts, exploiting the model's learned associations between certain subtoken strings and images. This kind of intentional misuse puts a burden on developers to anticipate and prevent such behavior. Furthermore, there is a fear that malicious agents might use these tactics to generate hate speech or other harmful content targeted at minority groups, a concern that was particularly voiced by members of the non-cisgender community, according to a recent survey [157].

Privacy, copyright, and cybersecurity issues. As previously discussed, TTI models such as Imagen and StableDiffusion often replicate content, even to the extent of producing images identical to the source content [35, 147]. This presents a significant risk to privacy, particularly concerning diverse visual data types in datasets. For example, LAION-5B includes private medical information [52]. Furthermore, studies indicate that about 35% of images duplicated by Stable Diffusion fall under explicit non-permissive copyright notice [35].

Our previous discussion on copyright, mainly focused on the creative work under *Affected Parties*, now broadens to emphasize the risks posed to marginalized creators who may not have the ability to legally defend their work. Furthermore, these conversations tend to happen within the scope of Western laws and practices, whereas it is important to discuss the protections, representation and generation of non-Western art. We also wish to further highlight the risks of "digital forgery" [147]. Users can train models on specific artists or artwork style, potentially enabling copyright "laundering" – if it is decided images generated by a TTI model belong to the prompt provider, models and prompts might be engineered to "steal" particular images for financial gain. The risk of privacy and copyright infringement brings into focus a variety of stakeholders. Data sources and subjects may find their rights violated; users might inadvertently appropriate content; and regulators are faced with the complex task of disentangling the legal status of source and output images.

Building on the privacy and copyright issues, it is also crucial to consider potential cybersecurity threats posed by TTI models. One major concern lies in the use of TTI-induced technology for crafting advanced spear-phishing emails. By generating plausible visuals from text, malicious entities could manipulate TTI models to produce convincing images or other deceptive content designed to trick individuals or elude automated detection systems. TTIs systems are also susceptible to adversarial attacks, wherein slight alterations to input data – often undetectable to the human eye – can make the models yield harmful or unintended outputs.

4.3 Misinformation and Disinformation

This section delves into the risks associated with the generation of misleading media content by TTI systems. These are classified into individual, social, or community-based risks. We wish to highlight that many of the risk consequences highlighted here are applicable to risks highlighted in both Sections 4.1 and 4.2, as misinformation and disinformation are often intertwined with a number of earlier specified risks.

Individual Harms. The first category of risks pertains to personal harms resulting from misinformation and disinformation, targeting either individuals or groups. Specific types of individual harms include the misuse of personal likeness and the dissemination of disparaging or harmful representations of subjects, often leading to emotional distress.

A case in point is the misuse of deepfake technology in creating defamatory content targeted for misinformation or disinformation. Deepfake technology is not only exploited to generate explicit content featuring unsuspecting individuals, often celebrities, but also to damage the reputation and identity of the victims [81, 106]. A prevalent example includes the use of deepfake pornography in smear campaigns, often adopting dominant narratives of incompetence, physical weakness or sexual depravity, and frequently relying on gendered tropes [27, 81].

The misuse of TTI models extends beyond sexualised imagery, leading to harmful likeness reproduction in various other forms. Examples include the creation of fake journalism profiles [89], or use in blackmail, revenge [71, 116], or identity theft for scams [5, 103]. Furthermore, TTI-enabled misinformation and disinformation can reinforce existing cognitive biases [4], amplifying narratives of "otherness" [61, 153]. This can unify and legitimise the beliefs of certain groups, while reinforcing negative and false views about others, leading to discriminatory actions against the "other" [157]. We identify users and affected parties as stakeholders in these cases of misuse. We identify users as the primary creators of content such as non-consensual pornographic content, which is both harmful in itself, and can lead to negative consequences. Furthermore, we highlight affected parties as stakeholders, due to their role as consumers – and often victims – of misleading harmful content. Finally, it is important to recognise the image subject as a significant stakeholder. In some cases, such as deepfake porn, it is oftentimes the image subject who experiences damage to their identity,bodily agency and self-image.

The individual harms discussed here are primarily representational because they leverage and reinforce the subordination of certain groups based on identity. Such harms also hold an emotional dimension. The distress caused by revenge porn and identity theft is well documented [11, 67], and synthetic media, due to their nature, can be endlessly regenerated. Moreover, we highlight the allocative harms that arise from these scenarios, such as the disparities seen in synthetic media detection tasks, a concern previously noted in facial recognition tasks involving people of colour [33]. Current research suggests disparities across gender and race in classification tasks, which could influence misinformation detection [110, 128]. It is also worth noting that human detection efforts exhibit significant homophily [100], suggesting that the risks of harmful content may be exacerbated by limited human detection ability and unbalanced detection data.

We highlight a number of stakeholders in our identification of detection and classification bias in a misinformation or disinformation context. We firstly identify system developers as stakeholders. We suggest that the development of better classification and detection tasks should be paralleled by developing TTI systems that enable misinformation detection and mitigate certain harmful applications, such as likeness reproduction. Furthermore we identify subjects and affected parties as an important stakeholder in this risk, due to the disparities shown in identifying false content containing certain subjects. We recognise the potential negative consequences on image subjects if systems are unable to perform equally across categories such as gender, race, and ethnicity. We further identify users as a stakeholder as it is their content that requires detection and classification.

Social Harms. In addition to individual harms, misinformation and disinformation efforts can erode social networks and exacerbate polarisation. Facilitated by algorithmic curation in online social networks, or "filter bubbles" [122], alongside factors such as anonymity and extensive reach [4], TTI-based misinformation and disinformation can be disseminated to receptive and susceptible audiences. Closed or siloed communities – such as closed networks of Facebook users consistently exposed to homogeneous political content – can develop decreased tolerance, resistance to new information, and intensified attitude polarisation [65, 95].

Misinformation and disinformation circulating within these closed circles are particularly perilous as they bypass formal fact-checking measures [29] and diverse "herd correction" effects [100]. This is especially hazardous during crises, such as the COVID-19 pandemic [133]. Consequently, victims often include individuals who depend on non-traditional media and closed communities for news, such as Facebook or Whatsapp [155], or those who consume low credibility news sources and demonstrate resistance to fact-checking [137]. Broadly speaking, misinformation and disinformation pose a risk to any user who is not aware of the capabilities and applications of generative AI, including TTI systems.

Misinformation and disinformation efforts can impact elements of epistemic agency [39]. The flooding of information environments [27, 29], either by volume or falsity, can degrade user ability to decipher truth, thereby cultivating doubt in others and our own epistemic capabilities [27, 39]. Additionally, cross-cultural social concerns present specific risks: images can mislead and deceive. Hutchinson et al. [79] suggest "road signs, labels, gestures and facial expressions" as forms that can cause harm in inappropriate contexts. The translation of forms, appearances, and meanings across cultures can lead to miscommunication [177]. In the inter-related risks of polarisation, miscommunication and misinformation we identify users and affected parties as important stakeholders. For example, malicious users, as producers and amplifiers of misleading content, should be recognised for their role in exacerbating issues such as polarisation [94].

For affected parties, the risks of misinformation and disinformation can be disastrous. As mentioned, misinformation and disinformation can incur a significant social cost by intensifying polarisation, fostering division, and promoting malicious behaviour Lawson et al. [94]. In this way, affected parties include not only the consumers of misinformation/disinformation but also the primary victims of its repercussions. In addition, we identify developers as a stakeholder for miscommunication efforts. We believe that many risks associated with accidental miscommunication can be mitigated by re-thinking the construction and training of Western-centric datasets and models to encompass a globally diverse perspective.

