Protecting Intellectual Property of Artists from Generative AI

Sadia Nourin, Neal Machado, Kiran Muthusamy
Contact: snourin@umd.edu, nmachado@umd.edu, kiran7@umd.edu

Abstract

Modern technology has long contributed powerful tools for creating art, recently culminating in the use of generative AI models for text-to-image creation (in this paper referred to as AI art). The dynamics between these models and present-day artists are complex, which we investigate in two parts; in Part I we focus on describing the impact of these models on the intellectual property rights and livelihoods of living artists whose works are used in the training of said models, and in Part II we propose a policy intervention to remedy the problems faced by artists as their intellectual property contributes to systems which threaten their existence, without their consent.

I. Background and Context

Overview

The stakeholders surrounding the issue of artists' intellectual property being used to train AI art models without the artists' consent are far and wide. They include Big Tech and AI firms that create these technologies, Artists, Consumers (including both businesses who use these models as well as everyday people who use and/or consume the output of these models), computer science researchers, and policymakers. The arguments surrounding this technology are varied and nuanced, but some basic summaries are noted below.

The main argument that Big Tech uses as justification for this technology is the promise that these tools will "significantly democratize creativity," providing those without the means or the time to spend honing an artistic craft to be able to still "translate their ideas into images," as described by OpenAI (creator of image generation tool DALL·E) in their public comment to the U.S. Office of Copyright's 2023 Notice of Inquiry (NOI) about AI and copyright [16]. With regards to copyright, they claim that "the training of AI models qualifies as a fair use," saying that when they train their models on images they are extracting "factual metadata and fundamental information" that "are not protected by copyright law" [16]. However, they also mention that they work hard to "prevent memorization or repetition" of training data, are working to add opt-out abilities for artists, and have asked rights holders to "identify sites on the internet that reproduce their copyrighted works," in order to not train on them [16].

The viewpoint of most artists is a very disenfranchised one, as without the works of artists (and all of their investments and sacrifices of time, resources, etc.) these generative models would not exist at all, and the effects of these very models are very damaging to the livelihoods of actual human artists. There are many examples of artists rebuking these models, such as Singaporean illustrator Nur Sabrina commenting that "AI art in Singapore will essentially destroy local art and urban culture" [38], and American freelance artist Caryn Chong lamenting how generative AI is "oversaturating the market with products made with generative AI, thereby decreasing the visibility of less established creators to potential clients" [17].

Everyday consumers generally tend to be more surprised, optimistic, but also confused and worried when viewing Generative AI images [35]. This is as opposed to being more happy, joyful, and content when viewing human created images [35]. Furthermore, consumers are concerned about manipulation, misuse, and plagiarism, especially as it pertains to AI-generated images in marketing [35]. Computer science researchers are generally split

into two camps: defensive researchers who are trying to make a technical solution to protect artists' works from being trained on, and attackers who prefer a non-technical solution. Furthermore, we can surmise that a third camp of researchers exist, those who are only focused on the improvement of Generative AI art models and do not consider the repercussions that training these models unrestrictedly can have on artists. On the other hand, policymakers are very focused on protecting artists, yet are still stuck in the phase of researching this issue, such as through the 2023 NOI mentioned previously (and further elaborated in the following section, History of Regulation).

Furthermore, these Generative AI art models can mirror the existing inequities present within our world. The images these models are trained on (as well as the images left out of the training datasets) have a huge effect on the outputs of the models and can end up propagating various cultural ideals. For example, if a model was trained using only European works of art from the 1800s-1900s, a prompt for a "beautiful woman in a dress" will likely result in only European ideals of beauty and fashion trends from the time. Our report does not focus on the aspects of how Generative AI art models can perpetuate existing inequities, nor does our proposed policy solution combat this, but we believe it's important to note that, like with all AI models, Generative AI art models are not exempt from perpetuating biases.

History of Regulation

There is a very long and well-documented history of artists copying and being inspired by the works of their predecessors and contemporaries. From an educational standpoint, copying allows "the young artist to acquire a vocabulary" to be able to translate experience and the physical world onto a surface as art [1]. Through the centuries, much debate (and subsequent regulation) have come about in various societies, such as the Engravers' Copyright Act passed in London in 1735, important to our discussion since much of American laws extend from the precedents set by British Common Law. This Act, also known as the Hogarth Act (after famous British artist William Hogarth) was directed specifically towards engravers, granting them the "exclusive rights for a period of 14 years" for the works they created [2]. This act served to extend the concept of copyright from the literary world to the fine arts world as well.

