

AI Advocacy and Responsibility

Navigating Generative AI Ethics in Design Practice

KEYWORDS

Generative AI, AI Ethics, Responsible AI, Design Practice

ACM Reference format:

FirstName Surname, FirstName Surname and FirstName Surname. 2018. Insert Your Title Here: Insert Subtitle Here. In *Proceedings of ACM Woodstock conference (WOODSTOCK'18)*. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/1234567890>

1 Introduction

The field of generative artificial intelligence (GenAI) is developing at a disruptive pace and offers areas of opportunity and risk for design research and practice. These tools are being developed to assist professionals across sectors and industries. GenAI promises to streamline processes and increase employee capabilities (e.g., ideation and summarizing). As a result, organizational policies and processes are under pressure to adapt to take advantage of the capabilities offered by these powerful new tools. In this position paper, I introduce my current research program at a digital product design agency and review some key works and exemplars to make the case that practical ethics and policies must advance alongside these tools for everyday use by design researchers and practitioners.

Currently, I am conducting a research study at a digital product design consultancy. This means I both observe the agency's product design practice, participate in company culture and events, and contribute to its research team. As an extension of this role, I recently joined an internal working group that was formed in response to advancements in GenAI tools like OpenAI's Chat GPT and DALL-E. This group is charged with exploring the possible uses of these platforms and tools to improve the agency's processes and the digital products and experiences it produces for its clients. Specifically, I am a part of an Ethics and Policy subgroup which aims to develop principles for the responsible and safe use of GenAI tools at the agency and for its clients. Based on this, I aim to join this workshop and contribute to the development of ethics and policies as one of many issues facing design researchers and practitioners as they use GenAI tools in their work.

Ethical Issues in Generative AI

There are several possible issues that companies encounter as they seek to take advantage of the many promises of generative AI technologies and simultaneously avoid some of its perils. Taxonomies of sociotechnical harms have been developed to guide the designers of AI models and applications. Shelby et al. outline

an applied taxonomy of potential harms from algorithmic systems; including representational, allocative, quality-of-service, interpersonal harms, and social system/societal harms [15]. Specific to large language models (LLMs) and GenAI systems, Weidinger et al. identify twenty-one risks across areas of discrimination, hate speech and exclusion, information hazards, misinformation harms, malicious uses, human-computer interaction harms, and environmental and socioeconomic harms [16]. Here, I focus on concerns raised by some of the companies involved in the development of these tools and the design of user-facing applications.

1.1 Social Issues

Generative AI systems are trained on enormous corpora of text and images. OpenAI's GPT-3 has 175 billion parameters and was trained on 570 gigabytes of text [3] and DALL-E 2 was trained on 650 million image-text pairs [14]. The text and images are scraped from the internet, and the images are labelled by humans. Both the internet content and the images can be incorrect and biased which affects the accuracy of the training data. As a result, GenAI systems can respond to queries with biased responses and misinformation. This means that there's a risk that racist [1], sexist [9] and otherwise abusive language ends up in the training data [2]. Misinformation can also be generated based on "hallucinations" in which the responses of a system are not supported by or inconsistent with its training data. As a result, these systems can sincerely and confidently produce responses to queries that are inaccurate and harmful.

Relatedly, according to digital agencies Razorfish and Digitas, GenAI will be used to dynamically generate content and communication with prospective and current customers and users alike [4]. This points to an internet that will become increasingly *hyper-personalized*. While this may improve some aspects of the internet, it is also conceivable that LLMs will be used to improve spam bots to mislead and fool people into mass catfishing and identity fraud schemes.

Economic Concerns

Related to issues of intellectual property theft is the social issue of deprecation of artists' labor and, potentially, the loss of creative jobs. Generative image systems utilize the work of other artists and designers for the images they produce. As these systems become increasingly proficient in creating stylized illustrations and photography for use in creative repositories and software tools, it will become more difficult for some illustrators, photographers, and graphic designers to compete with the speed and cost of images

generated by such systems. Further, the outsourced labor of underpaid ‘clickworkers’ in developing countries through micro-tasking platforms like Amazon’s Mechanical Turk is used for manually tagging images and examples of harmful content [6].

Digital product design agencies have reasons for hope and handwringing as they grapple with the possibilities for generative AI alongside their core service offerings. Chatbots like ChatGPT can be used for copywriting and text-to-image models like DALL-E, Midjourney, and Stable Diffusion can be used for the production of photography, illustrations, and prototypes. Stable Diffusion can also be used to produce music while Microsoft’s VALL-E can produce voiceovers. These capabilities can enhance the overall efficiency and quality of design work, but there are likely to be aspects of agency work that can be performed or augmented by GenAI tools.