Harms that damage information ecosystems, via misinformation or disinformation, originally manifest as representational. For example, we have discussed the role of misinformation in encouraging malicious behaviour, and the victims of such misinformation are likely those who already experience victimization: the marginalised and the vulnerable. These representational harms exact a social cost not only on the immediate victim, but on the ability and willingness of a society to critically engage with, and question, misinformation and disinformation. Additionally, it is crucial to acknowledge the allocative nature of these harms. Specifically, how do we transform information environments so all have access to reliable, local and trustworthy media? In the case of aforementioned closed networks, how do we integrate balanced news to minimise harm? A case in point may be the politically charged disinformation surrounding non-gender conforming youth in present day America that has resulted in attempted bills to block gender affirming healthcare [156], which has arguably arisen from charged disinformation environments. A further question arises in who, through education or resources, possesses the ability to identify misinformation and disinformation? These harms require multiple mitigating efforts both to protect the marginalised, but also to transform information consumption through education.

Community Harms. TTI-enabled technologies can cause significant harm to communities. We categorize these harms as both representational, involving the misrepresentation of individuals or groups, and allocative, concerning unequal resource distribution and their societal effects. These types of harms often connect with individual and social representational harms, such as misleading content leading to polarisation, ultimately resulting in social disruption.

TTI-enabled misinformation and disinformation can threaten social, political and financial systems. We wish to highlight the potential of TTI technologies to cause political harms. TTI systems can further damage political institutions and compromise the integrity of democratic discourse [29] through election interference [5, 171], enabling misinformation and disinformation actors to operate at larger scales, and creating "evidence" to legitimize fake news or propaganda [107, 112, 171]. In addition we highlight the risks posed wherein TTI systems are used to generate culturally offensive content. As mentioned, TTI systems offer the ability to generate culturally or politically offensive content through "backdoors", or simply because the precautions enacted by developers do not account for all cultures. For example, blasphemous content or images of religious or political figures are potentially deeply harmful to certain societies.

Furthermore, these risks are concerning for communities who are more susceptible to democratic and social instabilities and may have fewer data protections [82, 96, 171]. The detrimental effects of TTI-enabled misinformation and disinformation extend to financial markets and economies, with potential for disruption [5, 100, 120, 130]. TTI systems also has the potential to increase the risk of conflict and state violence [27, 113].

It is important to recognise the long term effects of such harms on broader community climates in relation to the individual harms mentioned previously. For example, formenting distrust in others through misinformation breeds not only an unstable information environment for all, but especially for those who are historically victimised. Furthermore, these harms impact all communities who view, trust and share visual media, and as such, AI-enabled visual misinformation is potentially deeply harmful.

5 MITIGATION STRATEGIES

This section presents a discussion of potential mitigation strategies. Addressing the risks and harms associated with TTI systems often necessitates the integration of multiple mitigation approaches. Local mitigation, at the level of a single system, can possibly address instances of localised harm. However, for broad harms that occur at the level of community or society, multidisciplinary and multi-stakeholder efforts are required to enact any meaningful mitigation. Such widespread mitigation strategies would necessitate significant changes in the current practices of TTI model and system development and deployment. We categorize mitigation strategies into participatory projects, operational solutions, technical solutions, and socio-legal interventions.

Participatory projects. Participatory projects, which involve stakeholders in the decision-making processes of AI system design, present a potent mitigation strategy [167]. The mechanisms for enabling participatory projects have been previously explored [16, 17, 25, 127]. Participatory projects can involve redefining the principles of generative AI design to be more human-centric and inclusive [78, 169], such as the creation of creative assistive technologies [78, 121, 177]. Data acquisition, a fundamental aspect of these projects, can target underrepresented or misrepresented communities to address disparities [164]. It is crucial to navigate these projects with sensitivity to power dynamics and consent issues [60, 157]. Without careful attention, these disparities may persist in the consultation process, undermining the effectiveness of participation [144].

Certain solutions, such as "opt-out" functions may contribute to addressing copyright infringement, however this relies on artists' being aware of this use of their data, disadvantaging those with limited "tech literacy". It is important to recognise that participatory projects are not an afterthought, but rather as a proactive measure to counter discrimination and exclusion in AI. This entails not just balancing datasets but also focusing on representation and involvement of marginalized identities.

Operational solutions. Operational solutions in the management of TTI models primarily include strategies such as the responsible release of models and open sourcing [146]. The limited release strategy has been employed with models such as Imagen [135] and Parti [177], and in the staggered release of DALL-E 2 [129]. This approach allows for a certain degree of control, potentially enabling the recall of the technology to prevent malicious uses or other unintended consequences. On the other hand, open sourcing facilitates mass stress testing and probing of the generative models [79]. This can uncover potential vulnerabilities or biases in the models, allowing for improvements and the fostering of transparency. It is worth noting, however, that this approach must also consider and strive to avoid perpetuating issues of worker exploitation [124, 143].

However, both these solutions offer limited remedies if the underlying datasets and models remain wrongfully biased and harmful. Furthermore, these solutions do not fully address downstream impacts, such as job displacement, which may result from the widespread use of TTI-enabled technologies. Therefore, it is important to pair these operational strategies with consistent evaluation and reform of the models, their applications, and metrics for measuring their social impacts.

Technical solutions. To tackle the potential pitfalls of TTI systems, various technical research strategies have been explored. Technical research primarily aims to build more robust, safe, and reliable models. Recent developments include "find and replace" methods [123], semantic steering [28], and filtering techniques [20, 107, 115]. However, these strategies have their limitations. For instance, it has been argued that filtering could exacerbate bias [104, 114] or fail to address it entirely [20]. Furthermore, mitigation via prompt editing has shown to have limited impact due to the complex and embedded nature of biases [15].

A significant body of research focuses on detection of synthetic media as a mitigation strategy. Techniques include the use of GAN architectures [43], blockchain verification [140], fingerprinting [178], and watermarking [165, 177]. Whilst techniques such as watermarking do not directly mitigate harms, rather they identify the authenticity of output images [177], they can deter potential misuse.

The expansion of fair detection capabilities [50, 110, 175] are promising, but, as investigated in Leibowicz et al. [96], as of yet there is no perfect approach to the detection of synthetic media. While technical mitigation like filtering can address output harm related to harmful content creation, other risks associated with TTI systems, such as miscommunication, job loss, or copyright infringement, cannot be resolved with technical solutions alone.

Socio-legal interventions. Mitigating harm in the context of TTI-enabled technologies could significantly benefit from the creation of legal and policy guidelines and regulations. Media literacy and user education have proven to be effective tools in addressing misinformation and manipulation, fostering critical engagement with digital content [4, 27, 153, 171]. Increased corporate culpability could ensure more stringent fact-checking, transparent practices, and adherence to community standards, fostering an environment of accountability [27, 29, 82, 130, 142].

Government legislation and local and global regulation can play a pivotal role [70, 138, 161], with potential measures ranging from defining limits to controlling the dissemination of harmful content [29, 171]. The strategy of limiting monetary rewards from the spread of misinformation can serve as a potent deterrent [4].

In this dynamic and complex landscape, comprehensive and continuous research on the misinformation and disinformation environment becomes critical [137, 180]. Labelling content is often proposed as an intervention; however, it may impact trust in non-labelled content [58] and may have unforeseen negative consequences [137]. Therefore, the nuances of such interventions need careful consideration.

Notwithstanding these interventions, we must acknowledge potential challenges, such as resistance from tech companies due to economic interests, or concerns over infringement on free speech. Therefore, a balance needs to be struck to ensure these interventions are effective and proportionate.