In the United States, probably the most important piece of legislation surrounding the rights an artist has with respect to their works was Section 107 of the Copyright Act of 1976 (17 U.S. Code § 107). This act established the concept of "fair use," first enumerating the scenarios under which "fair use" can apply, as well describing four factors to be evaluated if the initial conditions are met. Namely, copyrighted works can be fairly used only "for purposes such as criticism, comment, news reporting, teaching, …, scholarship, or research" [3], with the four factors including the purpose/character of the work (including whether commercial or educational), the "nature" of the work, the substantiality of the copyrighted work used in the fair use, and the "effect of the use upon the potential market for or value of the copyrighted work" [3]. Most notable to the stakeholders in this issue are likely the first and fourth factors, since if commercial work is found "transformative" in nature it is sometimes permissible fair use, and the market impacts and harms by generative AI are salient [4].

In addition to the intellectual rights of their works, the concept of an artist's Moral Rights (a.k.a. Droit moral) is pertinent to this issue. Moral rights recognize that a right of "integrity" to a work remains with the original artist, even after the work has been sold [5, page 490]. This integrity is a "moral or non property attribute of intellectual and moral character" realizing a relationship between artist and art that transcends the sale of a piece. Formalized in the Berne Convention for Protection of Literary and Artistic Works of 1886, such moral rights allowed the artist to "claim authorship" of their works as well as "object to any mutilation, deformation or other modification…or other derogatory action" to their artworks, specifically in cases which would be "prejudicial to the author's honor or reputation" [6]. The U.S. joined Berne more than a century later in 1990

with an amendment to copyright law called the Visual Artists Rights Act (VARA) [7]. These rights may come into play with regard to the opt-out nature of much scraping and training techniques for generative models. Even if works are sold by an artist, they may have the right to object to said works being used for training purposes, if they see it as damaging to the reputation or meaning of their works.

The U.S. Copyright Office has begun to offer some literature and review regarding AI and copyright laws. In 2023, the Office began an initiative to examine the subject, allowing for open public comments from stakeholders – the NOI (notice of inquiry) previously mentioned [7]. In 2024 the Copyright Office published the first section of its report (focused on Digital Replicas e.g. Deep Fakes), with future sections to detail "registration of works containing AI," "training AI models on copyrighted works," and more originally slated for a 2024 release [8]. We expect the latter release to encompass a lot of the issues which we focus on within this report. The U.S. Copyright Office has also published a document detailing more about its Human Authorship requirement for registering a copyrighted work, and, like many of the Office's stipulations, describe the process being a case-by-case basis depending on the amount of work each entity contributes to the piece. They do mention, however, that solely formulating a prompt, for example, is not enough to constitute Human Authorship since the generative model is doing the actual "traditional elements of authorship" [9, page 4].

Existing Legal Regimes

With regards to existing legal regimes, the EU has the strongest one, with the passing of the EU Copyright Directive in 2019 [29]. This directive establishes a robust framework for protecting creators, particularly through the implementation of Article 17, which makes platforms and services more accountable for content infringement. Artists have a solid chance of obtaining redress due to strong copyright protections, as moral rights are more expansive in the EU compared to in other regimes [39]. Platforms are required to ensure that any copyrighted content is used with the appropriate licenses or permissions. The combination of collective licensing systems and strict interpretations of copyright law positions the EU as one of the most favorable regions for artists seeking justice.

In the US, copyright laws are also quite strong, especially regarding derivative works. If an AI model creates content based on an artist's work without permission, the artist may have grounds to claim infringement, depending on the nature of the use and how transformative the new work is. Additionally, the Right of Publicity in various states provides further protection for artists, particularly when their name or likeness is involved. Although the Fair Use Doctrine introduces some uncertainty, recent legal rulings in the US generally favor artists, especially in cases of clear commercial exploitation or minimal transformation of the original work by AI [30].

In Singapore, the Copyright Act (2021 Amendments) outlines the protection of artists' intellectual property. In contrast to the US, Singapore features exceptions for Text and Data Mining (TDM), which permit AI model training for non-commercial research without requiring permission from rights holders. However, copyright protections are still robust for commercial applications. The Intellectual Property Office of Singapore (IPOS) is responsible for enforcing IP rights, ensuring a balance between fostering innovation in AI development and safeguarding artists from unauthorized use of their creations [31].

Op-eds, News Articles, and Magazine Articles

Unfortunately, reading through legal codices leaves much to be desired when it comes to understanding the perspectives of the artists that the laws are trying to protect. In this section, we try to understand how artists feel about the advent of Generative AI and whether they believe enough is being done to protect their livelihood.

Sarah Andersen, a cartoonist who is suing Midjourney, Stability AI and DeviantArt for using her work without consent, emphasizes

the importance of safeguarding artists' intellectual property from AI models [32]. She is worried about AI's ability to replicate her distinctive artistic style without her permission, which she views as a breach of her identity. She points out that AI models rely on datasets that often contain copyrighted material, frequently sourced without consent, which endangers artists' livelihoods. Although she doesn't completely reject AI technology, Sarah feels overlooked by the disregard for artists' rights and cautions that unregulated AI usage could result in significant exploitation of creative work, especially affecting marginalized communities.

