A Pathway Towards Responsible AI Generated Content

Chen Chen[1] , **Jie Fu**[2] , **Lingjuan Lyu**[1*]

[1]Sony AI   [2]BAAI

{ChenA.Chen,Lingjuan.Lv}@sony.com
fujie@baai.ac.cn

Abstract

AI Generated Content (AIGC) has received tremendous attention within the past few years, with content generated in the format of image, text, audio, video, etc. Meanwhile, AIGC has become a double-edged sword and recently received much criticism regarding its responsible usage. In this article, we focus on 8 main concerns that may hinder the healthy development and deployment of AIGC in practice, including risks from (1) privacy; (2) bias, toxicity, misinformation; (3) intellectual property (IP); (4) robustness; (5) open source and explanation; (6) technology abuse; (7) consent, credit, and compensation; (8) environment. Additionally, we provide insights into the promising directions for tackling these risks while constructing generative models, enabling AIGC to be used more responsibly to truly benefit society.

1 Introduction

Foundation models for generative AI. The success of high-quality AI Generated Content (AIGC) is strongly correlated with the emergence and rapid advancement of large foundation models. These models, with their vast capacity, enable the rapid development of domain-specific models, which are commonly employed for the production of various types of content, including images, texts, audio, video, etc. For instance, many text generators are built on the Generative Pre-trained Transformer (GPT) [Radford *et al.*, 2018] or its derivatives, such as GPT-2 [Radford *et al.*, 2019], GPT-3 [Brown *et al.*, 2020], GPT-3.5, GPT-4, etc. Similarly, numerous text-to-image generators rely on vision-language models such as CLIP [Radford *et al.*, 2021], Open-CLIP [Wortsman *et al.*, 2022], etc.

AIGC applications. In recent years, generative modeling has made rapid advances and tremendous progress. OpenAI's DALL·E [Ramesh *et al.*, 2021] was one of the first text-to-image models that had captured widespread public attention. It is trained to generate digital images from text descriptions, referred to as "prompts", using a dataset of text–image
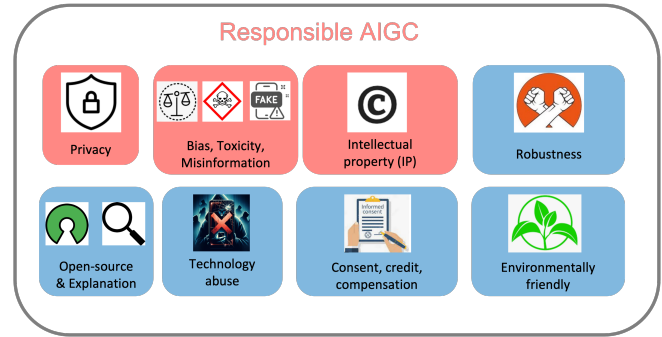

Figure 1: The scope of responsible AIGC. Note that some icons are from Shutterstock.

pairs [Brown *et al.*, 2020]. Its successor, DALL·E 2 [Ramesh *et al.*, 2022], which can generate more complex and realistic images, was unveiled in April 2022, followed by Stable Diffusion [Rombach *et al.*, 2022a], which was publicly released in August 2022. Google, as a rival to OpenAI, presented two text-to-image models that can generate photorealistic images: the diffusion-based model Imagen [Saharia *et al.*, 2022a], and the Pathways Autoregressive Text-to-Image model (Parti) [Yu *et al.*, 2022]. In addition to text-to-image tasks, diffusion models had been widely used for image-to-image [Saharia *et al.*, 2022b; Whang *et al.*, 2022] and text-to-video models, such as Runway [Runway, 2022], Make-A-Video [Singer *et al.*, 2022], Imagen Video [Ho *et al.*, 2022], and Phenaki [Villegas *et al.*, 2022]. Stable Diffusion has been adapted for various applications, from medical imaging [Chambon *et al.*, 2022] to music generation [Agostinelli *et al.*, 2023].

In addition to image and video generation, text generation is a popular generative domain. OpenAI's GPT-3 [Brown *et al.*, 2020] is a notable example of a large language model (LLM). With a simple text prompt, GPT-3 can produce a piece of writing or an entire essay. It can also assist programmers in writing code. OpenAI has further developed GPT-3.5, an improved version which is better at generating complex text and poetry. In 2022, OpenAI launched ChatGPT [OpenAI, 2022], a 175 billion parameter natural language processing (NLP) model that can produce responses in a conversational style. This model combines two popular AI topics: chatbots and GPT-3.5. ChatGPT is a specific chatbot

use case wherein the chatbot interacts with a GPT information source. The most recent version of ChatGPT integrated GPT-4 [OpenAI, 2023] – OpenAI's most advanced system, which can produce safer and more useful responses.

AIGC dispute. Despite its popularity, AIGC has raised a range of concerns such as privacy, bias, toxicity, misinformation, intellectual property (IP), and potential misuse of technology. The recent release of ChatGPT has sparked much conversation surrounding its capabilities and potential risks, such as its ability to debug code or compose essays for students [Elliot and DeLisi, 2022]. It is important to consider whether AIGC can be counted as unique creative works or simply replicate content from their training sets. Ideally, AIGC should produce original and distinct outputs, but the source and IP rights of the training data are often unknown due to the use of uncurated web-scale data [Somepalli et al., 2022]. Furthermore, the powerful memorization of large AIGC models [Carlini et al., 2022; Carlini et al., 2021] poses a risk of reproducing data directly from the training data [Butterick, 2023], which potentially violates privacy rights and raises legal concerns around copyright infringement and ownership. In addition to the aforementioned privacy and IP issues, as most AIGC models rely on text encoders that are trained on large amounts of data from the internet, hence these learned models may inherent social biases, toxicity, and produce misinformation.

Components in responsible AIGC. The essential components of responsible AIGC are summarized in Figure 1. Table 1 lists recent AIGC models and their associated issues related to privacy, bias, toxicity, misinformation, and IP, noting which models have taken proactive actions.

2 Privacy

2.1 Privacy leakage in foundation models

Large foundation models are known to be vulnerable to privacy risks, and it is possible that AIGC models that build upon these models could also be subject to privacy leakage. Previous research has demonstrated that large language models such as GPT-2 can be vulnerable to privacy attacks, as attackers can generate sequences from the trained model and identify those memorized from the training set [Carlini et al., 2021]. Kandpal et al. [Kandpal et al., 2022] have attributed the success of these privacy attacks to the presence of duplicated data in commonly used web-scraped training sets. It has been demonstrated that a sequence that appears multiple times in the training data is more likely to be generated than a sequence that occurred only once. This suggests that deduplication could be used as a potential countermeasure in privacy-sensitive applications.