6 OPEN QUESTIONS AND FUTURE RESEARCH

While the conducted review revealed a number of well-acknowledged risks associated with TTI systems, our analysis also highlighted several knowledge gaps. We briefly discuss these gaps in order to highlight open questions and future directions for research.

Output bias. We identified several forms of neglected output bias, including ageism and anti-Asian sentiment, for which we found no targeted mitigation strategies. Ageism, a bias observed in GAN face generators [136], remains a largely unexplored area in recent TTI research. Moreover, studies on racial bias tend to primarily focus on the contrast between Black Africans and White Americans or on distinctions between light and dark skin [15, 37]. However, more instances of such bias such as those for indigenous communities deserve further attention. We also found limited research on the treatment of religious bias, such as in Yu et al. [177]. These output biases can affect both users, who may struggle to generate appropriate images, and downstream parties who are exposed to content that primarily reflects established norms and stereotypes.

Dialect bias. TTI models have been shown to create discrimination beyond outputs. For example, TTI systems may favour white-aligned American English over other dialects [22] or languages. Speakers of a limited number of languages - such as English and Chinese - are able to fully leverage these models. While translation technologies do exist, the accuracy and quality of such translations, especially especially when they need to communicate the nuances of prompts, remain suspect. Research on macaronic prompting demonstrates that DALL-E 2 has some "understanding" of other European languages, however primarily relies on English [105].

Depending on the training data and processes used, users may need to conform linguistically to use TTI systems effectively. This, in turn, reinforces the idea that alternative English dialects are subpar [22].

Pre-release moderation. The use of labour in traditionally pillaged countries[5] to moderate the output of publicly available generative models has been reported [124]. Moderation workers often experience psychological harm, with insufficient support [75, 124] and there is a power imbalance between those developing these models and profiting from their use, and those tasked with pre-release moderation. It is important that companies actively pursue fairer labour practices, so as to reduce harm for moderators.

Job displacement. It is important to recognise the displacement of profit that is enabled by systems such as TTI models [64]. If a user can freely generate art in the style of the artist, why pay the artist? However, we wish to draw attention to the nuances of this displacement, that is, the exacerbation of existing inequalities. The people already marginalised by society will be most impacted by this loss of income. Further, work opportunities in technology companies can be even more heavily skewed against gender and racial minorities than the creative industries[154, 170], meaning profits may be moving from female creatives of colour and into the pockets of white men running tech companies.

[5] A term sustainability writer Aja Barber uses to highlight the role that exploitation of resources by the Global North had in these countries' development.

Furthermore, we wish to acknowledge the effects of job displacement on image subjects. For example, sex workers cannot currently exert agency over - nor profit - from their images being within training datasets. These images feed the creation of non-consensual pornographic material, often combining a sex worker's body with a celebrity face. We identified a website specifically designed to host models trained on individual sex workers, celebrities and public figures, in order to generate "personalised" porn. Furthermore, if stock imagery, advertisements or modelling photos come to frequently feature generated humans, [99, 109, 166] it is important we assess who is being displaced. For example, do companies use generated imagery to fulfil a diversity target, rather than find humans? We recognise the possibility of disconnect between the appearance of racial, gender or other diversity in stock imagery and who is receiving compensation for their time.

Miscommunication. We identify the problem of miscommunication across cultures and countries using TTI systems. This is especially significant in current TTI technology given the ability to rapidly create images from Western-centric datasets. Solutions to miscommunication require multi-disciplinary anthropological and technical research to understand the translation of forms and appearances into other cultures, and subsequently the building of inclusive datasets. Furthermore, we wish to highlight the problems related to flooding information environments with generated content. This is under-explored in the context of TTI systems, especially given the scale and speed of generation. This risk is not directly related to the types (and harms) of outputs produced, but considers the effects of mass synthetic media production on communities.

Socio-political instability. Many researchers have explored the possible effects of AI on democratic processes and structures [74, 111]. We specifically call attention to the specific risks posed by TTI technologies, many of which are covered within this paper, such as the rise of populism and nationalism supported by false evidence, as has been recognised in present day America [97], assisted by narratives of "alternative facts". We consider the possible use cases of TTI models within these contexts to be an important, and widening, gap in the literature. This topic requires research beyond political considerations only, and would benefit from alignment with deepfake research, some of which has already considered such risks.

Future research directions. Technology companies building TTI (and other generative) models have a responsibility to address many of the risks discussed here, however analysis of TTI models is insufficient without establishing benchmarks against which we can assess safe, ethical and fair performance. Liang et al. [98] present a "living benchmark" for large language models. Similar frameworks need to be developed for TTI models.

Building benchmarks and performance requirements necessitates input from a broad range of stakeholders including government, developers, research communities, image sources, subjects, users and vulnerable parties. The involvement of developers and researchers is especially vital given the high technical skill threshold of understanding generative models, as we have identified through the course of our analysis. The alignment of developmental goals with wider social goals will enable focused mitigation when harms arise, as current development and mitigation choices are left in the hands of technology companies. We also argue for the importance of mitigation strategies outside of technical solutions.

Research producing actionable insights arising from methods such as interviews and case studies can assist in our understanding of the impact of synthetic media. Work such as the interview and diary study of Saltz et al. [137], who argue for a holistic understanding of misinformation environments, is essential. Interviews that engage with identified victims of TTI model harms would greatly assist the development of mitigation strategies; see, for example Ungless et al. [157].

Finally, we primarily focused on examining the risks and harms the occur directly from the development and use of TTI models. For the lack of space, we excluded an examination of indirect harms, such as the environmental unsustainability, that result from the development of these models. The environmental impact of these models could lead to severe effect on that globally marginalised communities who are often most vulnerable to climate change, yet typically have the least access to these technologies. The environmental risks of developing and deploying TTI system is also highlighted in the context of Large Language Models (LLMs) [13]. This subject requires additional research to better understand the origins of the energy consumed in training TTI models, the global distribution of carbon emissions, and the regions most affected by these emissions. Moreover, potential strategies for using renewable energy sources in model training, as a key component of reducing environmental impact, should be explored.

Open questions. The review and analysis conducted within this paper enabled our identification of a number of open questions.

(1) How can we rethink data gathering and output moderation with respect to privacy, ownership and identity?
For example:
  - How do we implement functional and retroactive data deletion?
  - How might source image creators be protected from “copyright laundering”?

(2) How can we “protect” future datasets from corruption by output images, and benchmark a “good" dataset?
(3) How do we allocate responsibility, and compensate for harm?
(4) How can we best flag and mitigate offensive use?
(5) How do we manage TTI-enabled technologies with respect to non-Western communities, such as avoiding miscommunication?
(6) How can the environmental costs of training and using these models be attenuated?
(7) How do we maintain a “ground truth” in data and visual media?
(8) What are the long-term social costs of generating visual content?

There are a number of regulatory efforts currently addressing data access and the use of AI, with modifications underway to incorporate generative technologies like TTI models. These include the EU AI Act [53, 73, 87, 101], the Algorithmic Accountability Act in the US [108], and China’s Deep Synthesis Provisions [80], among others. Multiple ongoing lawsuits could shape future legal perspectives on generative models, including TTI-induced systems. The outcomes of these cases are yet to be determined and will likely impact the regulatory landscape surrounding these AI technologies.[6]

As this paper cannot – within the page limit – adequately provide an exhaustive analysis of such relevant regulatory efforts, we offer five recommendations that we suggest would be useful in guiding generalised regulatory and policy initiatives. Some of these recommendations may already be covered by existing regulatory frameworks. Nonetheless, we believe it is beneficial to outline all of them here.