Jason Allen, an artist who used AI and won first place in the digital art category at the Colorado State Fair, believes that artists who utilize AI tools, including himself, deserve recognition and protection under copyright laws. Allen sees AI as a tool, similar to a brush or camera, and stresses that human creativity is still at the heart of art creation, even when AI is involved [33].

Loish, another popular digital artist, argues that artists' intellectual property needs protection from AI models as numerous AI tools collect artwork without permission, which undermines the rights of creators. She supports the creation of ethically sourced databases for AI that honor copyright and guarantee fair compensation for artists. She stresses that artists should have the authority to manage how their work is used in training AI models and is against the exploitation of their creativity for profit without proper recognition or payment [34].

We can observe through these viewpoints of different artists that there are a myriad of harms involved with AI models training on artists' images without their consent. First, artists experience economic harm because they do not receive any type of compensation when their images are used to train AI art models. Second, artists experience reputational harm when AI art model users represent a concept or topic in the artist's style that the artist would never personally endorse or create. Finally, artists experience emotional harm simply when AI art model users create art in the style of a specific artist. Artists take years to hone their craft and when someone else replicates their style in the blink of an eye using AI, artists can feel as if identity has been stolen, as their identities are often so intricately linked to their art [32].

However, all of these harms are subsumed by the true harm that is caused by the current state of AI art model training—consent violation—as artists do not have any say on the matter of whether their images can or cannot be used for AI art model training. Consent violation spawns all of the various harms listed above (financial, reputational, emotional, etc.) [40]. Therefore, our proposed intervention attempts to solve this consent violation harm, and furthermore, the economic harms faced by artists whose artwork is being used to train AI art models.

Technical Research

Text-to-image diffusion models are a type of Generative AI model that has been developed in the past few years that, given a user prompt, can generate, often photorealistic images, pertaining to the prompt. Text-to-image diffusion models were first developed in 2017 by DALL-E, one of the first models on the market, using another new technology called transformers [22, 28]. However, DALL-E was having trouble with generating photo-realistic images. In comes stable diffusion, which allowed DALL-E 2, Imagen, and Midjourney to flourish [28]. However, in order to create these photo-realistic images, these models needed to have data to be trained upon. DALL-E 2 specifically states that it was trained upon "publicly available sources and sources that we licensed", Imagen uses the LAION-400M dataset, a dataset of websites from the Common Crawl project that scraped websites from 2014 and 2021, while Midjourney used the LAION5B dataset, a larger dataset than the LAION-400M [23, 24, 25, 26]. Some of these web scrapings were done with no human-in-the-loop, resulting in these models being trained upon many works of art from artists, without their consent [27].

In an effort to give power back to artists in this new era of Generative AI, there have been many technical developments within this space to protect artists, ranging from solutions

that try to protect the intellectual property of artists to those that try to break such solutions. Perhaps the most noteworthy of these developments include The Glaze Project, a project developed by researchers at the University of Chicago that aims to protect "human creatives" against the use of Generative AI. Their flagship paper, Glaze, aims to safeguard artists' creations from being used as training data for text-to-image diffusion models that generate art based on user prompts [19]. Glaze works by adding a "style cloak" to artwork before the artist shares the artwork online. This "style cloak" adds perturbations to the artwork that is nearly invisible to the human eye but which confuses generative AI models that try to mimic the artists' style into thinking that the image is actually of a different style. As a result, the model generates an image in a different style when given a prompt associated with the artist's original style. Since the release of the paper, many artists have begun to use Glaze to protect their art.

Soon after, Nightshade was released by The Glaze Project, an attack that artists can deploy against text-to-image diffusion models that scrape artists' images against their consent. Like Glaze, Nightshade perturbs the artists' image in such a way that is almost imperceptible to the human eye, but which manages to deceive text-to-image models into interpreting an image as one theme, when it actually depicts another [20]. When enough of an artist's images have Nightshade applied to them, the model begins to struggle with distinguishing between the text tied to the original theme, as it now associates that text with the altered theme instead. Nightshade is meant as an offensive position that artists can take against these text-to-image diffusion models while Glaze is more defensive. However, the researchers from The Glaze Project state that ideally, artists should use both tools for maximum protection.

It has yet to be determined whether tools such as Glaze and Nightshade can cause artists to face legal repercussions. According to the The Computer Fraud and Abuse Act (CFAA) in the U.S. (18 U.S. Code § 1030), anyone who accesses a computer system without proper authorization, or who damages a computer system, will be prosecuted [41]. Although artists are applying perturbations to their own art in an attempt to protect their art from being trained on AI art models, in theory, model developers can make the claim that artists have tampered and "damaged" their AI art models by using Glaze and Nightshade, and take artists to court under the CFAA. Whether such a claim will hold in court however, is not yet foreseeable.