2.2 Privacy leakage in generative models

The replication behavior in Generative Adversarial Networks (GANs) has been studied extensively [Meehan et al., 2020; Feng et al., 2021; Webster et al., 2021]. Due to the fact that AIGC models are trained on large-scale web-scraped data [Rombach et al., 2022a; Ramesh et al., 2022; Saharia et al., 2022a], the issue of overfitting and privacy leakage becomes especially relevant. For instance, the model card of Stable Diffusion recognized that it memorized duplicate images in the training data [Rombach et al., 2022c]. Somepalli et al. [Somepalli et al., 2022] also demonstrated that Stable Diffusion blatantly copies images from its training data, and the generated images are simple combinations of the foreground and background objects of the training dataset. Moreover, the system occasionally displays the ability to reconstruct memories, producing objects that are semantically equivalent to the original without being identical in pixel form. The existence of such images raises concerns about data memorization and the ownership of diffusion images.

Similarly, Melissa Heikkilä [Heikkilä, 2023] reported that Google's Imagen can leak photos of real people and copyrighted images. In Matthew Butterick's recent litigation [Butterick, 2023], he pointed out that because all visual information in the system is derived from copyrighted training images, the images produced are necessarily works derived from those training images, regardless of their outward appearance. DALL·E 2 also encountered similar problems. It can sometimes reproduce images from its training data rather than creating new ones. OpenAI found that this image regurgitation occurs due to images being replicated many times in the dataset [Nichol, 2022]. Similarly, when we asked ChatGPT "What is the privacy risk of ChatGPT", it responded with multiple potential privacy risks, as illustrated in Figure 3.

2.3 Privacy actions

Although a complete resolution to the privacy issues mentioned above has not been achieved, companies and researchers have taken proactive steps to address these issues, such as introducing warning messages and detecting replicated content.

At the industry level, Stability AI has recognized the limitations of Stable Diffusion, such as the potential for memorization of replicated images in the training data. To address this, they provide a website [Beaumont, 2022] to support the identification of such memorized images. In addition, art company Spawning AI has created a website called "Have I Been Trained" [1] to assist users in determining whether their photos or works have been used as AI training materials. OpenAI has taken steps to address privacy concerns by reducing data duplication through deduplication [Nichol, 2022]. Furthermore, companies such as Microsoft and Amazon have implemented measures to prevent employee breaches of confidentiality by banning the sharing of sensitive data with ChatGPT, given that this information could be utilized for training data for future versions of ChatGPT [Lopez, 2023]. At the academic level, researchers [Somepalli et al., 2022] have studied image retrieval frameworks to identify content duplication, while Dockhorn et al. [Dockhorn et al., 2022] have proposed differentially private diffusion models to guarantee privacy in generative models. Zhuang et al. [Zhuang et al., 2023] proposed to adopt federated learning for the privacy-preserving and responsible development of foundation models.

Existing privacy measures may not be inadequate to meet the demands of privacy. It is essential to explore more reliable

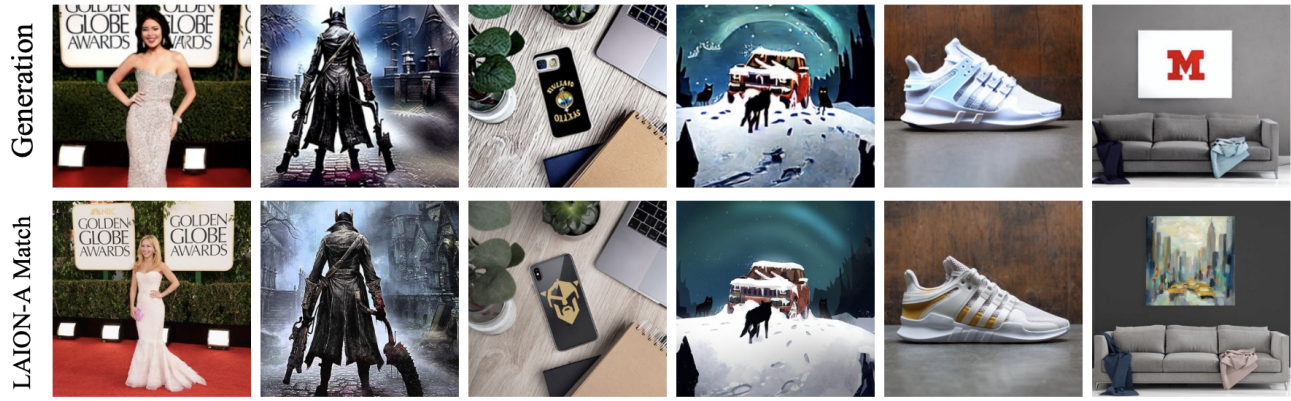---

[1] https://haveibeentrained.com

Figure 2: A comparison between training images and generated images (by Stable Diffusion). **Top row**: generated images. **Bottom row**: closest matches in the training dataset (LAION). The comparison shows that Stable Diffusion is able to replicate training data by combining foreground and background objects. Image source: [Somepalli *et al.*, 2022].
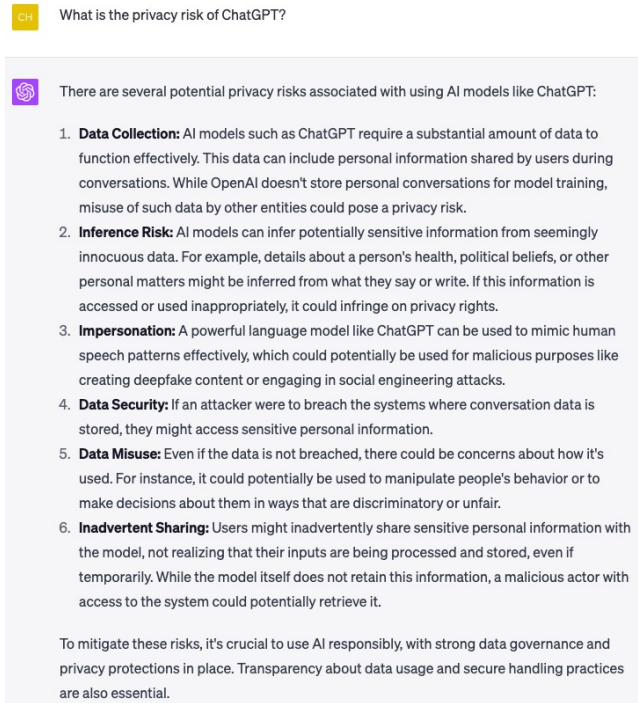

Figure 3: An answer to "What is the privacy risk of ChatGPT" by ChatGPT (GPT-4, May 12, 2023 version).

detection systems for data replication in generative models, and to further investigate memorization and generalization in current and future AIGC models. Designing more faithful metrics for the privacy assessment on the reconstructed or generated images is also worthwhile to explore [Sun *et al.*, 2023c].