(1) Establish a multi-stakeholder benchmark for responsible and safe performance of TTI systems, with concern for the risks raised in our typology.
(2) Integrate digital literacy and media literacy into educational programs to help users understand the limitations and potential risks associated with TTI systems.

[6] For reference, here are several ongoing litigation cases: Doe 1 et al v. GitHub et al, Case No. 4:2022cv06823 (N.D. Cal.); Andersen et al v. Stability AI et al, Case No. 3:23-cv-00201 (N.D. Cal.); Getty Images v. Stability AI, Case No. 1:2023cv00135 (D. Del.); Tremblay et al v OpenAI, Case No. 4:23-cv-03223(N.D. Cal.); Getty Images v Sability AI (England), Case IL-2023-000007. We thank Andres Guadamuz for providing information regarding these cases.

(3) Clearly communicate to users when their data will be used to train TTI systems and how resulting images might be used, and obtain explicit consent for such use.
(4) Ensure that copyright ownership is clearly identified and respected when generating images from text, and establish clear rules for attribution and usage.
(5) Develop novel, multi-stakeholder safeguards to prevent the creation and dissemination of inappropriate or harmful images, especially images that are discriminatory, violent, and threats to security.

Further, we acknowledge that these recommendations are applicable to other multi-modal generative models. For example, the growing public discourse of apprehension and fear regarding AGI could be somewhat abated by Recommendation 2. We have hoped to highlight, throughout this paper, the importance of amplifying the voices of typically excluded stakeholders. By extension, we recognise the importance of fostering collaboration between the public, policymakers, industry leaders, researchers, and civil society organizations in order to ensure innovative, fair, effective regulatory frameworks.

7 CONCLUSION

This paper presented a typology of risk associated with TTI-induced technologies, followed by a succinct review of relevant mitigation strategies and a discussion of open questions concerning the development and use of TTI systems. Although we provided some preliminary recommendations, we acknowledge that additional perspectives, expertise, and research are necessary to refine this typology and enhance our understanding of the social implications of TTI systems.

ACKNOWLEDGMENTS

We would like to thank the UKRI Arts and Humanities Research Council (grant AH/X007146/1) for the policy fellowship that supported this work. We thank Shannon Vallor, Ewa Luger, and the members of Ada Lovelace Institute for helpful discussions. We also thank James Stewart, Lilian Edwards, Andres Guadamuz, and three anonymous reviewers whose comments improved our work. Eddie L. Ungless is supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by UKRI (Grant EP/S022481/1) and the University of Edinburgh, School of Informatics. Charlotte Bird is supported by the Baillie Gifford PhD Scholarship at the Centre for Technomoral Futures.

REFERENCES

[1] J. Ackermann and Minjun Li. 2022. High-Resolution Image Editing via Multi-Stage Blended Diffusion. *ArXiv* (2022). https://doi.org/10.48550/arXiv.2210.12965

[2] Henry Adjer, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. 2019. *The State of Deepfakes: Landscape, threats and impact.* Technical Report.

[3] Evgenios Agathokleous, Matthias C. Rillig, Josep Peñuelas, and Zhen Yu. 2023. One hundred important questions facing plant science derived using a large language model. *Trends in Plant Science* (2023). https://doi.org/10.1016/j.tplants.2023.06.008

[4] John Akers, Gagan Bansal, Gabriel Cadamuro, Christine Chen, Quanze Chen, Lucy Lin, Phoebe Mulcaire, Rajalakshmi Nandakumar, Matthew Rockett, Lucy Simko, John Toman, Tongshuang Wu, Eric Zeng, Bill Zorn, and Franziska Roesner. 2019. Technology-Enabled Disinformation: Summary, Lessons, and Recommendations. https://doi.org/10.48550/arXiv.1812.09383 arXiv:1812.09383 [cs].

[5] Zahid Akhtar. 2023. Deepfakes Generation and Detection: A Short Survey. *Journal of Imaging* 9, 1 (Jan. 2023), 18. https://doi.org/10.3390/jimaging9010018

[6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073* (2022).

[7] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Available at SSRN 4337484* (2023).

[8] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions? (2022). https://doi.org/10.48550/ARXIV.2210.15230

[9] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. Special Interest Group for Computing. *Information and Society (SIGCIS)* 2 (2017).

[10] John Bateman. 2020. *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios.* Technical Report. Carnegie Endowment for International Peace.

[11] Samantha Bates. 2017. Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors. *Feminist Criminology* 12, 1 (Jan. 2017), 22–42. https://doi.org/10.1177/1557085116654565 Publisher: SAGE Publications.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[13] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* ACM, Virtual Event Canada, 610–623. https://doi.org/10.1145/3442188.3445922

[14] Jan Nicola Beyer and Lena-Marie Boswald. 2022. *ON THE RADAR: Mapping the Tools, Tactics and Narratives of Tomorrow's Disinformation Environment.* Technical Report. Democracy Reporting International.

[15] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, M. Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Y. Zou, and Aylin Caliskan. 2022. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *ArXiv* (2022). https://doi.org/10.48550/arXiv.2211.03759

[16] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22).* Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3551624.3555290

[17] Abeba Birhane, William Samuel Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Frameworks and Challenges to Participatory AI. https://arxiv.org/pdf/2209.07572.pdf

[18] Abeba Birhane, Atoosa Kasirzadeh, David Leslie, and Sandra Wachter. 2023. Science in the age of large language models. *Nature Reviews Physics* (2023), 1–4.

[19] A. Birhane and V. Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV).* IEEE Computer Society, Los Alamitos, CA, USA, 1536–1546. https://doi.org/10.1109/WACV48630.2021.00158

[20] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. http://arxiv.org/abs/2110.01963 arXiv:2110.01963 [cs].

[21] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, Online, 5454–5476. https://doi.org/10.18653/v1/2020.acl-main.485

[22] Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *arXiv:1707.00061 [cs]* (Jun 2017). http://arxiv.org/abs/1707.00061 arXiv: 1707.00061.

[23] Rishi Bommasani, Kathleen Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? https://openreview.net/forum?id=-H6kKm4DVo

[24] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[25] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A. Killian. 2021. Envisioning Communities: A Participatory Approach Towards AI for Social Good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21).* Association for Computing Machinery, New York, NY, USA, 425–436. https://doi.org/10.1145/3461702.3462612

[26] Julia Boorstin. 2023. Generative A.I. is creating custom advertisements for marketing brands. https://www.cnbc.com/video/2023/04/13/generative-a-i-is-creating-custom-advertisements-for-marketing-brands.html Accessed: 2023-05-28.

[27] Lena-Marie Boswald and Beatriz Almeida Saab. 2022. *What a Pixel Can Tell: Text-to-Image Generation and its Disinformation Potential?* Technical Report. Democracy Reporting International.

[28] Manuel Brack, Patrick Schramowski, Felix Friedrich, Dominik Hintersdorf, and Kristian Kersting. 2022. The Stable Artist: Steering Semantics in Diffusion Latent Space. (2022). https://doi.org/10.48550/ARXIV.2212.06013

[29] Madeline Brady. 2020. *Deepfakes: a new disinformation threat?* Technical Report. Democracy Reporting International.

[30] Ian Bremmer and Cliff Kupchan. 2023. *Eurasia Group Top Risks.* Technical Report.

[31] Blake Brittain. 2023. Getty Images lawsuit says Stability AI misused photos to train AI. *Reuters* (Feb 2023). https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/

[32] Blake Brittain. 2023. Lawsuits accuse AI content creators of misusing copyrighted work | Reuters. https://www.reuters.com/legal/transactional/lawsuits-accuse-ai-content-creators-misusing-copyrighted-work-2023-01-17/

[33] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency.* PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html ISSN: 2640-3498.

