Editorial: Copyright protection, artistic imagery, and the adoption of responsible artificial intelligence principles

1. Introduction

With the advent of ChatGPT (Generative Pre-Trained Transformer) – generative artificial intelligence (AI) – in November 2022, there has been a subsequent proliferation of proprietary generative AI technologies models used by consumers to create new content, which includes audio, code, images, text, simulations and videos (McKinsey and Company, 2023). These generative AI models are being continuously trained on database sets that learn to generate more objects that resemble the data or images that they were trained on (Zewe, 2023), including artists' work that is copyright-protected[1], intellectual property (IP). Several major AI companies, including OpenAI, Meta, Google, as well as smaller, entrepreneurial companies, such as Midjourney, Stability, Deviant Art and Runway AI, are contending with class-action lawsuits brought by artists claiming that their copyrighted material, i.e. artistic images, are being used by these companies in generative AI training exercises without their consent or receiving financial remuneration, i.e. licensing fees, for the right to use artistic imagery (Brittain, 2023; Heikkila, 2024).

In response to these alleged copyright infringements perpetrated by these AI companies, computer scientists have created digital tools useful to artists as a deterrent to AI companies who have been violating their copyrights (and intellectual property (IP)) to counter-act companies training their AI models without artist consent. Specifically, there are two digital tools available for artists to deter generative AI infringing on copyrighted images: Glaze and Nightshade.

Glaze is a system designed to protect human artists by disrupting style mimicry. At a high level, Glaze works by understanding the AI models that are training on human art, and using machine learning algorithms, computing a set of minimal changes to artworks, such that it appears unchanged to human eyes, but appears to AI models like a dramatically different art style (The Glaze Project, 2024a).

Nightshade works similarly as Glaze, but instead of a defense against style mimicry, it is designed as an offense tool to distort feature representations inside generative AI image models. Like Glaze, Nightshade is computed as a multi-objective optimization that minimizes visible changes to the original image. While human eyes see a shaded image that is largely unchanged from the original, the AI model sees a dramatically different composition in the image (The Glaze Project, 2024b).

There are three legal and ethical questions that arise from the above generative AI technologies scenario. First, for AI companies that publicly state that they abide by a set of voluntary "AI Principles", i.e. an organizational policy approach to "Responsible AI", does the indiscriminate infringement of artists' IP by such companies meet the ethical requirements of Responsible AI, much less than the legal protections recognized by copyright ownership? Second, is the use of Glaze by artists a legally and ethically acceptable method of protecting an artists' copyrighted images? Third, is the use of Nightshade ethically – or legally – acceptable to implement by artists to assist in protecting an artists' copyrighted images? To address the first question, the concept of Responsible AI

Journal of Ethics in
Entrepreneurship and Technology
Vol. 4 No. 1, 2024
pp. 2-6
© Emerald Publishing Limited
e-ISSN: 2633-7444
p-ISSN: 2633-7436
DOI 10.1108/JEET-06-2024-025

is next explained and its' ethical implications for IP infringement of artistic work further explored.

2. Responsible artificial intelligence and copyright protection

A working definition of "Responsible AI" has been developed by the Microsoft Corporation: "an approach to developing, assessing, and deploying AI systems in a safe, trustworthy, and ethical way" (Azure Machine Learning, 2024). Specifically, Responsible AI consists of policy principles which, in total, create a "Responsible AI Standard". In their research paper, Hemphill and Longstreet (2023), evaluated examples of major firm (Alphabet, Meta and Microsoft), industry (IT Industry Council) and cross-industry organization (U.S. Chamber of Commerce) responsible AI private governance standards. In their analysis, Hemphill and Longstreet (2023) found that there are seven common private governance categories of AI principles found among these five organizations: safety and reliability, responsible R&D, fairness and bias, privacy and security, transparency and accessibility, accountability and societal improvement.