3 Bias, toxicity, misinformation

3.1 Problematic datasets

Since the training data used in AI models are collected in the real world, they can unintentionally reinforce harmful stereotypes, exclude or marginalize certain groups, and contain toxic data sources, which can incite hate or violence and offend individuals [Weidinger *et al.*, 2021]. For example, the LAION dataset [Schuhmann *et al.*, 2021], which is used to train diffusion models, has been criticized for containing problematic content related to social stereotyping, pornography, racist slurs, and violence.

Although some AIGC models like Imagen [Saharia *et al.*, 2022a] try to filter out undesirable data, such as pornographic imagery and toxic language, the filtered data can still contain sexually explicit or violent content. Moreover, recent research works [Prabhu and Birhane, 2020; Birhane *et al.*, 2021] have pointed out that these unfiltered datasets utilized for training frequently encompass social biases, repressive perspectives, and derogatory connections towards underrepresented communities. Google's Imagen Video [Ho *et al.*, 2022] is trained on a combination of the LAION-400M image-text dataset and their internal dataset, and Google is concerned that its Imagen tool could be used to generate harmful content. Meanwhile, this dataset inherits social biases and stereotypes that are difficult to remove.

3.2 Problematic AIGC models

Models trained, learned, or fine-tuned on the aforementioned problematic datasets without mitigation strategies can inherit harmful stereotypes, social biases, and toxicity, leading to unfair discrimination and harm to certain social groups [Weidinger *et al.*, 2021].

Table 1: A summary of recent AIGC models and associated issues. We use dots with different colors to indicate different modalities involved in the models: Text, Image, Video.

Models	Developer(s)	Initial release	Format	Main technique	Release to public by Mar, 2023	Privacy	Bias, toxicity, misinformation	IP
●● DALL·E, DALL·E 2	OpenAI	Jan, 2021/ Apr, 2022	Text-to-image	CLIP, diffusion model	No	Deduplication	Data filtering and reweighting	—
●● Craiyon (DALL·E Mini)	Boris Dayma et al.	Jul, 2021	Text-to-image	CLIP, diffusion model	No	Deduplication	—	—
●● Stable Diffusion	CompVis; Runway; Stability AI	Aug, 2022	Text-to-image	CLIP, diffusion model	Yes	—	Data filtering	—
● ChatGPT	OpenAI	Dec, 2022	Text-to-text	GPT-3.5, reinforcement learning	No	Refusing to provide private information (e.g., phone number)	Data filtering, building tools to screen harmful model outputs, etc.	Classifier
●● Point-E	OpenAI	Dec, 2022	Text-to-3D model	GLIDE, diffusion model	No	—	—	—
●● Midjourney's algorithm	Midjourney	Mar, 2022	Text-to-image	Unknown	No	—	—	DMCA takedown policy
●● Imagen	Google Brain	Dec, 2022	Text-to-image	BERT, T5, CLIP, diffusion model	No	—	Data filtering	—
●● Parti	Google Brain	Dec, 2022	Text-to-image	ViT-VQGAN, autoregressive model	No	—	Prompt filtering, output filtering, and model recalibration	Adding watermark
●●● Video diffusion, Imagen Video	Google Brain	Dec, 2022	Text-to-video	Diffusion model	No	—	Prompt filtering and output filtering	—
●●● Make-A-Video	Meta	Dec, 2022	Text-to-video	CLIP, Pseudo-3D convolutions, diffusion model	No	—	Data filtering	Adding watermark
●● CogView, CogView 2	Tsinghua University, Alibaba, BAAI	May, 2021	Text-to-image	VQVAE, autoregressive model	No	—	—	—
●●● CogVideo	Tsinghua University, BAAI	May, 2022	Text-to-video	CogView 2	No	—	—	—

For example, Stable Diffusion v1 was trained primarily on the LAION-2B data set, which only contains images with English descriptions [Rombach *et al.*, 2022c]. As a result, the model was biased towards white, western cultures, and prompts in other languages may not be adequately represented. Follow-up versions of the Stable Diffusion model were fine-tuned on the filtered versions of the LAION dataset, but the bias issue still occurs [Rombach *et al.*, 2022b]. To illustrate the inherent bias in Stable Diffusion, we tested a toy example on Stable Diffusion v2.1. As shown in Figure 4, images generated with the prompt "Three engineers running on the grassland" were all male and none of them belong to the neglected racial minorities, indicating a lack of diversity in the generated images.

Similarly, DALLA·E and DALLA·E 2 exhibited negative stereotypes against minoritized groups [Johnson, 2022]. Google's Imagen [Saharia *et al.*, 2022a] also encoded several social biases and stereotypes, such as generating images of people with lighter skin tones and aligning with western gender stereotypes. These biases can lead to unfair discrimination and harm to certain social groups. Even when generating non-human images, Imagen has been shown to encode social and cultural biases [Miller, 2022]. Due to these issues, most companies decided not to make their AIGC models available to the public to avoid criticism and potential fine from government.

Figure 4: Images generated with the text "Three engineers running on the grassland" by Stable Diffusion v2.1. There are 28 people in the 9 images, all of them are male. Moreover, none of them belong to the neglected racial minorities. This shows a huge bias of Stable Diffusion v2.1.