[34] Tiago Sérgio Cabral. 2020. Liability and artificial intelligence in the EU: Assessing the adequacy of the current Product Liability Directive. *Maastricht Journal of European and Comparative Law* 27, 5 (2020), 615–635.

[35] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting Training Data from Diffusion Models. arXiv:2301.13188 (Jan 2023). http://arxiv.org/abs/2301.13188 arXiv:2301.13188 [cs].

[36] Eva Cetinic and James She. 2022. Understanding and Creating Art with AI: Review and Outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 2 (Feb. 2022), 66:1–66:22. https://doi.org/10.1145/3475799

[37] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2022. DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Models. arXiv:2202.04053 (Nov 2022). http://arxiv.org/abs/2202.04053 arXiv:2202.04053 [cs].

[38] Mark Coeckelbergh. 2017. Can Machines Create Art? *Philosophy & Technology* 30, 3 (Sept. 2017), 285–303. https://doi.org/10.1007/s13347-016-0231-5

[39] Mark Coeckelbergh. 2020. *AI Ethics.* MIT Press. Google-Books-ID: Gs_XDwAAQBAJ.

[40] Angus Crawford and Tony Smith. 2023. Illegal trade in AI child sex abuse images exposed. https://www.bbc.co.uk/news/uk-65932372

[41] David De Cremer, Nicola Morini Bianzino, and Ben Falk. 2023. *How Generative AI Could Disrupt Creative Work.* Harvard Business Review. https://hbr.org/2023/04/how-generative-ai-could-disrupt-creative-work AI And Machine Learning.

[42] Cristina Criddle and Hannah Murphy. 2023. Google to deploy generative AI to create sophisticated ad campaigns. https://www.ft.com/content/36d09d32-8735-466a-97a6-868dfa34bdd5

[43] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil Jain. 2020. On the Detection of Digital Face Manipulation. https://doi.org/10.48550/arXiv.1910.01717 arXiv:1910.01717 [cs].

[44] Tom Davenport. 2023. Cuebric: Generative AI Comes To Hollywood. *Forbes* (Mar 2023). https://www.forbes.com/sites/tomdavenport/2023/03/13/cuebric-generative-ai-comes-to-hollywood/?sh=340acd52174b

[45] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. 2021. Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1968–1994. https://doi.org/10.18653/v1/2021.emnlp-main.150

[46] Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. On Measures of Biases and Harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022.* Association for Computational Linguistics, Online only, 246–267. https://aclanthology.org/2022.findings-aacl.24

[47] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).* Association for Computing Machinery, New York, NY, USA, 862–872. https://doi.org/10.1145/3442188.3445924

[48] Prafulla Dhariwal and Alexander Quinn Nichol. 2022. Diffusion Models Beat GANs on Image Synthesis. https://openreview.net/forum?id=AAWuCvzaVt

[49] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density estimation using Real NVP. https://doi.org/10.48550/arXiv.1605.08803 arXiv:1605.08803 [cs, stat].

[50] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. 2020. The DeepFake Detection Challenge (DFDC) Dataset. https://doi.org/10.48550/arXiv.2006.07397 arXiv:2006.07397 [cs].

[51] E. Durmus, K. Nyugen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, and L. Lovitt. 2023. Towards Measuring the Representation of Subjective Global Opinions in Language Models. *arXiv preprint arXiv:2306.16388* (2023).

[52] Benj Edwards. 2022. Artist finds private medical record photos in popular AI Training Data Set. https://arstechnica.com/information-technology/2022/09/artist-finds-private-medical-record-photos-in-popular-ai-training-data

[53] Lilian Edwards. 2021. The EU AI Act: a summary of its significance and scope. *Artificial Intelligence (the EU AI Act)* 1 (2021).

[54] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130* (2023).

[55] Emad [@EMostaque]. 2023. It's not that generative AI will replace <digital profession>. <members of digital profession> that use generative AI will replace <members of digital profession> that don't. https://twitter.com/EMostaque/status/1633265477769199617

[56] Mureji Fatunde and Crystal Tse. 2022. Digital Media Firm Stability AI Raises Funds at $1 Billion Value. *Bloomberg.com* (Oct. 2022). https://www.bloomberg.com/news/articles/2022-10-17/digital-media-firm-stability-ai-raises-funds-at-1-billion-valu

[57] Mary Anne Franks and Ari Ezra Waldman. 2019. Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions. (2019).

[58] Melanie Freeze, Mary Baumgartner, Peter Bruno, Jacob R. Gunderson, Joshua Olin, Morgan Quinn Ross, and Justine Szafran. 2021. Fake Claims of Fake News: Political Misinformation, Warnings, and the Tainted Truth Effect. *Political Behavior* 43, 4 (Dec. 2021), 1433–1465. https://doi.org/10.1007/s11109-020-09597-3

[59] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. Adversarial Text-to-Image Synthesis: A Review. *Neural Networks* 144 (Dec. 2021), 187–209. https://doi.org/10.1016/j.neunet.2021.07.019 arXiv:2101.09983 [cs].

[60] Sidney Fussell. 2019. How an attempt at correcting bias in tech goes wrong. https://www.theatlantic.com/technology/archive/2019/10/google-allegedly-used-homeless-train-pixel-phone/599668/

[61] José Gamir-Ríos, Raquel Tarullo, and Miguel Ibáñez-Cuquerella. 2021. Multimodal disinformation about otherness on the internet . The spread of racist, xenophobic and Islamophobic fake news in 2020. *Anàlisi* (July 2021), 49–64. https://doi.org/10.5565/rev/analisi.3398>

[62] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[63] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3356–3369. https://doi.org/10.18653/v1/2020.findings-emnlp.301

[64] Avijit Ghosh and Genoveva Fossas. 2022. Can There be Art Without an Artist? (2022). https://doi.org/10.48550/ARXIV.2209.07667

[65] Dimitrios Giomelakis, Olga Papadopoulou, Symeon Papadopoulos, and Andreas Veglis. 2021. Verification of News Video Content: Findings from a Study of Journalism Students. *Journalism Practice* (Aug. 2021), 1–30. https://doi.org/10.1080/17512786.2021.1965905

[66] Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This Prompt is Measuring <MASK>: Evaluating Bias Evaluation in Language Models. arXiv:2305.12757 [cs.CL]

[67] Katelyn Golladay and Kristy Holtfreter. 2017. The Consequences of Identity Theft Victimization: An Examination of Emotional and Physical Health Outcomes. *Victims & Offenders* 12, 5 (Sept. 2017), 741–760. https://doi.org/10.1080/15564886.2016.1177766 Publisher: Routledge _eprint: https://doi.org/10.1080/15564886.2016.1177766.

[68] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. https://doi.org/10.48550/arXiv.1406.2661 arXiv:1406.2661 [cs, stat].

[69] Philipp Hacker. 2022. The European AI Liability Directives–Critique of a Half-Hearted Approach and Lessons for the Future. *arXiv preprint arXiv:2211.13960* (2022).

[70] Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating ChatGPT and other large generative AI models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1112–1123.

[71] Rune Klingenberg Hansen. 2022. AI Image Generator: This Is Someone Thinking About Data Ethics · Dataetisk Tænkehandletank. https://dataethics.eu/ai-image-generator-this-is-someone-thinking-about-data-ethics/

[72] Ryuichiro Hataya, Han Bao, and Hiromi Arai. 2022. Will Large-scale Generative Models Corrupt Future Datasets? (2022). https://doi.org/10.48550/ARXIV.2211.08095

[73] Natali Helberger and Nicholas Diakopoulos. 2023. ChatGPT and the AI Act. *Internet Policy Review* 12, 1 (2023).

