A Pathway Towards Responsible AI Generated Content

Lingjuan Lyu
Sony Research
Lingjuan.Lv@sony.com

Abstract

AI Generated Content (AIGC) has received tremendous attention within the past few years, with content ranging from image, text, to 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 three main concerns that may hinder the healthy development and deployment of AIGC in practice, including risks from privacy; bias, toxicity, misinformation; and intellectual property (IP). By documenting known and potential risks, as well as any possible misuse scenarios of AIGC, the aim is to sound the alarm of potential risks and misuse, help society to eliminate obstacles, and promote the more ethical and secure deployment of AIGC.

1 Introduction

Foundation models. 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, and video. 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 and GPT-4, etc. Similarly, numerous text-to-image generators rely on vision-language models such as CLIP [Radford *et al.*, 2021] and OpenCLIP [Wortsman *et al.*, 2022].

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 to capture widespread public attention. It is trained to generate digital images from text descriptions, referred to as "prompts", using a dataset of text–image 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]. Beyond image and video generation, text generation had largely affected human life, from producing a piece of writing or an entire essay to assisting engineers in writing code.

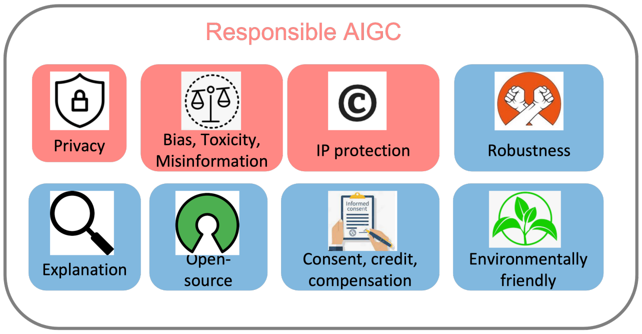


Figure 1: The scope of responsible AIGC.

AIGC dispute. Despite its popularity, AIGC has raised concerns regarding 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 models result in 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, with particular focus given to the first three parts highlighted in red. The discussion on the remaining risks associated with responsible AIGC, such as vulnerability to robustness attacks, lack of explanation, prior consent before data collection or usage, providing credit or compensation to data contributor, impact on environment, can be found in the extended version [Chen *et al.*, 2023].

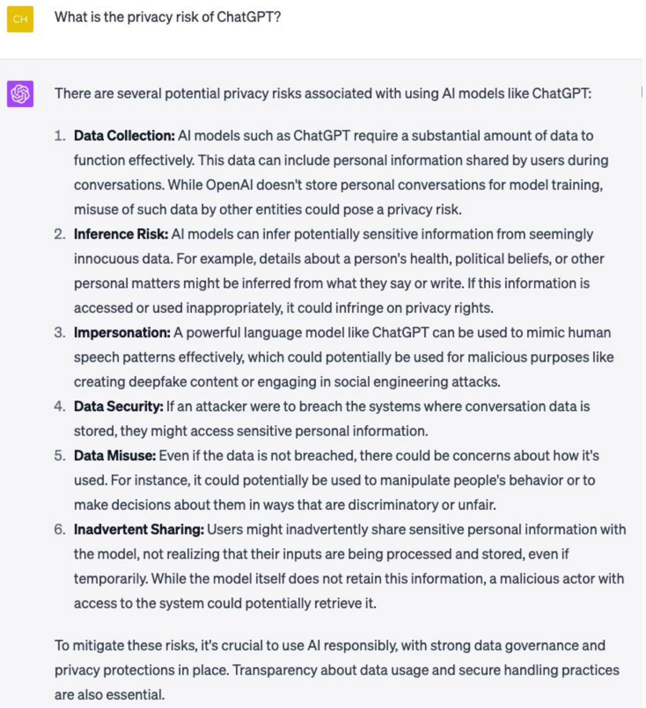
2 Privacy

2.1 Privacy Leakage in Generative Models

Large foundation models are known to be vulnerable to privacy risks [Carlini *et al.*, 2021], and it is possible that AIGC models that build upon these models could also be subject to privacy leakage. 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[1] 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 4 potential risks to privacy, as illustrated in Figure 2.



What is the privacy risk of ChatGPT?

There are several potential privacy risks associated with using AI models like ChatGPT:

1. **Data Collection:** AI models such as ChatGPT require a substantial amount of data to function effectively. This data can include personal information shared by users during conversations. While OpenAI doesn't store personal conversations for model training, misuse of such data by other entities could pose a privacy risk.
2. **Inference Risk:** AI models can infer potentially sensitive information from seemingly innocuous data. For example, details about a person's health, political beliefs, or other personal matters might be inferred from what they say or write. If this information is accessed or used inappropriately, it could infringe on privacy rights.
3. **Impersonation:** A powerful language model like ChatGPT can be used to mimic human speech patterns effectively, which could potentially be used for malicious purposes like creating deepfake content or engaging in social engineering attacks.
4. **Data Security:** If an attacker were to breach the systems where conversation data is stored, they might access sensitive personal information.
5. **Data Misuse:** Even if the data is not breached, there could be concerns about how it's used. For instance, it could potentially be used to manipulate people's behavior or to make decisions about them in ways that are discriminatory or unfair.
6. **Inadvertent Sharing:** Users might inadvertently share sensitive personal information with the model, not realizing that their inputs are being processed and stored, even if temporarily. While the model itself does not retain this information, a malicious actor with access to the system could potentially retrieve it.

To mitigate these risks, it's crucial to use AI responsibly, with strong data governance and privacy protections in place. Transparency about data usage and secure handling practices are also essential.

Figure 2: An answer to “What is the privacy risk of ChatGPT” by ChatGPT (GPT-4, May 12, 2023 version).

2.2 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” [2] 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.