Beyond above issues, there is also a risk of misinformation when AIGC models provide inaccurate, false even harmful answers or responses [Sun *et al.*, 2023a; Xie *et al.*, 2023]. For example, ChatGPT and its derivatives are notorious for their hallucination issues, i.e., the generated content may appear to be accurate and authoritative, but it could be completely inaccurate. Therefore, it can be used for misleading purposes in schools, laws, medical domains, weather forecasting, and anywhere else. For example, the answer on medical dosages that ChatGPT provides could be inaccurate or incomplete, potentially leading to the user taking dangerous or even life-threatening actions [Bickmore *et al.*, 2018]. Prompted misinformation on traffic laws could cause accidents and even death if drivers follow the false traffic rules. ChatGPT also exhibits verbosity and overuse of certain phrases. For instance, it repeatedly states that it is a language model trained by OpenAI. These issues are due to biases inherent in training data, as trainers tend to prefer longer answers that appear more comprehensive [OpenAI, 2022].

3.3 Bias, toxicity, misinformation mitigation

The quality of the content generated by AIGC models is inextricably linked to the quality of the training corpora. One noticeable point is that while problems such as biases, toxicity and stereotypes can be reduced in the source datasets, they can still be propagated even exacerbated during the training and development of AIGC models. For example, although some companies such as Google try to filter out undesirable data before training Imagen [Saharia *et al.*, 2022a], such as pornographic imagery and toxic language, the filtered data can still contain sexually explicit or violent content. OpenAI took extra measures to ensure that any violent or sexual content was removed from the training data for DALLA·E 2 by carefully filtering the original training dataset. However, filtering can introduce biases into the training data that can then be propagated to the downstream models. To address this issue, OpenAI developed pre-training techniques to mitigate the consequent filter-induced biases [Nichol, 2022]. Overall, it is crucial to evaluate the existence of bias and toxicity throughout the entire lifecycle of data usage, rather than staying solely at the data source level. Additionally, there is a challenge in defining a truly fair and non-toxic dataset. The extent and nature of these issues within AIGC models have not yet been comprehensively investigated.

In terms of misinformation and hallucination prevention, it is vital to analyze the root reasons behind them. AI hallucinations can occur for several reasons [Alston, 2023], including: (1) Insufficient, outdated, or low-quality training data. An AI model is only as good as the data it's trained on. If the AI tool doesn't understand the input prompt or doesn't have sufficient information, it'll rely on the limited dataset it's been trained on to generate a response—even if it's inaccurate; (2) Overfitting. When an AI model is trained on a limited dataset, it may memorize the inputs and appropriate outputs. This leaves it unable to effectively generalize new data, resulting in AI hallucinations; (3) Use of idioms or slang expressions. If a prompt contains an idiom or slang expression that the AI model hasn't been trained on, it may lead to nonsensical outputs; (4) Adversarial attacks. Prompts that are deliberately designed to confuse the AI can cause it to produce AI hallucinations.

To defend against misinformation and hallucination, Elena [Alston, 2023] recommended 6 ways to prevent AI hallucinations, including (1) Limit the possible outcomes; (2) Pack in relevant data and sources unique to you; (3) Create a data template for the model to follow; (4) Give the AI a specific role—and tell it not to lie; (5) Tell it what you want—and what you don't want; (6) Experiment with the temperature which controls the randomness of model results. Sun *et al.* [Sun *et al.*, 2023a] recently adopted the self-verification strategy to address the hallucination issue of LLMs. Xie *et al.* [Xie *et al.*, 2023] proposed the psychologically inspired self-reminder technique that can efficiently and effectively mitigate against jailbreaks without further training.

It is also essential to regularly update the training corpora

used by AIGC models with the most recent information to ensure that AI-driven models reflect the current state of society, thus reduce misinformation and hallucination. This will help prevent information lag and hallucination, and ensure that the models remain updated, relevant, and beneficial to society. Lazaridou *et al.* [Lazaridou *et al.*, 2021] showed that transformer models cannot accurately predict data that did not fall into training data period. This is because test data and training data come from different periods, and increasing model size does not improve performance. It is thus essential to incorporate new training data and update the model regularly. Gathering user feedback is also an effective way to keep models updated. Companies such as OpenAI actively seek feedbacks from users to identify harmful outputs that could arise in real-world scenarios, as well as to uncover and mitigate novel risks [OpenAI, 2022] in a timely manner. Actually, GPT-4 had incorporated an additional safety reward signal during Reinforcement Learning from Human Feedback (RLHF) training to reduce harmful outputs by training the model to refuse requests for such content [OpenAI, 2023]. By involving users in the feedback loop, AIGC developers can better understand the potential consequences of their models and take corrective actions to minimize any negative impacts.

4 IP Protection

As AIGC continues to advance in sophistication and popularity, it raises questions about the origin of content for copyright purposes and whether AI-generated content should be entitled to the same intellectual property protections as content created by humans.

4.1 Difficulty of IP infringement detection

Traditional understanding of copyright. Copyright law generally protects original works of authorship that are created by human authors and are fixed in a tangible form [Office, 2023]. For a work to be eligible for copyright protection, it needs to be expressed in a tangible form, either physical or digital, such as a book, painting, or computer file.

Difficulty of copyright definition in AIGC. The ownership and protection of generated content have raised a significant amount of concern and debate. It remains unclear whether such generated content should be considered original works eligible for copyright protection under current laws.

There are many different notions of replication from AIGC. Somepalli *et al.* [Somepalli *et al.*, 2022] gave an (informal) definition as follows: *An image is considered to contain replicated content if it includes an object that is identical to an object in a training image, regardless of minor variations in appearance resulting from data augmentation, whether the object is in the foreground or background.*

In fact, addressing AI copyright issues is a complex task that involves several factors, including: (1) unclear regulations on data collection, usage, rights confirmation, and commercial use of data; (2) the need for a fair benefit distribution mechanism for contributors; (3) the lack of a unified legal understanding of AIGC copyright worldwide, with disputes over ownership still unresolved; and (4) difficulties in identifying all original works used to train AIGC models, as these models can generate an unlimited amount of content, making it impossible to test all of it.

4.2 IP infringement examples

There is a risk of copyright infringement with the generated content if it copies existing works, whether intentionally or not, raising legal questions about IP infringement.

In November 2022, Matthew Butterick filed a class action lawsuit against Microsoft's subsidiary GitHub, accusing that their product Copilot, a code-generating service, violated copyright law [Butterick, 2022]. The lawsuit centers around Copilot's illegal use of licensed code sections from the internet without attribution. Texas A&M professor Tim Davis also provided examples of his code being copied verbatim by Copilot [Jennings, 2022]. Although Microsoft and OpenAI have acknowledged that Copilot is trained on open-source software in public GitHub repositories, Microsoft claims that the output of Copilot is merely a series of code "suggestions" and does not claim any rights in these suggestions. Microsoft also does not make any guarantees regarding the correctness, security, or copyright of the generated code.