[74] Dirk Helbing, Bruno S. Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hagner, Yvonne Hofstetter, Jeroen van den Hoven, Roberto V. Zicari, and Andrej Zwitter. 2019. Will Democracy Survive Big Data and Artificial Intelligence? In *Towards Digital Enlightenment: Essays on the Dark and Light Sides of the Digital Revolution*, Dirk Helbing (Ed.). Springer International Publishing, Cham, 73–98. https://doi.org/10.1007/978-3-319-90869-4_7

[75] Alex Hern. 2019. Revealed: Catastrophic effects of working as a facebook moderator. https://www.theguardian.com/technology/2019/sep/17/revealed-catastrophic-effects-working-facebook-moderator

[76] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. (June 2020). https://doi.org/10.48550/arXiv.2006.11239

[77] Adrian Horton. 2023. Marvel faces backlash over ai-generated opening credits. https://www.theguardian.com/tv-and-radio/2023/jun/21/marvel-ai-generated-credits-backlash

[78] Stephanie Houde, Vera Liao, Jacquelyn Martino, Michael Muller, David Piorkowski, John Richards, Justin Weisz, and Yunfeng Zhang. 2020. Business (mis)Use Cases of Generative AI. http://arxiv.org/abs/2003.07679 arXiv:2003.07679 [cs].

[79] Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. Underspecification in Scene Description-to-Depiction Tasks. arXiv. https://doi.org/10.48550/ARXIV.2210.05815

[80] Giulia Interesse. 2022. *China to Regulate Deep Synthesis (Deepfake) Technology Starting 2023.* https://www.china-briefing.com/news/china-to-regulate-deep-synthesis-deep-fake-technology-starting-january-2023/ Accessed: 2023-07-08.

[81] Nina Jankowicz, Sandra Pepera, and Molly Middlehurst. 2021. *Addressing Online Misogyny and Gendered Disinformation: A How-To Guide.* Technical Report. National Democracy Institution.

[82] Alfonsas Jursenas, Kasparas Karlauskas, Gediminas Maskeliunas, and Julius Ruseckas. 2021. *The Double-Edged Sword of AI: Enabler of Disinformation.* Technical Report. Nato Strategic Communications.

[83] Ivan Kapelyukh, Vitalis Vosylius, and Edward Johns. 2022. DALL-E-Bot: Introducing Web-Scale Diffusion Models to Robotics. arXiv:2210.02438 (Nov 2022). http://arxiv.org/abs/2210.02438 arXiv:2210.02438 [cs].

[84] Atoosa Kasirzadeh. 2022. Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. In *AIES '22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society.* 349–356. https://doi.org/10.1145/3514094.3534188

[85] Atoosa Kasirzadeh and Iason Gabriel. 2023. In conversation with Artificial Intelligence: aligning language models with human values. *Philosophy & Technology* 36, 2 (2023), 1–24. https://doi.org/10.1007/s13347-023-00606-x

[86] Atoosa Kasirzadeh and Colin Klein. 2021. The ethical gravity thesis: Marrian levels and the persistence of bias in automated decision-making systems. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 618–626.

[87] Emre Kazim, Osman Güçlütürk, Denise Almeida, Charles Kerrigan, Elizabeth Lomas, Adriano Koshiyama, Airlie Hilliard, and Markus Trengove. 2022. Proposed EU AI Act—Presidency compromise text: select overview and comment on the changes to the proposed regulation. *AI and Ethics* (2022), 1–7.

[88] Dilara Keküllüoğlu, Nadin Kökciyan, and Pınar Yolum. 2016. Strategies for Privacy Negotiation in Online Social Networks. In *Proceedings of the 1st International Workshop on AI for Privacy and Security.* ACM, The Hague Netherlands, 1–8. https://doi.org/10.1145/2970030.2970035

[89] Brandon Khoo, Raphaël C.-W. Phan, and Chern-Hong Lim. 2022. Deepfake attribution: On the source identification of artificially generated images. *WIREs Data Mining and Knowledge Discovery* 12, 3 (2022), e1438. https://doi.org/10.1002/widm.1438 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1438.

[90] Diederik P. Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. https://doi.org/10.48550/arXiv.1312.6114 arXiv:1312.6114 [cs, stat].

[91] Timur Kuran and Edward J. McCaffery. 2004. Expanding Discrimination Research: Beyond Ethnicity and to the Web*. *Social Science Quarterly* 85, 3 (2004), 713–730. https://doi.org/10.1111/j.0038-4941.2004.00241.x

[92] Nadin Kökciyan, Nefise Yaglikci, and Pinar Yolum. 2017. An Argumentation Approach for Resolving Privacy Disputes in Online Social Networks. *ACM Transactions on Internet Technology* 17, 3 (Aug 2017), 1–22. https://doi.org/10.1145/3003434

[93] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (Jul 2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[94] M. Asher Lawson, Shikhar Anand, and Hemant Kakkar. 2023. Tribalism and tribulations: The social costs of not sharing fake news. *Journal of Experimental Psychology: General* 152, 3 (2023), 611–631. https://doi.org/10.1037/xge0001374 Place: US Publisher: American Psychological Association.

[95] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (March 2018), 1094–1096. https://doi.org/10.1126/science.aao2998 Publisher: American Association for the Advancement of Science.

[96] Claire R. Leibowicz, Sean McGregor, and Aviv Ovadya. 2021. The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 736–744. https://doi.org/10.1145/3461702.3462584

[97] David Leonhardt. 2022. 'A Crisis Coming': The Twin Threats to American Democracy. *The New York Times* (Sept. 2022). https://www.nytimes.com/2022/09/17/us/american-democracy-threats.html

[98] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110* (2022).

[99] Natasha Lomas. 2022. Shutterstock to integrate OpenAI's DALL-E 2 and launch fund for contributor artists | TechCrunch. https://techcrunch.com/2022/10/25/shutterstock-openai-dall-e-2/

[100] Juniper Lovato, Laurent Hébert-Dufresne, Jonathan St-Onge, Randall Harp, Gabriela Salazar Lopez, Sean P. Rogers, Ijaz Ul Haq, and Jeremiah Onaolapo. 2022. Diverse Misinformation: Impacts of Human Biases on Detection of Deepfakes on Networks. (2022). https://doi.org/10.48550/ARXIV.2210.10026 Publisher: arXiv Version Number: 2.

[101] Tambiama Madiega and Samy Chahri. 2023. *Artificial intelligence act*. EU Legislation in Progress. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf BRIEFING.

[102] Nina Markl. 2022. Mind the data gap(s): Investigating power in speech and language datasets. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics, Dublin, Ireland, 1–12. https://doi.org/10.18653/v1/2022.ltedi-1.1

[103] Sherin Mathews, Shivangee Trivedi, Amanda House, Steve Povolny, and Celeste Fralick. 2023. An explainable deepfake detection framework on a novel unconstrained dataset. *Complex & Intelligent Systems* (Jan. 2023). https://doi.org/10.1007/s40747-022-00956-7

[104] Gianluca Mauro and Hilke Schellmann. 2023. 'There is no standard': investigation finds AI algorithms objectify women's bodies. *The Guardian* (Feb 2023). https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies

[105] Raphaël Millière. 2022. Adversarial Attacks on Image Generation With Made-Up Words. *ArXiv* (2022). https://doi.org/10.48550/arXiv.2208.04135

[106] Raphaël Millière. 2022. Deep learning and synthetic media. *Synthese* 200, 3 (May 2022), 231. https://doi.org/10.1007/s11229-022-03739-2

[107] Pamela Mishkin, Lama Ahmad, Miles Brundage, Gretchen Krueger, and Girish Sastry. 2022. DALL·E 2 Preview - Risks and Limitations. (2022). [https://github.com/openai/dalle-2-preview/blob/main/system-card.md](https://github.com/openai/dalle-

[108] Jakob Mökander, Prathm Juneja, David S Watson, and Luciano Floridi. 2022. The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other? *Minds and Machines* 32, 4 (2022), 751–758.