As with any security defense, attacks will soon follow that try to break the defense, resulting in a never ending cat and mouse game. Glaze and Nightshade are great defenses against text-to-image diffusion models that steal artists' work, but even the researchers who created these tools knew that the defenses would not last forever. Researchers from Google Deepmind and ETH Zurich discovered that text-to-image diffusion models can bypass Glaze and mimic artists' work by simply switching the JPEG compression of artists' images, originally used to mimic artists' work, to JPEG compression with Gaussian noising [21]. This simple switch leaves artists' work vulnerable to being trained on again. The researchers behind the attack emphasized the importance of continuing to develop technical solutions to protect artists, and they hope that their work will inform others about also improving the non-technical protections that artists need in this new world with Generative AI.

Proposed Legislation and Rules

In terms of current regulation there are a couple of federal legislation that has been proposed over the past year regarding artists and generative AI, but not too much on the local level. When examining the list compiled by the National Conference of State Legislature, of all 2023 state legislation regarding AI there were very few efforts which focused on generative AI [10]. A few states attempted (and sometimes succeeded) in creating task forces to investigate more into how AI is affecting constituents, such as a proposed study by the New York Department of Labor, which included a stipulation that the state would not use AI to displace employees in any way until after the report is completed [11]. Michigan, New York and Wisconsin state legislatures all had proposed

legislation regarding the disclosure of generated media in any political communication [10].

With regard to the federal level, first and foremost the U.S. Copyright Office report on training AI models will likely be very important, whenever that is released. There are a couple notable attempts to introduce legislation protecting artists from generative AI, namely the Generative AI Copyright Disclosure Act and the COPIED Act, introduced in April and July of 2024, respectively. The former piece of legislation, if passed, would require a "sufficiently detailed summary of any copyrighted works used" in either training or tuning of any dataset "used in building a generative AI system" [12]. In theory this increase in transparency is a good thing for artists and is supported by industry groups such as RIAA, SAG-AFTRA, and the Writers Guilds [13], however the ambiguity in the wording of the bill leaves much in the air (e.g. does this apply for all model creators at all scales?). The second piece of legislation involves NIST developing "guidelines…for content provenance" (e.g. watermarking of generated content) and also provides for artists to be able to attach nonremovable provenance information which would allow them to "protect their work" and "set terms of use" for their pieces [14]. As always there are tradeoffs to every piece of proposed legislation; one criticism of the COPIED act is that this new declaration of provenance completely goes in the face of all fair use, even though there are many cases when copyrighted works can in fact be used fairly [15].

II. Proposed Intervention

Overview

Our proposed intervention to protect the intellectual property of artists from AI models consists of a policy solution step and a technological solution step, both of which intertwine together. The first policy solution step is multi-faceted and requires artists to tag their online artwork with metadata that will indicate whether they want to opt-in or opt-out each piece of their artwork from AI art model training datasets. When AI art model developers scrape the internet to obtain images for their training datasets, they are required to omit the images tagged with the opt-out metadata. Furthermore, AI art model developers will be legally required to publish their training datasets. These datasets will be hosted on a platform where artists can query the datasets to determine whether any of their artwork was used to train AI art models without their consent. If artists notice that the artwork that they had opted-out of appears in the training dataset(s), artists have a right to sue the AI art model developers under our proposed regulation.

As a result, AI art models will only be developed using artwork that artists have already consented to being used for AI art model training. However, AI art model developers will also need to compensate the artists whose images have been used in the training dataset. The compensation will be on a per-image basis. Each time a user generates an image using the AI art model, artists whose images have been used to create the image will be compensated. This compensation will be distributed on a microcent scale. This is where the second step of our proposed intervention, the technological solution portion, comes into play. Researchers will need to understand exactly which images were used to generate the image outputted by the AI art model. Ideally, multiple artists will get compensated based on what percentage of their images were used to generate each image outputted by the model. This is unfortunately still an active area of research within the AI explainability community, as computer science researchers do not understand how these AI art models work under the hood.

Assuming researchers make substantial headway in understanding the explainability of AI art models in the near future, we encounter a win-win situation with our proposed intervention where AI art model developers are allowed to keep training on artists' artwork and artists get compensated for allowing models to be trained on their artwork.

Below we dive into the specifics of our proposed intervention a bit more.