Noteworthy among the seven common private governance AI principles that make up a Responsible AI Standard is the category of "responsible R&D". As it pertains to an artist's copyright protection of creative work, training generative AI models entails a company "development" activity which can potentially legally infringe upon an artist's IP. Ethically, companies having a Responsible AI Standard, e.g. Meta and Microsoft, not only should be legally abiding by copyright law and respecting IP rights of an artist's work, but also be ethically responsible for following its own Responsible AI Standard as to its pre-training of generative AI models on such data and artistic images. For these companies, they should be held "accountable" for acquiring artistic consent and appropriately compensating artists for the pre-training use of their copyrighted artistic images.

How have AI companies responded to artists' complaints and lawsuits? Stability.AI (in December 2022) and OpenAI (in September 2023) have offered to let artists "opt-out" of having their copyright-protected artistic images used in training future versions of their AI generative text-to-image models (Heikkila, 2022; Wong, 2023). Artists, however, argue that this "opt-out" option is insufficient, as these policies "require artists to jump through hoops and still leave tech companies with all the power" (Heikkila, 2024: 8). Moreover, the "opt-out" option on artistic images requires "that every single artist in the world is automatically opted in and our choice is taken away", says Karla Ortiz, an artist and a board member of the Concept Art Association (Heikkila, 2022). The existing AI companies "opt-out" response does not appear to meet what artists argue should be found in a "fair" and ethical Responsible AI Standard.

3. Glaze and nightshade: legal and ethical responses?

What is interesting - and controversial - about the two digital tools (Glaze and Nightshade) is that Glaze is considered "defensive" in nature, while Nightshade "offensive" in nature. By "defensive", it is meant that an AI model "might see the glazed version of a charcoal portrait as a modern abstract style, versus a human viewing a *glazed* charcoal portrait with an actual unchanged realism style" (The Glaze Project, 2024a). Therefore, this "glaze" effect disrupts the AI generative model. Thus, when a consumer "prompts an AI generative model to mimic the original charcoal artist image, that humans will get an output much different than expected" (The Glaze Project, 2024a). It is defensive in nature because it protects the artist's IP artistic image from copyright infringement, without causing any harmful effects to the company's AI generative model or database - only to the artist's Glaze consented IP. Yet, there are potentially ethical concerns with Glaze - especially for downstream, unknowing consumers of the AI generative model.

In contrast, Nightshade is “offensive” in nature. For example:

> […] Nightshade transforms images into “poison” samples, so that models training on them without consent will see their models learn unpredictable behaviors that deviate from expected norms, e.g. a prompt that asks for an image of a cow flying in space might instead get an image of a handbag floating in space (The Glaze Project, 2024b).

Nightshade is therefore considered a deterrent to AI companies who “disregard copyrights, opt-out lists, and do-not-scrape/robots.txt directives” (The Glaze Project, 2024b). Nightshade “associates a small incremental price on each piece of data scraped and trained without authorization” (The Glaze Project, 2024b). What is Nightshade’s ultimate goal? That goal is “to increase the cost of training on unlicensed data, such that licensing images from their creators becomes a viable alternative” (The Glaze Project, 2024b). How is this goal to be accomplished? By allegedly protecting the IP rights of artists whose artistic images are being infringed through what is referred to as “prompt-specific poisoning” (Shan *et al.*, 2024), thus altering a company’s AI generative models so they “learn unpredictable behaviors” and the copyright infringing companies incur “a small incremental price.” Nightshade may have long-term, negative financial consequences for these AI companies; however, this cost is intended to “incentivize” companies to request consent to access and license the artists’ images.

A “prompt-specific poisoning” tool such as Nightshade initially appears to be a legal activity. However, depending on the jurisdiction, the “intent” of this digital tool may be legally questionable depending on whether the so-called “poisoning” has an intent to cause harm or deceive users. This intent, however, has yet to be tested in the courts criminally or civilly concerning the application of Nightshade, for example, as this digital tool has only recently been publicly made available to artists. Nevertheless, there could be impacts of “prompt-specific poisoning” activity that could conceivably violate the terms of service on a digital platform that hosts an AI generative model. While this “prompt-specific poisoning” activity may not be *per se* illegal, the AI generative model could be banned or there could be other negative consequences imposed on an AI company by a platform service provider under its terms of service contract language.