[109] Johan Moreno. 2022. Shutterstock Will Soon Offer Licensed DALL-E 2 Images, Showing What The Future Of Generative AI Might Look Like. https://www.forbes.com/sites/johanmoreno/2022/10/26/shutterstock-will-soon-offer-ai-generated-images-showing-what-the-fu

[110] Aakash Varma Nadimpalli and Ajita Rattani. 2022. *GBDF: Gender Balanced DeepFake Dataset Towards Fair DeepFake Detection.*

[111] Paul Nemitz. 2018. Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2133 (Oct. 2018), 20180089. https://doi.org/10.1098/rsta.2018.0089 Publisher: Royal Society.

[112] Alexis Newton and Kaustubh Dhole. 2023. Is AI Art Another Industrial Revolution in the Making? (2023). https://doi.org/10.48550/ARXIV.2301.05133

[113] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M. Nguyen. 2022. Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding* 223 (Oct. 2022), 103525. https://doi.org/10.1016/j.cviu.2022.103525

[114] Alex Nichol. 2022. Dall·E 2 pre-training mitigations. https://openai.com/research/dall-e-2-pre-training-mitigations

[115] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. https://doi.org/10.48550/arXiv.2112.10741 arXiv:2112.10741 [cs].

[116] Nika Nour and Julia Gelfand. 2021. Deepfakes: A Digital Transformation Leads to Misinformation. (2021).

[117] Evgeny Obedkov. 2023. *Game illustrator jobs in China down 70% due to rapid AI adoption.* Game World Observer. https://gameworldobserver.com/2023/04/12/game-artist-jobs-china-down-70-percent-gen-ai-adoption

[118] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. 2016. Conditional Image Generation with PixelCNN Decoders. https://doi.org/10.48550/arXiv.1606.05328 arXiv:1606.05328 [cs].

[119] Jonas Oppenlaender. 2022. The Creativity of Text-to-Image Generation. In *Proceedings of the 25th International Academic Mindtrek Conference (Academic Mindtrek '22).* Association for Computing Machinery, New York, NY, USA, 192–202. https://doi.org/10.1145/3569219.3569352

[120] Donie O'Sullivan and Jon Passantino. 2023. *'Verified' Twitter accounts share fake image of 'explosion' near Pentagon, causing confusion.* CNN. https://edition.cnn.com/2023/05/22/tech/twitter-fake-image-pentagon-explosion/index.html

[121] Britt Paris and Joan Donovan. 2019. DEEPFAKES AND CHEAP FAKES. (2019).

[122] Eli Pariser. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think.* Penguin. Google-Books-ID: wcalrOI1YbQC.

[123] Seongbeom Park, Suhong Moon, and Jinkyu Kim. 2022. Judge, Localize, and Edit: Ensuring Visual Commonsense Morality for Text-to-Image Generation. https://doi.org/10.48550/arXiv.2212.03507 arXiv:2212.03507 [cs].

[124] Billy Perrigo. 2023. OpenAI used Kenyan workers on less than 2 per hour: Exclusive. https://time.com/6247678/openai-chatgpt-kenya-workers/

[125] Muxin Pu, Meng Yi Kuan, Nyee Thoang Lim, Chun Yong Chong, and Mei Kuan Lim. 2022. Fairness Evaluation in Deepfake Detection Models using Metamorphic Testing. In *2022 IEEE/ACM 7th International Workshop on Metamorphic Testing (MET).* 7–14. https://doi.org/10.1145/3524846.3527337

[126] Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. 2022. Are Multimodal Models Robust to Image and Text Perturbations? (2022). https://doi.org/10.48550/ARXIV.2212.08044

[127] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt, Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23).* Association for Computing Machinery, New York, NY, USA, 1882–1895. https://doi.org/10.1145/3593013.3594134

[128] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020 arXiv:2103.00020 [cs].

[129] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. https://doi.org/10.48550/arXiv.2204.06125 arXiv:2204.06125 [cs].

[130] Mehul S Raval, Mohendra Roy, and Minoru Kuribayashi. 2022. Survey on Vision based Fake News Detection and its Impact Analysis. *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (Nov. 2022), 1837–1841. https://doi.org/10.23919/APSIPAASC55919.2022.9980089 Conference Name: 2022 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) ISBN: 9786165904773 Place: Chiang Mai, Thailand Publisher: IEEE.

[131] Scott E. Reed, Zeynep Akata, Xinchen Yan, L. Logeswaran, B. Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. *ArXiv* (May 2016). https://www.semanticscholar.org/paper/6c7f040a150abf21dbcefe1f22e0f98fa184f41a

[132] Karen Renaud, Merrill Warkentin, and George Westerman. 2023. *From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI.* MIT Sloan Management Review.

[133] Yasmim Mendes Rocha, Gabriel Acácio de Moura, Gabriel Alves Desidério, Carlos Henrique de Oliveira, Francisco Dantas Lourenço, and Larissa Deadame de Figueiredo Nicolete. 2021. The impact of fake news on social media and its influence on health during the COVID-19 pandemic: a systematic review. *Zeitschrift Fur Gesundheitswissenschaften* (Oct. 2021), 1–10. https://doi.org/10.1007/s10389-021-01658-z

[134] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* 10684–10695.

[135] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. http://arxiv.org/abs/2205.11487 arXiv:2205.11487 [cs].

[136] Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J. Jansen. 2020. Analyzing Demographic Bias in Artificially Generated Facial Pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382791

[137] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3411763.3451807

[138] Pamela Samuelson. 2023. Legal Challenges to Generative AI, Part I. *Commun. ACM* 66, 7 (2023), 20–23.

[139] Glorin Sebastian. 2023. Do ChatGPT and other AI chatbots pose a cybersecurity risk?: An exploratory study. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)* 15, 1 (2023), 1–11.

[140] Oshani Seneviratne. 2022. Blockchain for Social Good: Combating Misinformation on the Web with AI and Blockchain. *14th ACM Web Science Conference 2022* (June 2022), 435–442. https://doi.org/10.1145/3501247.3539016 Conference Name: WebSci '22: 14th ACM Web Science Conference 2022 ISBN: 9781450391917 Place: Barcelona Spain Publisher: ACM.

[141] Sachith Seneviratne, Damith Senanayake, Sanka Rasnayaka, Rajith Vidanaarachchi, and Jason Thompson. 2022. DALLE-URBAN: Capturing the urban design expertise of large text to image transformers. arXiv:2208.04139 (Oct 2022). http://arxiv.org/abs/2208.04139 arXiv:2208.04139 [cs].

[142] Jia Wen Seow, Mei Kuan Lim, Raphaël C.W. Phan, and Joseph K. Liu. 2022. A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing* 513 (Nov. 2022), 351–371. https://doi.org/10.1016/j.neucom.2022.09.135

[143] Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond Fair Pay: Ethical Implications of NLP Crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 3758–3769. https://doi.org/10.18653/v1/2021.naacl-main.295

[144] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation Is not a Design Fix for Machine Learning. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Arlington VA USA, 1–6. https://doi.org/10.1145/3551624.3555285

[145] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. https://doi.org/10.48550/arXiv.1503.03585 arXiv:1503.03585 [cond-mat, q-bio, stat].