Metadata Tagging, Public Training Datasets, and Artist Profiles

Artists will use EXIF metadata to tag how their art is used by AI developers. Artists can use either opt-in or opt-out tags on each of their images. As a result, artists can now choose whether or not to consent to allowing each piece of their artwork to be included within the AI art models' training data. This method directly solves the problem of unauthorized use of copyrighted content: a metadata tagging system can help artists protect their intellectual property and ensure that AI developers respect artists' decisions. Additionally, having a platform that mandates that AI developers publish their training datasets promotes transparency and accountability. By mandating that AI developers reveal the data they use, artists can check these datasets to see if their work has been used without their permission.

EXIF metadata is inherently linked to the specific file of an image that is uploaded to the internet. As a result, it is trivial to take a screenshot or make a copy of the original image file that an artist has uploaded on the internet and change the EXIF metadata tag to an opt-in tag, even though the original image contained an opt-out tag. As such, it is the model developer's responsibility to check whether their training data contains copies of an image that originally held the opt-out metadata flag. We place this burden on the model developer instead of the artist because the model developer is the one who has the entire training dataset, and it is much easier for them to conduct reverse image searches on the internet to determine whether they have opt-in copies of an original image with the opt-out tag.

However, we do understand that there will be substantial pushback from the AI community when it comes to publishing their training datasets. Although some AI models do publicly release their training datasets, others do not because of fear that AI developers would lose out on competition or that users may find illegal or sensitive information within their datasets [44]. We must work with policy makers and the AI community to ensure that we do not become a victim of regulatory capture when trying to implement transparency policies for training datasets.

At the same time, we do understand that having model developers publicly release their training data is not mutually exclusive with the transparency of the training data. We can also use PETs (Privacy Enhancing Technologies) to determine whether or not a specific image is a member of a training dataset. However, we choose not to use PETS as we believe that it would be easy for model developers to game the system. If we want to create a PET for set membership of images, then we may have to encode all of the pixels of each training image within the dataset. What if model developers edit a single pixel of each training image or screenshots each training image at a lower resolution, and then encodes the image? The resulting image within the training dataset would be almost identical to the original training image. If an artist were to query the training dataset with their own image, they would not be able to find it their image, and would believe that their opt-out request was sufficiently respected, even though in reality, it was not. As a result, it's much better to just be completely transparent and make the entire training dataset public. This way, artists can not only search for their own images, but also similar images, to ensure that model developers are properly respecting their opt-in and opt-out requests.

Our proposed intervention will also create another public platform for artists. This public platform will ideally be controlled by a government contractor or non-profit agency, and will allow users to register their artist profile on the platform. Artist profiles on the platform will allow the legitimate artists to manage their artist name (which will be used to query for their artwork on the platform that hosts the publicly available training datasets), their art, and their payment options (to be discussed in the Economics of Microcent-based Compensation for Artists section below). We understand that many artists use pseudonyms, as opposed to their real name, when publishing their artwork and would like them to continue to to remain anonymous on this platform. In general this is fine, except when artists would like to remain anonymous while receiving compensation.

Artists who would like to receive payment by anonymous means will have to do so through an NGO. This NGO will act as an intermediary and a payment handler. We will place a legal non-disclosure of identity on these NGOs. This setup will ensure that artists can maintain control over their profile and settings in an anonymous and decentralized, but still traceable manner, if they choose to do so.

If an artist queries the public training datasets and discovers that their image was used for AI model training without their consent, then the artist can collect damages. Artists can collect these damages under a new policy that will be created as a result of our proposed intervention, not under any existing part of the copyright law. We choose this method of redress for artists because the U.S. government still hasn't decided on how to address AI with regards to copyright infringement laws, and we can't wait for them to decide what can and cannot be sued — especially if they rule that users cannot sue if their work was used to train an AI model.

Furthermore, artists can claim damages based on the scale of the model. If an artists' image has been used to generate many outputs from a model, then the artist can claim damages on every single image that has been outputted. Model developers will need to train their entire existing model without the image that the artist is collecting damages for. This will be expensive for the model developer, but this punishment seems just due to how simple it is to just not include images that artists have opted out of for AI training.

Tracking Influence of Artists' Images in AI Art Models

Monitoring the extent to which an artist's work contributes to the production of an AI model's output is the most difficult aspect of this intervention. The challenge is to develop a comprehensive explainability framework that can determine whether images in a training dataset influence the generation of a particular AI outcome. These results comply with the current research on AI's explainability, which try to understand how models make decisions based on training data.

There are a few existing studies out there that attempt to tackle this problem of tracking influence within image generation. One such study by Carmichael et al. uses Pixel-Grounded Prototypical Part Networks to try to understand how specific sections of an image are related to another image (i.e., such as an image in the training dataset) [42]. Furthermore, another study done by Carlini et al. showed how diffusion models can "memorize" training dataset images and output such images exactly [43]. This suggests the potential for future methods to trace which specific training images contributed to a given output, provided the model hasn't memorized the data entirely.