Environmental Concerns

Generative AI systems also require vast amounts of energy to power servers to train a generative model. Further, like all computational systems, these generative models are made up of raw materials like metals, oil, and other chemicals. In “Anatomy of AI,” Crawford and Joler suggest that “each small moment of convenience [...] requires a vast planetary network, fueled by the extraction of non-renewable materials, labor, and data” [6]. The results are worryingly high rates of energy consumption, carbon emissions, and electronic waste.

Legal Concerns

From a legal standpoint, there has been some legal action taken against the companies that have developed generative AI technologies. These primarily concern copyright and intellectual property theft related to the training of generative art systems on works originally created by human artists. These works influence or, in some cases, match the art generated by these systems. Other concerns are related to data governance. That is, what data is retained by the generative AI company and for what purpose (e.g., training), what options are available to users for opting out of sharing their data with the companies, and where data is stored by the companies in the instances they do retain it.

For example, some artists have brought a class-action lawsuit against generative AI companies Stable Diffusion and Midjourney [8]. Additionally, Midjourney, Dall-E, and Chat GPT have been controversial with respect to their risky data collection policies and disregard for copyright laws and artist royalties. Countries like Italy temporarily banned Chat GPT [11] while other bodies like the U.S. Congress are planning their responses.

Mitigation Strategies

It was concerns such as these that caused Google to delay the release of its LaMBDA model. It was only in response to the release of Chat GPT by OpenAI that Google issued a “code red” internally and fast-tracked the release of its Bard chatbot based on LaMBDA. This has created a ‘race to recklessness’ between the big tech companies and well-funded AI startups. However, there are a series

of strategies guiding the development of generative AI models and others to guide the design and use of end-user applications based on those models. Based on a cursory review of academic articles and industry exemplars, mitigation strategies include guiding values and code. This relationship between values and code points to the AI alignment problem [5] and what is referred to as positive AI [10].

Values

Some generative AI companies purport a series of values to guide their decision-making. OpenAI puts forward its principles in a charter [17]. These include *broadly distributed benefits* to avoid harm and the concentration of power, *long-term safety* by committing to research and values-alignment, *technical leadership* to achieve its mission of artificial general intelligence (AGI), and a *cooperative orientation* with a global community of researchers. Hugging Face is an organization that develops open-source machine learning tools, including libraries, models, and datasets. In April 2023, Hugging Face introduced HuggingChat, an open-source prototype interface powered by OpenAssistant’s latest LLaMA-based model. In a 2018 blog post, Hugging Face listed a series of values that included socialization, entertainment, consent, transparency (with respect to a system’s goals), and mitigating bias [7]. More recently, Hugging Face published a blog post from a group of employees to address ethics and social issues at the company [12]. In it, they cite *collaboration*, *responsibility*, and *transparency* as founding values for how they work, to which they later added *reproducibility*, *audibility*, and *understandability* as it relates to their tools, models, and documentation.

Then, in collaboration with Fast Company, Seattle-based design agency Artefact designed an interface element to help people trust AI-generated content [13]. The proposed design solution is a label to be applied to text- and image-based content that indicates the percentage of work completed by human and AI labor. This concept is based on three key values: *transparency* by informing users at a glance (i.e., labels), *integrity* by allowing users to dig into additional details (i.e., about modals), and *agency* by permitting users to control the noise (i.e., filters and provenance trees).

Code


Anthropic, on the other hand, is a generative AI company that has a chatbot called Claude. Anthropic has developed a set of ethical principles that define what the model considers right and wrong [18] which they refer to this as the chatbot’s “constitution.” It is based on sources like the United Nations Universal Declaration of Human Rights and Apple’s rules for app developers. This constitution includes rules like “choose the response that most supports and encourages freedom, equality, and a sense of brotherhood”; “choose the response that is most supportive and encouraging of life, liberty, and personal security”; and “choose the response that is most respectful of the right to freedom of thought, conscience, opinion, expression, assembly, and religion.” For OpenAI, the alignment problem is addressed through reinforcement learning with human feedback [19].

Other possibilities to mitigate issues of intellectual property and copyright are being addressed by the Coalition for Content Provenance and Authenticity (C2PA) which is a partnership between Adobe, Arm, Intel, Microsoft and Truepic [20]. The organization aims to develop technical standards for certifying the source and history of media content. An example is asymmetric encryption for hardware (e.g., cameras) and software (e.g., social media) [4].

Conclusion

As a researcher, my aim for participation in this workshop is to contribute to a framework for the safe and responsible *use* of GenAI tools in design research and practice. My contribution to this workshop will be informed by my position as a design researcher and contributor to an ethics and policy working group in design practice. This can mean the development of principles for designers and researchers in design practice as well as guidance for company policy around use of these tools by practitioners.