[146] Irene Solaiman. 2023. The gradient of generative AI release: Methods and considerations. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 111–122.

[147] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2022. Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models. https://doi.org/10.48550/arXiv.2212.03860 arXiv:2212.03860 [cs].

[148] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. (Nov. 2020). https://doi.org/10.48550/arXiv.2011.13456

[149] Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2022. Rickrolling the Artist: Injecting Invisible Backdoors into Text-Guided Image Generation Models. (2022). https://doi.org/10.48550/ARXIV.2211.02408

[150] Naofse Mac Sweeney. 2009. Beyond Ethnicity: The Overlooked Diversity of Group Identities. *Journal of Mediterranean Archaeology* 22, 1 (Jun 2009), 101–126. https://doi.org/10.1558/jmea.v22i1.101

[151] Dean Takahashi. 2023. AI Games and AI Film Festival will highlight how generative AI is taking root. https://venturebeat.com/games/ai-games-and-ai-film-festival-will-highlight-how-generative-ai-is-taking-root/

[152] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion* 64 (Dec. 2020), 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

[153] Nenad Tomasev, Jonathan Leader Maynard, and Iason Gabriel. 2022. Manifestations of Xenophobia in AI Systems. (2022). https://doi.org/10.48550/ARXIV.2212.07877 Publisher: arXiv Version Number: 1.

[154] Chad M. Topaz, Jude Higdon, Avriel Epps-Darling, Ethan Siau, Harper Kerkhoff, Shivani Mendiratta, and Eric Young. 2022. Race- and gender-based under-representation of creative contributors: art, fashion, film, and music. *Humanities and Social Sciences Communications* 9, 11 (Jun 2022), 1–11. https://doi.org/10.1057/s41599-022-01239-9

[155] Inga Trauthig. 2022. WhatsApp, Misinformation, and Latino Political Discourse in the U.S. https://techpolicy.press/whatsapp-misinformation-and-latino-political-discourse-in-the-u-s/

[156] Daniel Trotta and Brendan Pierson. 2023. US judges halt healthcare bans for Transgender Youth. https://www.reuters.com/legal/us-judges-halt-healthcare-bans-transgender-youth-2023-07-03/

[157] Eddie L. Ungless, Björn Ross, and Anne Lauscher. 2023. Stereotypes and Smut: The (Mis)representation of Non-cisgender Identities by Text-to-Image Models. arXiv:2305.17072 (May 2023). http://arxiv.org/abs/2305.17072 arXiv:2305.17072 [cs].

[158] Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society* 6, 1 (Jan. 2020), 2056305120903408. https://doi.org/10.1177/2056305120903408 Publisher: SAGE Publications Ltd.

[159] Henriikka Vartiainen and Matti Tedre. 2023. Using artificial intelligence in craft education: crafting with text-to-image generative models. *Digital Creativity* (2023), 1–21.

[160] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. (June 2017). https://doi.org/10.48550/arXiv.1706.03762

[161] Michael Veale, Kira Matus, and Robert Gorwa. 2023. AI and Global Governance: Modalities, Rationales, Tensions. *Annual Review of Law and Social Science* 19 (2023). https://doi.org/10.1146/annurev-lawsocsci-020223-040749 Review in Advance first posted online on June 28, 2023. (Changes may still occur before final publication.).

[162] Luisa Verdoliva. 2020. Media Forensics and DeepFakes: an overview. http://arxiv.org/abs/2001.06564 arXiv:2001.06564 [cs].

[163] Nikhil Vyas, Sham Kakade, and Boaz Barak. 2023. Provable Copyright Protection for Generative Models. arXiv:2302.10870 (Feb 2023). http://arxiv.org/abs/2302.10870 arXiv:2302.10870 [cs, stat].

[164] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. 2022. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. *International Journal of Computer Vision* 130, 7 (Jul 2022), 1790–1810. https://doi.org/10.1007/s11263-022-01625-5

[165] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. 2022. Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting. https://doi.org/10.48550/arXiv.2212.06909 arXiv:2212.06909 [cs].

[166] Jess Weatherbed. 2023. Levi's will test AI-generated clothing models to "increase diversity". https://www.theverge.com/2023/3/27/23658385/levis-ai-generated-clothing-model-diversity-denim

[167] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. https://doi.org/10.48550/arXiv.2112.04359 arXiv:2112.04359 [cs].

[168] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 214–229. https://doi.org/10.1145/3531146.3533088

[169] Justin D. Weisz, Michael Muller, Jessica He, and Stephanie Houde. 2023. Toward General Design Principles for Generative AI Applications. (2023). https://doi.org/10.48550/ARXIV.2301.05578 Publisher: arXiv Version Number: 1.

[170] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. *Discriminating Systems: Gender, Race and Power in AI.* Retrievedfromhttps://ainowinstitute.org/discriminatingsystems.html.

[171] Mika Westerlund. 2019. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review* 9, 11 (2019), 40–53. https://doi.org/10.22215/timreview/1282 Place: Ottawa Publisher: Talent First Network.

[172] David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. Limits and Possibilities for &#x201c;Ethical AI&#x201d; in Open Source: A Study of Deepfakes. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2035–2046. https://doi.org/10.1145/3531146.3533779

[173] Kyle Wiggers. 2023. The current legal cases against generative AI are just the beginning. https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning/

[174] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. 2022. Contrastive Language-Vision AI Models Pre-trained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias. arXiv:2212.11261 (Dec 2022). http://arxiv.org/abs/2212.11261 arXiv:2212.11261 [cs].

[175] Ying Xu, Philipp Terhörst, Kiran Raja, and Marius Pedersen. 2022. A Comprehensive Analysis of AI Biases in DeepFake Detection With Massively Annotated Databases. (2022). https://doi.org/10.48550/ARXIV.2208.05845 Publisher: arXiv Version Number: 1.

[176] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2022. Diffusion Models: A Comprehensive Survey of Methods and Applications. https://doi.org/10.48550/ARXIV.2209.00796

[177] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. (2022). https://doi.org/10.48550/ARXIV.2206.10789 Publisher: arXiv Version Number: 1.

[178] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. 2020. Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. (July 2020). https://doi.org/10.48550/arXiv.2007.08457

[179] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2979–2989. https://doi.org/10.18653/v1/D17-1323

[180] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3544548.3581318

A TAXONOMY METHODOLOGY

We conducted our searches utilising the Semantic Scholar API. Semantic Scholar index over 200 million academic papers. To capture relevant papers we selected five seed papers covering biased training data, biased image generation and bias in text-to-image models [8, 15, 20, 37, 136]. To capture papers relevant to misinformation harms, we selected three papers relevant to either deep fakes or synthetic media [152, 171] or diffusion technology and evaluation [176]. Our search returned over 300 papers. 43 of these papers provided substantial and useful discussions of text-to-image technologies. Through extensive manual searches we identified a further 40 papers, most of which were technical papers. Collected papers were then analysed for stakeholders, risks, empirical investigations and open research questions.

Our taxonomy of risks initially adopted an inductive-deductive approach, in that we preempted the existence of three broad categories (discrimination and exclusion, harmful misuse, misinformation) and derived subcategories from analysis of the papers. We then retroactively identified potential "gaps" in the literature, based in part on analogous research into the harms of other technologies, plus identifying key stakeholders that have not been addressed. These gaps are clearly identified in the table.