Advancements in AI art explainability have been exciting in recent years, yet have not fully solved the problem. Explanation is still an active area of research and additional development is needed to create a system that can monitor the precise contribution of each image required to adequately compensate an artist. As a result, our proposed intervention still needs more research to implement properly.

Although we cannot say how this technological portion of our solution will be implemented by researchers and model developers, we do have a suggestion for what an appropriate explainability notion would reveal: the "influence" of some image x within the training corpus on an output image Y is the percentage difference between Y and Y', where Y' is the same generated image (i.e. with the same prompt) with x eliminated from the training corpus. In this way, the influence of x is the amount it "contributes" to a certain output image, where percentage difference can be calculated in a number of ways (any type of appropriate norm/metric for images such as feature level differences, pixel level difference, etc). Of course, this is a very loose definition that still needs to be flushed out more, but its quantifiable nature and similarity to differential privacy leads to some interesting potential, as discussed in the discussion section.

In order to incentivize research in this area, we will host a NIST competition, similar to the NIST competitions hosted for post-quantum cryptography [37].

Economics of Microcent-based Compensation for Artists

The main goal of our proposal is to help address the imbalance between the allocation of the revenue earned from model outputs and the external costs placed on artists for their work. This is where the trackable influence of training data is essential: for every image outputted, we require a minimum portion $0 < \lambda \leq 1$ of the associated revenue to be split proportionately among the artists whose works have a predominant influence on the output image. Since there may be a huge number of artists' works that contribute only minimally to an output image, we can require a minimum threshold (decided by legislators) for how much influence is required – say, >1% – for an artist to be proportionally compensated for each training image.

The value λ can also be decided by legislators, although we recommend that a majority of the profits be allotted for artists ($\lambda > 0.5$). These technologies wouldn't exist in magnitude and scale without the R&D and powerful compute that large corporations provide, but they would be utterly obsolete without the hard work of the artists whose works are used for training. We chose a system where artists receive a portion of revenue instead of an upfront, fixed payout since we feel like this offers the most fair compensation and mitigates the scenario where artists are paid only a small amount upfront while the model makers profit perpetually. The reason why revenues (and not strictly profits) are to be split is because the external costs to artists persist no matter how profitable these models are; these splits address the negative externalities in any case.

These measures may seem very costly to the companies creating these models and as such we expect these corporations to be the largest parties in opposition. Furthermore, we also expect some pushback from non-profit generative AI companies as well as businesses who utilize such models, as they may experience heightened prices. We understand that all regulation is a compromise and we want both artists and AI companies alike to find success under our proposal; this is why we introduced a flexible component λ left to the discretion of policymakers. That being said, if AI companies are unwilling to adhere to the compensation policies for modern artists, there are centuries of work by master artists which can be used as training data for no additional cost (assuming that these master artists' works are open for fair use).

We also imagine a new market developing if this proposal is implemented, in which artists and companies can negotiate and decide on market prices and splits. We hypothesize that many hobbyist artists or smaller artists making little money may choose to opt in, while more distinguished artists would likely opt out in order to avoid contributing to tools which cannibalize their own market. This may incentivize some companies to offer better splits (higher values for λ), with the eventual emergence of different models for different desired qualities of art.

Discussion and Limitations

There are many instances reliant on trust under our proposed intervention. Artists will have to trust that model developers are honest in their disclosure of the images used in their training. Our current system places the burden on the artists themselves to query through the public training datasets and ensure that their decisions to opt out of training are respected. As discussed in the Metadata Tagging, Public Datasets and Artist Profiles section there are many ways in which model developers can act in bad faith, and artists are not necessarily equipped to discern them all. In the future, we'd like to see enforcement in the hands of a government agency instead. Further, a lot of trust will be placed in the intermediary NGO, in order to handle payments and maintain the privacy of artists.

One of the difficulties of regulating this type of technology is that models can be easily separated from the entity that creates it, creating an accountability vacuum. A model can be trained with images taken without consent, and then its parameters can be dumped online (e.g. on GitHub or HuggingFace) for anyone to use. In this case, as long as the model is not used to

generate revenue our proposed regulation would not apply. We are looking to correct a failure of for-profit model developers not compensating artists for the value they create; if no value is being generated then our proposed intervention has no standing.

An entity could also train a model and run it locally or download such parameters from an online host, generating images (and potentially value for their enterprise) without selling access to some generation API. In this case we would defer to the existing copyright regime (dependent on the U.S. Copyright Office's report), with courts deciding the appropriate outcomes if artists claim their rights have been violated. Again, since our proposed policy intervention revolves around artists being compensated a certain percentage of revenue generated from these models, if there is no explicit measure of revenue per image then there is no value to apply λ to – artists will have to seek redress through traditional means.