In addition to code generation, text-to-image generative models like Stable Diffusion also faced accusations of infringing on the creative work of artists, as they are trained on billions of images from the Internet without the approval of the IP holders, which some argue is a violation of their rights. This is evident in Stable Diffusion, which has generated images with the Getty Images' watermark on them [Vincent, 2023]. Somepalli *et al.* [Somepalli *et al.*, 2022] also presented evidence suggesting that Stable Diffusion copies from the data on which it was trained on. While Stable Diffusion disclaims any ownership of generated images and allows users to use them freely as long as the image content is legal and non-harmful, this freedom raises questions about ownership ethics.

4.3 IP problem mitigation

To mitigate IP concerns, many companies have started implementing measures to accommodate content creators. Midjourney, for instance, has added a DMCA takedown policy to its terms of service, allowing artists to request the removal of their work from the dataset if they suspect copyright infringement [Midjourney, 2022]. Similarly, Stability AI plans to offer artists the option of excluding themselves from future versions of Stable Diffusion [Heikkilä, 2022a]. OpenAI has released a classifier that can distinguish between text generated by AI and that written by humans. However, this tool should not be relied exclusively on for critical decisions.

In addition to above attempts, watermarks can be extremely useful in tracking IP violations or detecting the origin of the generated content [He *et al.*, 2022a; He *et al.*, 2022b; Peng *et al.*, 2023]. Wang *et al.* [Wang *et al.*, 2023b; Wang *et al.*, 2023a] recently utilized novel watermark techniques to conduct origin attribution of AI-generated images and detect the unauthorized data usages in text-to-image diffusion models. In light of the growing popularity of AIGC, the need for watermarking is becoming increasingly pressing. OpenAI is developing a watermark to identify text generated by its GPT model. It could be a valuable tool for educators

and professors to detect plagiarism in assignments generated with such tools. Google has already applied a Parti watermark to all images it releases. John Kirchenbauer *et al.* [Kirchenbauer *et al.*, 2023] proposed a watermark to detect whether the text is generated by an AI model. Still, they only tested it on the smaller open-source language model OPT-6.7B from Meta, leaving its performance on the larger and more widely used ChatGPT model unknown.

In general, the emergence of AIGC presents significant IP concerns and challenges that demand immediate attention. It is essential for technologists, lawyers, and policymakers to recognize these issues and work together to ensure that the intellectual property rights of human creators are protected.

5 Robustness

Previous studies have demonstrated that large models or foundation models trained on the unlabelled data can be backdoored [Pan *et al.*, 2023; Shejwalkar *et al.*, 2023]. This poisoning effect could cause catastrophic damage to downstream applications that depend on the compromised foundation or generative models. For example, a diffusion model with a hidden "backdoor" could carry out malicious actions when it encounters a specific trigger pattern during data generation [Chou *et al.*, 2022; Zhang *et al.*, 2022; Sun *et al.*, 2023b]. How to sift out clean data for training [Zeng *et al.*, 2022] matters a lot for model robustness.

Beyond the poisoning attack during training phase, the recent emergence of jailbreak attacks [Daryanani, 2023; Albert, 2023] which use adversarial prompts to bypass the deployed ChatGPT's ethics safeguards and engender harmful responses notably threatens the responsible and secure use of ChatGPT [Xie *et al.*, 2023]. Xie *et al.* [Xie *et al.*, 2023] proposed the psychologically inspired self-reminder technique that can efficiently and effectively mitigate against jailbreaks without further training. Unfortunately, research on the robustness of foundational and fine-tuned models is still limited.

6 Responsible Open Source and Explanation

Lack of transparency of models behind AIGC can lead to a series of unsatisfactory results. It is frequently challenging to explain why and how the models generate different content and determine the information used to generate a model's output. For example, social and cultural bias is introduced and potentially amplified at multiple stages of model development and deployment. However, how the biases are propagated through these models remain unclear. Similarly, while deduplication can be an effective method of preventing memorization, it does not completely explain why or how models like DALL·E 2 memorize training data. As most of the code and models behind AIGC are not transparent to the public, and their downstream applications are diverse and may have complex societal impacts, it is challenging to determine the potential harms they may cause. Therefore, the need for responsible open source becomes critical in determining whether the benefits of AIGC outweigh its potential risks in specific use cases. Open-sourcing can also facilitate explanation of the behaviours of the models behind AIGC.

Currently, most companies chose not to release their models or open-source their code before solving all the potential risks associated with their models. OpenAI has been criticized for not sharing more about how the most recent GPT-4 was created. Stable Diffusion [Rombach *et al.*, 2022b] and Meta's LLAMA [Meta, 2023] are few generative AI models that provide the source code and pretrained model (weights). The risk is that anyone can use these open-sourced models for free, even for commercial or malicious purposes. To promote a healthy open-sourcing environment, communities have put a lot of joint efforts. In Dec, 2023, IBM And Meta Launch the AI Alliance for safe and open AI with the belief that open and transparent innovation is crucial for harnessing AI advancements in a way that prioritizes safety, diversity, and widespread economic opportunity [Forbes, 2023].

7 Limit Technology Abuse

AIGC can be used for malicious purposes such as spreading fake news, hoaxes, and harassment. The foundation models that power AIGC have made it easier and cheaper to create deepfakes that are close to the original, posing additional risks and concerns. In fact, many models are still far from satisfactory and some of them have gained negative reputations for producing useless, biased, or harmful information.

For example, on the 4chan online forum, there are numerous discussions about images of naked celebrities and other forms of fake pornographic content generated by Stable Diffusion [Wiggers, 2022a]. The misuse of these technologies could lead to the spread of misinformation, harm the reputations of individuals, or even break the law. The potential negative impact of ChatGPT on education is significant, as students could use it to write homework or solve math problems, thus compromising the integrity of their work. Moreover, as ChatGPT is a chatbot, it lacks the necessary emotional connection that a human teacher can provide, which could lead to a diminished learning experience. In light of these concerns, New York City public schools have recently banned the use of ChatGPT [Rosenblatt, 2022]. Stack Overflow, a Q&A platform for coders and programmers, temporarily prohibited the sharing of ChatGPT information, acknowledging its potential to cause significant harm to the site and users who rely on it for accurate answers [Overflow, 2022]. Writing and editing tools that rely on ChatGPT also face the risk of losing customers if they inadvertently introduce errors into the output.

