Poisoning the Palette: Evaluating Glaze, Nightshade, and Watermarking as Defenses Against Unauthorized AI Training

Negotiating Creativity: Artists, Institutions, and Policymakers in the Age of Generative AI

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Your Major

By
Aaditya Ghosalkar
November 8, 2024

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Kent Wayland, Department of Engineering and Society

Introduction:

As AI continues to evolve, it dramatically changes the digital landscape, especially within the creative sector. Generative AI allows users to compile images and artworks to create something similar to that art style. Think of taking all of Van Gogh's paintings and creating similar art based on his style just by feeding program information on how he decided to make specific pieces and then using that to create one's images without any experience or learning the process of drawing. This functionality is readily accessible through applications like MidJourney and DALLE. Generative AI models, which learn from vast datasets and user feedback, can now create content in diverse formats, including text, images, audio, or combinations. As a result, roles centered on content production—such as writing, illustration, coding, and other knowledge-intensive tasks—appear poised to be significantly influenced by these emerging technologies. (Reference below). Due to this, many ethical concerns exist regarding applications using images for their training model.

The internet is free, and using artists' styles without the right to is all too familiar. Events like tracing and copying others' work and reselling it for profit are examples of how people benefit from other people's work without just compensation. The issue we have right now is the taking of artists' ability to even profit off of their art by making images for free labor and effort. Artists have their works stolen to make works that look like their own art, which is an entirely different beast of an issue.

Although multiple industries face this issue, my STS project will focus on image generation and what guidelines are in place to protect artists from AI. Still, there are industries we can take similar ideas from, including voice acting and script writing, which have recently faced similar

problems with AI replicating performances to suit a producer's needs. These industries can give us insight into how they tackle this issue and how we can further set guidelines for art.
My technical portion of this project will explore how AI replicates images and how it works on a high level. I will then explore the effectiveness of specific applications that abuse models training off images to combat this growing concern of art theft.

Technical Problem Statement:

As AI becomes more of a prevailing topic, this section aims to highlight how training a model leads to creating images or replicating voices or writing and how specific applications and people exploit these processes to prevent data collection. My project will focus on said techniques and their effectiveness; for images specifically, Glazing and Nightshade are among the most common methods. Glaze makes subtle, almost imperceptible alterations to an artwork so that AI models perceive it as a drastically different style. At the same time, it looks unchanged to human eyes, making it difficult for models to accurately recognize features such as eyes, faces, or other objects. Nightshade, on the other hand, transforms images into “poison” samples, causing models trained on these images without consent to learn unpredictable and deviant behaviors. Both techniques have limitations, and this project’s technical portion will involve testing these methods on images from consenting artists, interviewing them to determine whether they can perceive the changes, and comparing the altered images against a control set. Alongside Glaze and Nightshade, other protective measures like visible watermarks will be explored for their potential effectiveness in preventing models from extracting helpful training data. The experiments will involve a test set of images modified using these techniques and a control set of

unaltered photos, and an image-generation tool—currently, Dall-E 2—is planned for evaluating how well models pick up on Glaze, Nightshade, and watermarking. Subsequently, these outputs will be tested among various groups to see if the resulting images can be identified as AI-generated.

STS Research Problem:

It has been a bit of a debate recently; with the rise of AI in the creative fields, what exactly is creativity? What separates the work of a human versus the work of an AI? "Creative tasks generally require some degree of original thinking, extensive experience, and an understanding of the audience, while production tasks are, in general, more repetitive or predictable, making them more amenable to being performed by machines." (Anantrasirichai & Bull, 2022) Machines, by definition, do not have that level of creative thinking and personality to put their thought into work. At its core, AI is a technology that predicts where to set colors and strokes on a canvas based on what data it is given to replicate. This distinction is essential for this argument. Right now, AI has become a heated topic in the entertainment industry, with some of the major players in the scene being voice actors and screenwriters for big production studios, fearing their previous works being used as data for machines to create new pieces. This encroachment of AI in this industry exists in the art sector, where publically available services allow individuals to generate art based on massive images. More often than not, the images these models take from are without the creator's consent, and these tools allow consumers a low-cost, low-effort, and time-efficient alternative to hiring an artist for whatever artwork they want. Now, artists have to compete with AI models that have both their works and those of other artists and can generate images in a fraction of the time it takes them to make one.(Jiang et al., 2023)

How are artists, industry organizations, and policymakers navigating and negotiating the new conditions introduced by generative AI tools?

Background

As investments in AI continue to soar—global private investment in AI reached approximately $93.5 billion in 2021 (Lynch, 2022) [7], more than double the previous year's total—so does the technology's presence in creative domains. This surge has enabled new capabilities, including generating imagery and audio that closely mimic existing styles, voices, and aesthetics. Such tools can produce content far faster than human artists, often drawing from massive datasets of copyrighted work without the creators' consent. The consequences extend beyond efficiency gains; they challenge how we define creativity, authorship, and value in the digital era.

To better understand these shifts, it is helpful to consider three overlapping groups of actors within the evolving ecosystem of online creative production:

These are visual artists, illustrators, and other creatives who typically showcase and distribute their work online. Unlike unionized entertainment professionals—such as the voice actors and performers represented by SAG-AFTRA—many internet-based artists lack strong collective bargaining structures. Traditionally, artists have relied on copyright law to protect their work, allowing them to license their art, negotiate commissions, and ensure they receive credit and payment. Platforms like Pateon, Pixiv, and Twitter help facilitate these transactions. However, generative AI tools complicate these protections by blending myriad styles, including those of artists who never agreed to have their work used as training material. This unauthorized appropriation raises questions about authorship and fair use that current legal frameworks struggle to address.