Another potential flaw is the idea of using the outputted images of a model where artists are properly compensated in order to train another model, in which no artists will be entitled to any compensation since the corpus is entirely generated by AI. However, due to model collapse developers are unlikely to want to rely on much (if any) synthetic training data, since doing so will very quickly cause the quality of the model to deteriorate [36].

One of the harms of this technology that our proposed intervention aims to address is the reputational harm and breach of identity that generative AI can cause. Art is a form of communication; generating art in the style of another artist may appear as if said artist is endorsing the message of the generated work. Our remedy for this problem is the ability for artists to completely opt out of this technology, since enumerating a monetary value for this harm is very case-by-case. This is where further research into trackable influence and differential privacy would be potentially fruitful. A differential privacy approach to training a generative model – in which any output image is essentially the same no matter whether or not any one input image is in the training set – would alleviate worries about models learning an artist's style too closely.

III. Conclusion

In this report, we present a deep dive into the history, concerns, stakeholders, and legislation surrounding the issue of using the intellectual property of artists without their consent to train AI art models. We then propose a policy intervention which tackles the consent violation issue and the economic harms that artists experience by having their art trained on by these models. We note that our proposed solution has limitations. Currently, it only addresses two harms, consent violation and economic harms, and causes artists to trust the institutions that will be executing and enforcing the policy. However, it is definitely possible for institutions and model developers to game the system via regulatory capture, modifying the AI art model so that the metric of explainability shows that no training image had a majority influence on an output image even if visually, there are undeniable similarities, etc. Although this proposed intervention has limitations, we believe that this is a first step forward towards protecting the intellectual property of artists in this new era of Generative AI.

References

1. Homburg, C. (1996). *The Copy Turns Original*. John Benjamins Publishing Company.
2. Engravers' Copyright Act, London (1735), Primary Sources on Copyright (1450-1900), eds L. Bently & M. Kretschmer, www.copyrighthistory.org
3. U.S. Copyright Office. (2010). *Chapter 1 - Circular 92 | U.S. Copyright Office*. Copyright.gov. https://www.copyright.gov/title17/92chap1.html#107
4. Jiang, H., Brown, L. T., Cheng, J., Khan, M., Gupta, A., Workman, D., Hanna, A., Flowers, J., & Gebru, T. (2023). AI Art and Its Impact on Artists. AIES '23: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 363–374.

https://doi.org/10.1145/3600211.3604681
5. Fishman, J. J. (1977). The emergence of art law. Clev. St. L. Rev., 26, 481.
6. World Intellectual Property Organization. (2019). *Summary of the Berne Convention for the Protection of Literary and Artistic Works* (1886). World Intellectual Property Organization. https://www.wipo.int/treaties/en/ip/berne/summary_berne.html
7. H.R.2690 - 101st Congress (1989-1990): Visual Artists Rights Act of 1990. (1990, June 11). https://www.congress.gov/bill/101st-congress/house-bill/2690
8. Scheland, N. (2024, March 26). *Looking Forward: The U.S. Copyright Office's AI Initiative in 2024 | Copyright*. The Library of Congress. https://blogs.loc.gov/copyright/2024/03/looking-forward-the-u-s-copyright-offices-ai-initiative-in-2024/
9. Published Document: 2023-05321 (88 FR 16190)
    a. https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence
10. NCSL. (2023, July 20). *Artificial Intelligence 2023 Legislation*. www.ncsl.org; NCSL. https://www.ncsl.org/technology-and-communication/artificial-intelligence-2023-legislation
11. https://custom.statenet.com/public/resources.cgi?id=ID:bill:NY2023000A7838&ciq=ncsl&client_md=bbfdb6333015b967d68de540b40ba8ce&mode=current_text
12. Text - H.R.7913 - 118th Congress (2023-2024): Generative AI Copyright Disclosure Act of 2024. (2024, April 9). https://www.congress.gov/bill/118th-congress/house-bill/7913/text
13. *Rep. Schiff Introduces Groundbreaking Bill to Create AI Transparency Between Creators and Companies*. (2024, April 9). Schiff.house.gov. https://schiff.house.gov/news/press-releases/rep-schiff-introduces-groundbreaking-bill-to-create-ai-transparency-between-creators-and-companies
14. *Cantwell, Blackburn, Heinrich Introduce Legislation to Increase Transparency, Combat AI Deepfakes & Put Journalists, Artists & Songwriters Back in Control of Their Content.* (2024, July 11). U.S. Senate Committee on Commerce, Science, & Transportation. https://www.commerce.senate.gov/2024/7/cantwell-blackburn-heinrich-introduce-legislation-to-combat-ai-deepfakes-put-journalists-artists-songwriters-back-in-control-of-their-content
15. Macpherson, L. (2024, July 24). *The COPIED Act Is an End Run around Copyright Law.* Public Knowledge. https://publicknowledge.org/the-copied-act-is-an-end-run-around-copyright-law/
16. https://www.regulations.gov/comment/COLC-2023-0006-8906
17. https://www.regulations.gov/comment/COLC-2023-0006-7524
18. Hunter-Doniger, T. (2016). The eugenics movement and its impact on art education in the United States. *Arts Education Policy Review*, 118(2), 83–92. https://doi-org.proxy-um.researchport.umd.edu/10.1080/10632913.2015.1051256
19. Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, Ben Y. Zhao. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. Proceedings of 32nd USENIX Security Symposium, August 2023.
20. Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, Ben Y. Zhao. Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models. Proceedings of 45th IEEE Symposium on Security and Privacy, May 2024.
21. Robert Hönig, Javier Rando, Nicholas Carlini, Florian Tramèr. Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI.