Overall, the potential misuse of AIGC poses a threat to not only the users but also the whole creative industry. Therefore, it is crucial to use AIGC only in situations where the risk can be managed or corrected. To mitigate risks, it is also necessary to include governance mechanisms for AIGC models as soon as possible, such as establishing legal regulations. The most recent deal on comprehensive rules for trustworthy AI from EU [EU-, 2023] reflects the urgency to deal with concerns on the misuse of AIGC technologies.

8 Consent, credit, and compensation

Many AIGC models are trained on datasets without obtaining consent or providing credit or compensation to the original data contributors. For example, Simon Willison and

Andy Baio found that a large number of images in LAION were copied from DeviantArt and used to train Stable Diffusion [Willison and Baio, 2022]. This results in data contributors' works being learned by AI models and recreated by other users for profit, without their knowledge or permission. This practice damages the interests of the original data contributors. To avoid negative impacts, AIGC companies should obtain consent from data contributors and take proactive measures before training their models on any original or augmented works. Failure to do so could result in lawsuits against AIGC. Therefore, AIGC companies must ensure that data collection and model training are conducted in an ethical and responsible manner.

A potential solution to the issue of using creators' works for AI training is to notify them from the beginning and give them the option to benefit from subsequent creations based on their works generated by the model. Additionally, creators who give their consent for their data to be used can be rewarded based on how their creations contribute to AIGC each time the tool is queried. By incentivizing creators, companies can encourage creators to contribute more and accelerate the development of AIGC. For example, a more user-friendly version of Copilot could allow voluntary participation or compensate coders for contributing to the training corpus [Butterick, 2022].

9 Environment impact

The massive size of AIGC models, which can have billions even trillions of parameters, results in high environmental costs for both model training and operation. For example, GPT-3 has 175 billion parameters and requires significant computing resources to train. Narayanan *et al.* [Narayanan *et al.*, 2021] estimated that training GPT-3 with A100s would require 1,024 GPUs, 34 days, and cost 4.6 million dollars, with an expected energy consumption of 936 MWh [Charmaine Lai and Maver, 2022]. This raises important questions about how to reduce the energy consumption and carbon emission of AIGC models.

The upcoming GPT-4, with even more parameters than its predecessor, is expected to leave a more significant carbon emission. Failing to take appropriate steps to mitigate the substantial energy costs of AIGC could lead to irreparable damage to our planet. It is crucial to address these concerns and explore sustainable alternatives. Communities have started to explore more slim alternatives with decent performance as much larger ones.

10 Discussion

What about commercial usage: a vicious competition? Will AIGC replace humans and become a roadblock to human creativity? Many AIGC models are being utilized for art and graphic design commercially. For example, PromptBase [PromptBase, 2022] is an early marketplace for DALL·E, Midjourney, Stable Diffusion & GPT-3 prompts. Microsoft is using DALL-E 2 to power a generative art feature that will be available in Microsoft Edge. Microsoft and OpenAI are collaborating on ChatGPT-Powered Bing [Wiggers, 2022b]. Moreover, Microsoft had integrated OpenAI's

ChatGPT into Word, PowerPoint, Outlook, and other applications to allow users to automatically generate text using simple prompts [Holmes and McLaughlin, 2023]. While using the generated works for profit or commercial purposes is not recommended, there are no mandatory legal restrictions at this stage.

The use of AIGC has faced criticism from those who fear that it will replace human jobs. Insider has listed several jobs that could potentially be replaced by ChatGPT, including coders, data analysts, journalists, legal assistants, traders, accountants, etc [Mok and Zinkula, 2023]. Some artists worry that the wide use of image generation tools such as Stable Diffusion could eventually make human artists, photographers, models, cinematographers, and actors commercially uncompetitive [Heikkilä, 2022b]. For example, the images generated by Stable Diffusion can be sold on the market. This creates direct competition and poses a significant threat to creators, such as writers, artists, and programmers, who could suffer permanent damage to their businesses [Butterick, 2023]. Since Stable Diffusion can produce an unlimited number of infringing images, this threat is even more significant. However, David Holz, the founder of Midjourney, views artists as customers rather than competitors. Artists can use Midjourney to quickly prototype artistic concepts and show them to clients before starting work themselves [Holz and Claburn, 2022].

As AIGC models become more widespread, people may become too dependent on instant answers and less willing to think critically on their own, which could ultimately diminish or destroy human creativity and increase the risk of AI exerting control over humans. Overreliance on AIGC could create opportunities for malicious attackers to exploit user trust and access their private information.

Fairness of benefit distribution. It is important to recognize that AIGC models may have varying impacts on different groups of people depending on their environmental and individual abilities, which could further exacerbate global inequities [Weidinger *et al.*, 2021]. Addressing the issue of how to fairly distribute the benefits of AIGC models is an area that requires further exploration and attention.

Conflict among multiple goals. It is critical to ensure that the mitigation of one risk does not exacerbate another [Weidinger *et al.*, 2021]. For example, approaches to mitigate the use of toxic language in language models can introduce biases in model predictions against marginalized communities [Welbl *et al.*, 2021; Xu *et al.*, 2021]. Therefore, it is essential to explore effective mitigation strategies that can trade-off multiple goals.

11 Conclusion

Although AIGC is still in its infancy, it is rapidly expanding and will remain active for the foreseeable future. Current AIGC technologies only scratch the surface of what AI can create in the field of creativity. While AIGC offers many opportunities, it also carries significant risks. To acquire a thorough comprehension of these risks, we provide a synopsis of both current and potential threats in recent AIGC models, so that both the users and companies can be well aware of these

risks, and make the appropriate actions to mitigate them.

In order to promote responsible usage of AIGC tools and mitigate associated risks, we propose several steps that companies and users can take. It is important for companies to incorporate responsible AI practices throughout the whole life cycles during development of AIGC products. For example, proactive measures should be taken to mitigate potential risks in data sources, models, and pre/post-processing steps. Without proper safeguards, AIGC development may face significant challenges and regulatory hurdles. Note that this vision paper is not exhaustive, and it is essential for the wider community to contribute to the understanding and implementation of responsible AIGC. To facilitate this, it is necessary to build comprehensive benchmarks for measuring and evaluating the risks associated with different AIGC technologies.