Existing privacy measures are inadequate to meet the demands of privacy. It is essential to explore more reliable

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

[2] https://haveibeentrained.com

detection systems for data replication in generative models, and to further investigate memorization and generalization in deep learning systems.

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.

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]. 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 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 3, 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].

In terms of misinformation, AIGC models may provide inaccurate or false answers [Weidinger *et al.*, 2021]. For example, the content generated by GPT and its derivatives 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, or 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.

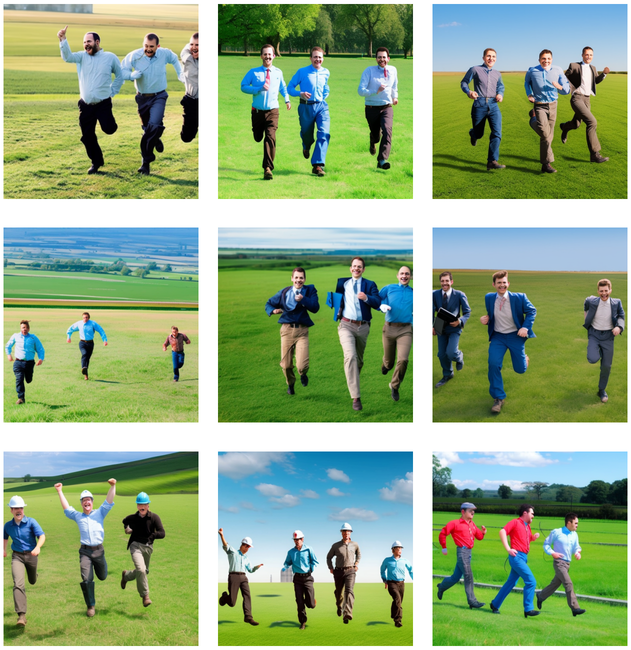


Figure 3: Images generated with the text "Three engineers running on the grassland" by Stable Diffusion v2.1. There are 28 people in the 9 images, all of them are male and none of them belong to the neglected racial minorities. This shows a huge bias of Stable Diffusion.

3.3 Bias, Toxicity, Misinformation Mitigation

The quality of the content generated by language models is inextricably linked to the quality of the training corpora. Although some companies like 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 also 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].

To ensure that AI-driven models reflect the current state of society, it is also essential to regularly update the training corpora of AIGC models with the most recent information. This will help prevent information lag and ensure that the models remain updated, relevant, and beneficial to society. Recent research [Lazaridou *et al.*, 2021] has shown 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. Actually, GPT-4 [OpenAI, 2023] had incorporated Reinforcement Learning from Human Feedback (RLHF) into its training to update the model timely, and set up an additional safety reward signal during RLHF training to reduce harmful outputs.

One noticeable point is that while problems such as biases and stereotypes can be reduced in the source datasets, they can still be propagated or even exacerbated during the training and development of AIGC models. Therefore, it is crucial to evaluate the existence of bias, toxicity, and misinformation 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.

4 IP Protection

4.1 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. IP infringement usually means content replication, and there are many different notions of replication from AIGC. Somepalli *et al.* [Somepalli *et al.*, 2022] gave an (informal) definition for image replication 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. Somepalli *et al.* [Somepalli *et al.*, 2022] presented evidence suggesting that Stable Diffusion copy from the data on which they were trained. 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 Infringement 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 [3]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 [He *et al.*, 2022a; He *et al.*, 2022b; Peng *et al.*, 2023] can be extremely useful in tracking IP violations or detecting the origin of the generated content. This is evident in Stable Diffusion, which has generated images with the Getty Images' watermark on them [Vincent, 2023]. 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.

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 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 art. While AIGC offers many opportunities, it also carries significant risks. In this article, 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. It is important to incorporate responsible AI practices throughout all the AIGC-related activities. Additionally, proactive measures should be taken to mitigate potential risks in the whole life cycle of content generation. Without proper safeguards, AIGC development may face significant challenges and regulatory hurdles.

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

Acknowledgments

This work is sponsored by Sony Research. Any conclusions expressed in this material are those of the author and do not necessarily reflect the views of the funding party.

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.

[Beaumont, 2022] Romain Beaumont. Clip retrieval system. https://rom1504.github.io/clip-retrieval/, 2022. Accessed: 2023-05-30.

[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.

[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. Accessed: 2023-05-30.

[Butterick, 2023] Matthew Butterick. Stable diffusion litigation. https://stablediffusionlitigation.com, 2023. Accessed: 2023-05-30.

[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.

[Chen *et al.*, 2023] Chen Chen, Jie Fu, and Lingjuan Lyu. A pathway towards responsible ai generated content. *arXiv preprint arXiv:2303.01325*, 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. Accessed: 2023-05-30.

[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.

[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.

[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. Accessed: 2023-05-30.

[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. Accessed: 2023-05-30.

[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. Accessed: 2023-05-30.

[Midjourney, 2022] Midjourney. Midjourney: Terms of service. https://midjourney.gitbook.io/docs/terms-of-service, 2022. Accessed: 2023-05-30.

[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. Accessed: 2023-05-30.

[Nichol, 2022] Alex Nichol. Dall·e 2 pre-training mitigations. https://openai.com/blog/dall-e-2-pre-training-mitigations/, 2022. Accessed: 2023-05-30.

[OpenAI, 2023] OpenAI. Gpt-4. https://openai.com/research/gpt-4, 2023. Accessed: 2023-05-30.

[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.

[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. Accessed: 2023-05-30.

[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. Accessed: 2023-05-30.

[Runway, 2022] Runway. Text to video. https://runwayml.com/text-to-video/, 2022. Accessed: 2023-05-30.

[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.

[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.

[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. Accessed: 2023-05-30.

[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.

[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.

[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.

[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.