An ethical concern surrounding a “prompt-specific poisoning tool” such as Nightshade, or by a disrupting style tool such as Glaze, includes whether the “poisoning” or “disruption” is accomplished without transparency to the AI generative model company, and by extension, to the consumer utilizing the model who has not consented to using this model without knowledge of its limitations. This “poisoning” or “disruption” could be considered deceptive to the AI company and consumer. Specifically, however, the consumer is totally unaware that the AI generative model has these inherent modifications for certain AI image generated content. Whether using a deontological, teleological or virtue ethics theory of morality (MacKinnon and Fiala, 2018), there are ethical issues arising with a “prompt-specific poisoning tool” such as Nightshade, or from the “disruption” effects of Glaze, to the unwitting consumer.

In the case of a deontological moral approach, the *intention* of the artist using Nightshade or Glaze may be focused on negatively impacting the copyright infringing AI company, but its downstream effects are potentially negatively affecting unknowing consumers. In the case of a teleological moral approach, the artist’s negative *consequences* are to be focused on the copyright infringing AI company; however, the unknowing consumer of the AI generative model will be the ultimate victim of Nightshade’s and Glaze’s deception. Finally, a virtue ethics moral approach presumes that an artist will have a predisposition to behave in a morally righteous way, but in the case of Nightshade and Glaze, the character - or moral foundation - of an artist should focus on *not deceiving* the consumer with these copyright-protected, artistic images.

4. Discussion

AI companies – whether established or entrepreneurial – offering AI generative models in the marketplace need to abide by the Responsible AI Standard, specifically the "Responsible R&D" category, and meet the legal requirements of recognizing copyright protection for artists' artistic images, i.e. consent and licensing. As mentioned earlier, Stability.AI and OpenAI are allowing artists to "strictly opt-out" of having their copyright-protected artistic images used in training future versions of their AI generative text-to-image models (Rippy, 2021). This approach has been criticized in the U.S. artistic community, as it places the onus – or burden – on the artist to "protect' their copyright on their imagery by actively consenting to this notification to prevent copyright infringement. This is prima facie evidence that AI companies are not recognizing the IP protection of artistic imagery. To re-establish compliance with the "Responsible R&D" category, these AI companies need to establish a "strictly opt-in" approach, by which they establish a default rule whereby each company must receive active consent to allow – after agreeing to a negotiated licensing fee – the company to utilize the artists' copyright-protected imagery to be used for its AI generative model training purposes (Rippy, 2021). Moreover, AI technologies can also be used to accurately maintain and improve artist consent tracking for the purposes of AI generative model training (Porter, 2023). This "strictly opt-in" approach is particularly important for entrepreneurial enterprises, who often lack the deep financial "pockets" to engage in lengthy, expensive copyright infringement lawsuits concerning proprietary artistic images.

Ethically, there are potential issues concerning the potential negative impacts of artists using Glaze, and particularly, Nightshade, as it concerns consumers and their expectations of AI generative models under the existing operational environment for AI companies. The most effective solution to ethically treating consumers fairly and honestly, however, is to institute an effective "opt-in" approach to recognizing artists IP rights. This would change - to a better balance - the power asymmetry that presently exists between artists ("weak") and AI companies ("strong"), thus allowing for consumers to benefit from accurate text-to-image responses (for those artist's artistic imagery consensually included) when they query AI generative models.

Thomas A. Hemphill
Department of Management and Marketing,
University of Michigan-Flint, Flint, Michigan, USA

Note

1. According to the U.S. Copyright Office (2024), "[C]opyright is a type of intellectual property that protects original works of authorship as soon as an author fixes the work in a tangible form of expression. Under the current law, works created on or after January 1, 1978, have a copyright term of life of the author plus seventy years after the author's death."