References

[Agostinelli *et al.*, 2023] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text. *arXiv preprint arXiv: Arxiv-2301.11325*, 2023.

[Albert, 2023] Alex Albert. Jailbreak chatgpt. https://www.jailbreakchat.com/, 2023.

[Alston, 2023] Elena Alston. What are ai hallucinations and how do you prevent them? https://zapier.com/blog/ai-hallucinations/, 2023.

[Beaumont, 2022] Romain Beaumont. Clip retrieval system. https://rom1504.github.io/clip-retrieval/, 2022.

[Bickmore *et al.*, 2018] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. *Journal of medical Internet research*, 20(9):e11510, 2018.

[Birhane *et al.*, 2021] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Butterick, 2022] Matthew Butterick. Github copilot investigation. https://githubcopilotinvestigation.com/, 2022.

[Butterick, 2023] Matthew Butterick. Stable diffusion litigation. https://stablediffusionlitigation.com, 2023.

[Carlini *et al.*, 2021] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[Carlini *et al.*, 2022] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

[Chambon *et al.*, 2022] Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022.

[Charmaine Lai and Maver, 2022] Donna Dubinsky Charmaine Lai, Subutai Ahmad and Christy Maver. Ai is harming our planet: addressing ai's staggering energy cost. https://www.numenta.com/blog/2022/05/24/ai-is-harming-our-planet/, 2022.

[Chou *et al.*, 2022] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? *arXiv preprint arXiv:2212.05400*, 2022.

[Daryanani, 2023] Lavina Daryanani. How to jailbreak chatgpt. https://watcher.guru/news/how-to-jailbreak-chatgpt/, 2023.

[Dockhorn *et al.*, 2022] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *arXiv preprint arXiv:2210.09929*, 2022.

[Elliot and DeLisi, 2022] Bern Elliot and Meghan Rimol DeLisi. Why is chatgpt making waves in the ai market? https://www.gartner.com/en/newsroom/press-releases/2022-12-08-why-is-chatgpt-making-waves-in-the-ai-market, 2022.

[EU-, 2023] Eu ai act. https://www.nytimes.com/2023/12/08/technology/eu-ai-act-regulation.html, 2023.

[Feng *et al.*, 2021] Qianli Feng, Chenqi Guo, Fabian Benitez-Quiroz, and Aleix M Martinez. When do gans replicate? on the choice of dataset size. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6701–6710, 2021.

[Forbes, 2023] Forbes. ai-alliance. https://www.forbes.com/sites/stevemcdowell/2023/12/06/ibm--meta-launch-the-ai-alliance-for-safe-open-ai/?sh=5aeef6d9fb64, 2023.

[He *et al.*, 2022a] Xuanli He, Qiongkai Xu, Lingjuan Lyu, Fangzhao Wu, and Chenguang Wang. Protecting intellectual property of language generation apis with lexical watermark. AAAI, 2022.

[He *et al.*, 2022b] Xuanli He, Qiongkai Xu, Yi Zeng, Lingjuan Lyu, Fangzhao Wu, Jiwei Li, and Ruoxi Jia. Cater: Intellectual property protection on text generation apis via conditional watermarks. Advances in Neural Information Processing Systems, 2022.

[Heikkilä, 2022a] Melissa Heikkilä. Artists can now opt out of the next version of stable diffusion. https://www.technologyreview.com/2022/12/16/1065247/artists-can-now-opt-out-of-the-next-version-of-stable-diffusion/, 2022.

[Heikkilä, 2022b] Melissa Heikkilä. This artist is dominating ai-generated art. and he's not happy about it. https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/, 2022.

[Heikkilä, 2023] Melissa Heikkilä. Ai models spit out photos of real people and copyrighted images. https://www.technologyreview.com/2023/02/03/1067786/ai-models-spit-out-photos-of-real-people-and-copyrighted-images/, 2023.

[Ho et al., 2022] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[Holmes and McLaughlin, 2023] Aaron Holmes and Kevin McLaughlin. Ghost writer: Microsoft looks to add openai's chatbot technology to word, email. https://www.theinformation.com/articles/ghost-writer-microsoft-looks-to-add-openais-chatbot-technology-to-word-email, 2023.

[Holz and Claburn, 2022] David Holz and Thomas Claburn. David holz, founder of ai art generator midjourney, on the future of imaging. https://www.theregister.com/2022/08/01/david_holz_midjourney/, 2022.

[Jennings, 2022] Richi Jennings. Devs: Don't rely on github copilot — legal risk gets real. https://www.reversinglabs.com/blog/devs-dont-rely-on-github-copilot-legal-risk-is-real, 2022.

[Johnson, 2022] Khari Johnson. Dall-e 2 creates incredible images—and biased ones you don't see. https://www.wired.com/story/dall-e-2-ai-text-image-bias-social-media/, 2022.

[Kandpal et al., 2022] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. *arXiv preprint arXiv:2202.06539*, 2022.

[Kirchenbauer et al., 2023] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. *arXiv preprint arXiv:2301.10226*, 2023.

[Lazaridou et al., 2021] Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, C d M d'Autume, Sebastian Ruder, Dani Yogatama, et al. Pitfalls of static language modelling. *arXiv preprint arXiv:2102.01951*, 2021.

[Lopez, 2023] James Lopez. Microsoft, and amazon guard against chatgpt theft, ban employees from sharing sensitive data. https://www.techgoing.com/microsoft-and-amazon-guard-against-chatgpt-theft-ban-employees-from-sharing-sensitive-data/, 2023.

[Meehan et al., 2020] Casey Meehan, Kamalika Chaudhuri, and Sanjoy Dasgupta. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[Meta, 2023] Meta. Llama open source. https://ai.meta.com/llama/, 2023.

[Midjourney, 2022] Midjourney. Midjourney: Terms of service. https://midjourney.gitbook.io/docs/terms-of-service, 2022.

[Miller, 2022] Kirk Miller. Google admits its mind-blowing text-to-image ai is endlessly problematic. https://www.insidehook.com/daily_brief/tech/google-imagen-text-to-image, 2022.

[Mok and Zinkula, 2023] Aaron Mok and Jacob Zinkula. Chatgpt may be coming for our jobs. here are the 10 roles that ai is most likely to replace. https://www.businessinsider.com/chatgpt-jobs-at-risk-replacement-artificial-intelligence-ai-labor-trends-2023-02, 2023.

[Narayanan et al., 2021] Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, 2021.