arXiv:2406.12027. https://doi.org/10.48550/arXiv.2406.12027.
22. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems (2017).
23. https://github.com/openai/dalle-2-preview/blob/main/system-card.md#model
24. https://laion.ai/blog/laion-400-open-dataset/
25. https://laion.ai/blog/laion-5b/
26. The Upwork Team. "What Is Google Imagen? A Beginner's Guide | Upwork." Upwork, 8 Jan. 2024, www.upwork.com/resources/google-imagen.
27. Salvaggio, Eryk. "Laion-5B, Stable Diffusion 1.5, and the Original Sin of Generative AI." Tech Policy Press, Tech Policy Press, 2 Jan. 2024, www.techpolicy.press/laion5b-stable-diffusion-and-the-original-sin-of-generative-ai/.
28. Dipert, Brian. "From Dall·e to Stable Diffusion: How Do Text-to-Image Generation Models Work?" Edge AI and Vision Alliance, 13 Sept. 2023, www.edge-ai-vision.com/2023/01/from-dall%C2%B7e-to-stable-diffusion-how-do-text-to-image-generation-models-work/
29. New EU copyright rules that will benefit creators, businesses and consumers start to apply, 03 Jun 2021 https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1807
30. Generative Artificial Intelligence and Copyright Law, 29 Sep 2023 https://crsreports.congress.gov/product/pdf/LSB/LSB10922
31. Text and data mining in Singapore, 05 Feb 2024 https://www.reedsmith.com/en/perspectives/ai-in-entertainment-and-media/2024/02/text-and-data-mining-in-singapore
32. AI and Artists' IP: Exploring Copyright Infringement Allegations in Andersen v. StabilityAI, 26 Feb 2024 https://itsartlaw.org/2024/02/26/artificial-intelligence-and-artists-intellectual-property-unpacking-copyright-infringement-allegations-in-andersen-v-stability-ai-ltd/
33. AI-Generated Art Won a Prize. Artists Aren't Happy, 02 Sep 2022 https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html
34. Loish Blog, 15 Dec 2022 https://blog.loish.net/post/703723938473181184/theres-a-protest-going-on-against-ai-art-over-on
35. Conjointly. "What Do Consumers Think of Generative AI in Marketing?" Conjointly, 22 June 2023, conjointly.com/blog/generative-ai-in-marketing/.
36. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature, 631(8022)*, 755-759.
37. NIST. (2017, January 3). Post-Quantum Cryptography . CSRC. https://csrc.nist.gov/projects/post-quantum-cryptography.
38. Chan, G. (2023, May 26). *Ai-generated artwork by Bank employees raises questions of copyright and artistic licence*. The Straits Times. https://www.straitstimes.com/singapore/ai-generated-artwork-by-bank-employees-raises-questions-of-copyright-and-artistic-licence.
39. Wadsworth, T. J. (2021, November 23). *AI-created art and the moral rights of authors*. Columbia Journal of Transnational Law. https://www.jtl.columbia.edu/bulletin-blog/ai-created-art-and-the-moral-rights-of-authors
40. Corrall, M. (2024, January 16). *The Harm & Hypocrisy of ai art*. Matt Corrall. https://www.corralldesign.com/writing/ai-harm-hypocrisy
41. United States. (1986). *Computer Fraud and Abuse Act of 1986*, 18 U.S.C. § 1030.

https://uscode.house.gov/view.xhtml?req=(title:18%20section:1030%20edition:prelim)
42. Carmichael, Z., Lohit, S., Cherian, A., Jones, M. J., & Scheirer, W. J. (2024). Pixel-grounded prototypical part networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 4768-4779).
43. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., ... & Wallace, E. (2023). Extracting training data from diffusion models. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 5253-5270).
44. Hardinges, J., Simperl, E., & Shadbolt, N. (2024, May 31). We must fix the lack of transparency around the data used to train foundation models. Harvard Data Science Review. https://hdsr.mitpress.mit.edu/pub/xau9dza3/release/2