References

Azure Machine Learning (2024), "What is responsible AI? AI skills challenge", Microsoft Corporation, January 31, available at: https://learn.microsoft.com/en-us/azure/machine-learning/concept-ponsible-ai?view=azureml-api-2

Brittain, B. (2023), "Artists take new shot at stability, midjourney in updated copyright lawsuit.", *Reuters*, available at: www.reuters.com/legal/litigation/artists-take-new-shot-stability-midjourney-updated-copyright-lawsuit-2023-11-30/

Heikkila, M. (2024), “This new Data-Poisoning tool lets artists fight back against generative AI.”, *MIT Technology Review*, Vol. 127 No. 1, pp. 7-8.

Heikkila, M. (2022), “Artists can now opt-out of the next version of stable diffusion”, MIT Technology Review, December 16, available at: www.technologyreview.com/2022/12/16/1065247/artists-can-now-opt-out-of-the-next-version-of-stable-diffusion/.

Hemphill, T.A. and Longstreet, P. (2023), “How private governance mitigates AI risk”, Policy Paper, The Center for Growth and Opportunity, Jon M. Huntsman School, Utah State University, September 26.

MacKinnon, B. and Fiala, A. (2018), *Ethics: Theory and Contemporary Issues*, Cengage Learning, Boston, MA.

McKinsey and Company (2023), “What is generative AI?”, McKinsey Explainers, January 29, available at: www.mckinsey.com/~/media/mckinsey/featured%20insights/mckinsey%20explainers/what%20is%20generative%20ai/what%20is%20generative%20ai.pdf

Porter, A. (2023), “Opt-In vs Opt-Out? Understanding consent management”, BigId, October 17, available at: https://bigid.com/blog/opt-in-vs-opt-out-consent/

Rippy, S. (2021), “Opt-In vs. Opt-out approaches to personal information processing”, International Association of Privacy Professionals, May 10, available at: https://iapp.org/news/a/opt-in-vs-opt-out-approaches-to-personal-information-processing/

Shan, S., Ding, W., Passananti, J., Wu, S., Zhang, H. and Zhao, B.Y. (2024), “Prompt-Specific poisoning attacks on text-to-image generative models”, *Cornell University, February*, Vol. 16 https://arxiv.org/abs/2310.13828.

The Glaze Project (2024a), “What is glaze?”, University of Chicago, available at: https://glaze.cs.uchicago.edu/what-is-glaze.html

The Glaze Project (2024b), “What is nightshade?”, University of Chicago, available at: https://nightshade.csuchicago.edu/whatis.html

U.S. Copyright Office (2024), “What is copyright?”, available at: www.copyright.gov/what-is-copyright/

Wong, M. (2023), “Artists are losing the war against AI”, *The Atlantic*, October 2, available at: www.theatlantic.com/technology/archive/2023/10/openai-dall-e-3-artists-work/675519/

Zewe, A. (2023), “Explained: Generative AI”, MIT News, Massachusetts Institute of Technology, November 9, available at: https://news.mit.edu/2023/explained-generative-ai-1109

About the author

Thomas A. Hemphill, David M. French Distinguished professor of strategy, innovation and public policy, School of Management, University of Michigan-Flint, received his PhD in Business Administration with a primary field in Strategic Management and Public Policy and secondary field in Technology and Innovation Policy from The George Washington University. His technology and innovation management publications can be found in the *Bulletin of Science*, *Technology and Society*; *Innovation: Management, Policy and Practice*; *International Journal of Innovation Management*; *International Journal of Innovation and Technology Management*, *Journal of Responsible Innovation*, *Knowledge, Technology and Policy*; *Research-Technology Management*; *Science and Public Policy*; *Technology Analysis and Strategic Management*; *Technology In Society: An International Journal, and Technological Sustainability.*