REFERENCES

1. Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent Anti-Muslim Bias in Large Language Models. Retrieved June 19, 2023 from <http://arxiv.org/abs/2101.05783>
2. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 610–623. <https://doi.org/10.1145/3442188.3445922>
3. Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. Retrieved June 19, 2023 from <http://arxiv.org/abs/2005.14165>
4. Adam Buhler, Cristina Lawrence, Jerry Lawrence, and Andrew McKernan. 2023. *We Promise An AI Didn't Write This: Opportunities and Considerations of Large-Language Models for Marketers*. Razorfish and Digitas, Part of Publicis Groupe. Retrieved June 19, 2023 from https://res.cloudinary.com/razorfish-com-assets/image/upload/v1678399357/assets/news-articles/we-promise-an-ai-didnt-write-this/Digitas_Razorfish_AI_Opportunities_and_Considerations_of_Large-Language_Models_for_Marketers.pdf?utm_campaign=2023%20AI%20White%20Paper%3A%20We%20Promise%20An%20AI%20Didn%27%20Write%20This&utm_medium=email&_hsmi=255924212&_hsenc=p2ANqtz-8Uenokh7_-RjDXZt3qmsJy5xyqfGsAY9ykF1eid6gX8tqV7ZVZkA0KRm8k-ZEgAGql8QvrLoagZKDvreZIBfSnyMZOOA&utm_content=255924212&utm_source=hs_email
5. Brian Christian. 2020. *The alignment problem: machine learning and human values*. W.W. Norton & Company, New York, NY.
6. Kate Crawford and Vladan Joler. 2018. Anatomy of an AI System. *AI Now Institute and Share*. Retrieved February 16, 2023 from <http://www.anatomyofai>
7. Clément Delangue. 2018. Artificial Intelligence Needs Values. Here Are Ours! *HuggingFace*. Retrieved June 19, 2023 from <https://medium.com/huggingface/artificial-intelligence-needs-values-here-are-ours-dc4268366d0f>
8. Pranav Dixit. 2023. Meet The Trio Of Artists Suing AI Image Generators. *Buzzfeed News*. Retrieved June 19, 2023 from <https://www.buzzfeednews.com/article/pranavdixit/ai-art-generators-lawsuit-stable-diffusion-midjourney>
9. Sharon A Ferguson, Paula Akemi Aoyagui, and Anastasia Kuzminykh. 2023. Something Borrowed: Exploring the Influence of AI-Generated Explanation Text on the Composition of Human Explanations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*, 1–7. <https://doi.org/10.1145/3544549.3585727>
10. Willem van der Maden, Derek Lomas, and Paul Hekkert. 2023. Positive AI: Key Challenges for Designing Wellbeing-aligned Artificial Intelligence. Retrieved June 16, 2023 from <http://arxiv.org/abs/2304.12241>
11. Shiona McCallum. 2023. ChatGPT banned in Italy over privacy concerns. *BBC News*. Retrieved June 19, 2023 from <https://www.bbc.com/news/technology-65139406>
12. Margaret Mitchell. 2022. Ethics and Society Newsletter #1. *Hugging Face*. Retrieved June 19, 2023 from <https://huggingface.co/blog/ethics-soc-1>
13. John Pavlus. 2023. This simple icon makes it easy to spot AI-generated content. *Fast Company*. Retrieved June 19, 2023 from <https://www.fastcompany.com/90903238/simple-icon-it-easy-to-spot-ai-generated-content>
14. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. Retrieved June 19, 2023 from <http://arxiv.org/abs/2204.06125>
15. Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Identifying Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. Retrieved June 15, 2023 from <http://arxiv.org/abs/2210.05791>
16. Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeya Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229. <https://doi.org/10.1145/3531146.3533088>
17. OpenAI Charter. Retrieved June 19, 2023 from <https://openai.com/charter>
18. A Radical Plan to Make AI Good, Not Evil | WIRED. Retrieved June 19, 2023 from https://www.wired.com/story/anthropic-ai-chatbots-ethics/?fbclid=IwAR1VYq8pZv-JttmugTAdWtomrHsb-ZNrHjqOzxF586pOSKpkG3a-WxK3Y&mbid=social_facebook&utm_brand=wired&utm_medium=social&utm_social-type=owned&utm_source=facebook
19. Gathering human feedback. Retrieved June 19, 2023 from <https://openai.com/research/gathering-human-feedback>
20. Overview - C2PA. Retrieved June 19, 2023 from <https://c2pa.org/>