Another emerging category comprises people who use generative AI models to create images by carefully crafting prompts. They may present these machine-generated pieces as original “art,” building portfolios and selling outputs online. On platforms like Twitter, it can be challenging for patrons and fans to discern whether a piece results from human labor and skill or algorithmic generation guided by prompt engineers. These blurred boundaries undermine traditional notions of artistic integrity and have sparked debate about what should be considered authentic creative output. Traditional copyright approaches—centered on the idea of a human author—offer limited guidance on classifying works resulting from AI-driven processes.

Finally, the companies and developers behind AI art tools are pivotal players. They often focus on refining the capabilities and efficiency of their models, guided by market demands and technological ambition. While their aim may be to push the boundaries of innovation rather than exploit individual creators, the widespread use of training data scraped from countless online sources—including copyrighted artworks—raises significant legal and ethical questions. Current copyright law was never designed to address large-scale, automated ingestion of visual and audio works. This leaves regulators, courts, and legal scholars debating whether new policies are needed to ensure that developers acquire training data ethically and that rights holders have a say in how their works are employed.

Methods and framework

I will use Actor Network Theory (ANT) to analyze the issue of AI-generated art. ANT allows for modeling interactions between different entities, referred to as "actors," which can include

human and non-human participants (Sismondo, 2010, p. 81). According to ANT, interactions comprise the influences actors exert on each other within a network. For this topic, key actors include artists, users of AI art generators, their developers, the AI tools themselves, and the U.S. government's copyright laws. The dynamics between these actors will be explored in greater detail later on.

My research will primarily involve reviewing existing literature and examining social and public policies related to the topic. I want to dive deep into the cases of Sag-Aftra strikes since they are a prominent point in this discussion; in addition to reviewing copyright law that does let us impact this scene, since this can be categorized as free use, I want to examine how that might be applied to cases like this one and whether said formalities and rulings apply to Internet Artists.

Timeline

To research this paper, I plan to incorporate insights from the technical portion of the project to understand how AI generation tools function. In addition, I will conduct a comprehensive review of existing copyright laws and examine relevant agreements negotiated by organizations like SAG-AFTRA and media companies. By mapping out the legal protections currently in place and considering what recent collective actions have achieved, this question aims to clarify the evolving environment in which artists and AI-generated works interact.

In conclusion, the increasing presence of AI in the creative sector complicates traditional understandings of artistry, authorship, and the structures intended to protect creative labor. As generative tools become more capable and prevalent, the relationships among artists, developers, legal frameworks, and industry organizations continue to evolve, often in unpredictable ways. Rather than offering clear-cut resolutions, these developments invite ongoing inquiry into how

value is assigned, who holds the rights to creative outputs, and what role collective action can play. By further examining cases like SAG-AFTRA's negotiations and investigating how current legal mechanisms adapt—or fail to adapt—to new forms of production, researchers can gain deeper insight into the socio-technical networks that shape and redefine creativity in the age of AI.

References:

University of Chicago. (2024). What is Glaze? *Glaze Project*. https://glaze.cs.uchicago.edu/what-is-glaze.html

Shan, S., Ding, W., Passananti, J., Zheng, H., & Zhao, B. Y. (2024). Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. *Proceedings of the 45th IEEE Symposium on Security and Privacy*. https://doi.org/10.48550/arXiv.2310.13828

Stanford Institute for Human-Centered Artificial Intelligence (HAI). (n.d.). *The state of AI in 9 charts*. Retrieved December 16, 2024, from https://hai.stanford.edu/news/state-ai-9-charts

Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. 2023. AI Art and its Impact on Artists. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23). ACM, New York, NY, USA, 363–374. https://doi.org/10.1145/3600211.3604681

Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: A review. *Artificial Intelligence Review, 55*(1), 589–656. https://doi.org/10.1007/s10462-021-10039-7

Vimpari, V., Kultima, A., Hämäläinen, P., & Guckelsberger, C. (2023). "An adapt-or-die type of situation": Perception, adoption, and use of text-to-image-generation AI by game

industry professionals. *Proceedings of the ACM on Human-Computer Interaction, 7*(CHI PLAY), Article 379. https://doi.org/10.1145/3611025

SAG-AFTRA. (2024, October 28). *New SAG-AFTRA and Ethovox agreement empowers actors and secures essential AI guardrails*. https://www.sagaftra.org/new-sag-aftra-and-ethovox-agreement-empowers-actors-and-secures-essential-ai-guardrails

Cerullo, M. (2023, November 14). *The SAG-AFTRA strike is over. Here are 6 things actors got in the new contract*. CBS News. https://www.cbsnews.com/news/sag-aftra-contract-deal-agreement-actors-ai/

Hosanagar, K., & Saxena, A. (2023, April 11). *How generative AI could disrupt creative work*. Harvard Business Review. https://hbr.org/2023/04/how-generative-ai-could-disrupt-creative-work

Murray, M. D. (2023). *Generative and AI authored artworks and copyright law*. Hastings Communications and Entertainment Law Journal, 45(1). https://doi.org/10.2139/ssrn.4152484