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

[Office, 2023] U.S. Copyright Office. What is copyright? https://www.copyright.gov/what-is-copyright/, 2023.

[OpenAI, 2022] OpenAI. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2022.

[OpenAI, 2023] OpenAI. Gpt-4. https://openai.com/research/gpt-4, 2023.

[Overflow, 2022] Stack Overflow. Temporary policy: Chatgpt is banned. https://meta.stackoverflow.com/questions/421831/temporary-policy-chatgpt-is-banned, 2022.

[Pan et al., 2023] Minzhou Pan, Yi Zeng, Lingjuan Lyu, Xue Lin, and Ruoxi Jia. Asset: Robust backdoor data detection across a multiplicity of deep learning paradigms. *arXiv preprint arXiv:2302.11408*, 2023.

[Peng et al., 2023] Wenjun Peng, Jingwei Yi, Fangzhao Wu, Shangxi Wu, Bin Zhu, Lingjuan Lyu, Binxing Jiao, Tong Xu, Guangzhong Sun, and Xing Xie. Are you copying my model? protecting the copyright of large language models for eaas via backdoor watermark. *arXiv preprint arXiv:2305.10036*, 2023.

[Prabhu and Birhane, 2020] Vinay Uday Prabhu and Abeba Birhane. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*, 2020.

[PromptBase, 2022] PromptBase. Promptbase official website. https://promptbase.com, 2022.

[Radford et al., 2018] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[Ramesh *et al.*, 2021] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[Rombach *et al.*, 2022a] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[Rombach *et al.*, 2022b] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable diffusion github repository. https://github.com/CompVis/stable-diffusion, 2022.

[Rombach *et al.*, 2022c] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. Stable diffusion v1 model card. https://github.com/CompVis/stable-diffusion/blob/main/Stable_Diffusion_v1_Model_Card.md, 2022.

[Rosenblatt, 2022] Kalhan Rosenblatt. Chatgpt banned from new york city public schools' devices and networks. https://www.nbcnews.com/tech/tech-news/new-york-city-public-schools-ban-chatgpt-devices-networks-rcna64446, 2022.

[Runway, 2022] Runway. Text to video. https://runwayml.com/text-to-video/, 2022.

[Saharia *et al.*, 2022a] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[Saharia *et al.*, 2022b] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Schuhmann *et al.*, 2021] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[Shejwalkar *et al.*, 2023] Virat Shejwalkar, Lingjuan Lyu, and Amir Houmansadr. The perils of learning from unlabeled data: Backdoor attacks on semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4730–4740, 2023.

[Singer *et al.*, 2022] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[Somepalli *et al.*, 2022] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. *arXiv preprint arXiv:2212.03860*, 2022.

[Sun *et al.*, 2023a] Xiaofei Sun, Linfeng Dong, Xiaoya Li, Zhen Wan, Shuhe Wang, Tianwei Zhang, Jiwei Li, Fei Cheng, Lingjuan Lyu, Fei Wu, et al. Pushing the limits of chatgpt on nlp tasks. *arXiv preprint arXiv:2306.09719*, 2023.

[Sun *et al.*, 2023b] Xiaofei Sun, Xiaoya Li, Yuxian Meng, Xiang Ao, Lingjuan Lyu, Jiwei Li, and Tianwei Zhang. Defending against backdoor attacks in natural language generation. In *AAAI*, 2023.

[Sun *et al.*, 2023c] Xiaoxiao Sun, Nidham Gazagnadou, Vivek Sharma, Lingjuan Lyu, Hongdong Li, and Liang Zheng. Privacy assessment on reconstructed images: Are existing evaluation metrics faithful to human perception? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Villegas *et al.*, 2022] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

[Vincent, 2023] James Vincent. Getty images is suing the creators of ai art tool stable diffusion for scraping its content. https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit, 2023.

[Wang *et al.*, 2023a] Zhenting Wang, Chen Chen, Yuchen Liu, Lingjuan Lyu, Dimitris Metaxas, and Shiqing Ma. How to detect unauthorized data usages in text-to-image diffusion models. *arXiv preprint arXiv:2307.03108*, 2023.

[Wang *et al.*, 2023b] Zhenting Wang, Chen Chen, Yi Zeng, Lingjuan Lyu, and Shiqing Ma. Where did i come from? origin attribution of ai-generated images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Webster *et al.*, 2021] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. This person (probably) exists.

identity membership attacks against gan generated faces. *arXiv preprint arXiv:2107.06018*, 2021.

[Weidinger *et al.*, 2021] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[Welbl *et al.*, 2021] Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, 2021.

[Whang *et al.*, 2022] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022.

[Wiggers, 2022a] Kyle Wiggers. Deepfakes for all: Uncensored ai art model prompts ethics questions. https://techcrunch.com/2022/08/24/deepfakes-for-all-uncensored-ai-art-model-prompts-ethics-questions/, 2022.

[Wiggers, 2022b] Kyle Wiggers. Image-generating ai can copy and paste from training data, raising ip concerns. https://techcrunch.com/2022/12/13/image-generating-ai-can-copy-and-paste-from-training-data-raising-ip-concerns/, 2022.

[Willison and Baio, 2022] Simon Willison and Andy Baio. Exploring the training data behind stable diffusion. https://simonwillison.net/2022/Sep/5/laion-aesthetics-weeknotes/, 2022.

[Wortsman *et al.*, 2022] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

[Xie *et al.*, 2023] Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, pages 1–11, 2023.

[Xu *et al.*, 2021] Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. Detoxifying language models risks marginalizing minority voices. *arXiv preprint arXiv:2104.06390*, 2021.

[Yu *et al.*, 2022] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[Zeng *et al.*, 2022] Yi Zeng, Minzhou Pan, Himanshu Jahagirdar, Ming Jin, Lingjuan Lyu, and Ruoxi Jia. How to sift out a clean data subset in the presence of data poisoning? *arXiv preprint arXiv:2210.06516*, 2022.

[Zhang *et al.*, 2022] Zhiyuan Zhang, Lingjuan Lyu, Xingjun Ma, Chenguang Wang, and Xu Sun. Fine-mixing: Mitigating backdoors in fine-tuned language models. In *EMNLP*, 2022.

[Zhuang *et al.*, 2023] Weiming Zhuang, Chen Chen, and Lingjuan Lyu. When foundation model meets federated learning: Motivations, challenges, and future directions. *arXiv preprint arXiv:2306.15546*, 2023.