Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain (The Short Version)

Katherine Lee*
The GenLaw Center
San Francisco, CA, USA
Cornell University
Ithaca, NY, USA

A. Feder Cooper*
The GenLaw Center
San Francisco, CA, USA
Cornell University
Ithaca, NY, USA

James Grimmelmann*
The GenLaw Center
San Francisco, CA, USA
Cornell Tech & Cornell Law School
New York, NY, USA

ABSTRACT

"Does generative AI infringe copyright?" is an urgent question. It is also a difficult question, for two reasons. First, "generative AI" is not just one product from one company. It is a catch-all name for a massive ecosystem of loosely related technologies. These systems behave differently and raise different legal issues. Second, copyright law is notoriously complicated, and generative-AI systems manage to touch on a great many corners of it. They raise issues of authorship, similarity, direct and indirect liability, and fair use, among much else. These issues cannot be analyzed in isolation, because there are connections everywhere. We aim to bring order to the chaos. To do so, we introduce the **generative-AI supply chain**: an interconnected set of stages that transform training data into generations. The supply chain reveals all of the places at which companies and users make choices that have copyright consequences. It enables us to trace the effects of upstream technical designs on downstream uses, and to assess who in these complicated sociotechnical systems bears responsibility for infringement when it happens. Because we engage so closely with the technology of generative AI, we are able to shed more light on the copyright questions. We identify the key decisions that courts will need to make as they grapple with these issues, and point out the consequences that would likely flow from different liability regimes. This article is a much-abbreviated version of a forthcoming law review article at *The Journal of the Copyright Society*.

1 INTRODUCTION

Generative-AI systems like ChatGPT, Bard, DALL·E, and Ideogram can turn a user-supplied prompt like `"give three arguments why marbury v. madison was wrongly decided"` into a persuasive essay, or `"a cowboy riding a rocket ship"` into a work of digital art. They are unpredictable and complex; they break out of existing legal categories. In particular, because generative-AI systems are trained on millions of examples of human creativity, they raise serious copyright issues. This has not gone unnoticed. Numerous plaintiffs have sued leading generative-AI companies for copyright infringement, with potential damages reaching into the billions of dollars.

This article looks systematically at how copyright applies to generative-AI systems. Our first contribution is to be precise about what "generative AI" is. It is a catch-all term for a massive ecosystem of loosely related technologies, including conversational text chatbots like ChatGPT, image generators like Midjourney and DALL·E, coding assistants like GitHub Copilot, and systems that compose music, create videos, and suggest molecules for new medical drugs. Generative-AI models have different technical architectures and are trained on different kinds of data using different algorithms. Some take months and cost millions of dollars to train; others can be spun up in a weekend. Some models are offered through paid online services; others are distributed open-source, such that anyone could download and modify them.

We take the complexity and diversity of generative-AI systems seriously. We introduce the **generative-AI supply chain**: an interconnected set of stages that transform training data (millions of pictures of cats) into generations (a picture that may never have been seen before of a cat that may not exist). We conceive of eight stages: 1) production of creative works, 2) conversion of creative works into quantified data, 3) creation and curation of training datasets, 4) base model (pre-)training, 5) model fine-tuning to adapt to a specific problem domain, 6) model release or deployment within a software system, 7) generation, and 8) alignment, i.e., adjusting the model and system to advance goals (such as helpfulness, safety, legal compliance). The supply chain is not a simple cascade from data to generations. Instead, each stage is regularly adjusted to better meet the needs of the others. Breaking down generative AI into these constituent stages reveals all of the places at which companies and users make choices that have copyright consequences.

We then work systematically through the copyright analysis of these different stages. Copyright law is notoriously complicated, and generative-AI systems manage to touch on a great many corners of it. They raise issues of authorship, similarity, direct and indirect liability, fair use, and licensing, among much else. These issues cannot be analyzed in isolation, because there are connections everywhere. We trace the effects of upstream technical designs on downstream uses, and assess who in these complicated sociotechnical systems bears responsibility for infringement when it happens. Because we engage so closely with the technology of generative AI, we are able to shed more light on the copyright questions. We identify the key decisions that courts will need to make as they grapple with these issues, and point out the consequences that would likely flow from different liability regimes.

We proceed in three parts. We:

- Describe the generative-AI supply chain in detail, including what happens at each stage, the diversity of variations on the basic theme, and the design choices that the various actors must make to create and use a generative-AI system (Section 2).
- Provide examples of how the supply-chain framing facilitates detailed copyright analysis, covering substantial similarity, direct infringement, and fair use. We ask *what* might possibly be an infringing technical artifact, *who* might be an infringing

*Equal contribution

CSLAW '24, March 12-13, 2024. Boston, MA, USA

ACM ISBN 979-8-4007-03333-1/24-03.
http://doi.org/10.1145/3614407.3643696

actor, and *when* infringement may occur, and discuss how the choices made by actors at one point in the supply chain affect the copyright risks faced by others (Section 3).
- Detail broader lessons, including the options courts have and how they should conceptualize generative AI (Section 4).

Altogether, we argue that copyright pervades the generative-AI supply chain, that fair use is not a silver bullet, that the ordinary business of copyright litigation will continue even in a generative-AI age, and that courts should beware of metaphors that provide too-easy answers to the genuinely hard problems before them. We note that this article is a shortened version of a summer 2023 law review submission.

2 THE GENERATIVE-AI SUPPLY CHAIN

We assume introductory familiarity with machine learning (ML) and generative AI, and delve right into our discussion of the generative-AI supply chain. To begin, we note that one of the big enablers of today's generative-AI systems is scale. Notably, scale complicates *what* technical and creative artifacts are produced, *when* these artifacts are produced and stored, and *who* exactly is involved in the production process. In turn, these considerations are important for how we reason about copyright implications: *what* is potentially an infringing artifact, *when* in the production process it is possible for infringement to occur, and *who* is potentially an infringing actor [Cooper et al. 2022; Yew and Hadfield-Menell 2023].[1]

To provide some structure for reasoning about this complexity, we introduce our abstraction for reasoning about generative AI as a supply chain. We conceive of the **generative-AI supply chain** as having eight stages (see Figure 1): the creation of expressive works (Section 2.1), data creation (Section 2.2), dataset collection and curation (Section 2.3), model (pre-)training (Section 2.4), model fine-tuning (Section 2.5), system deployment (Section 2.6), generation (Section 2.7), and model alignment (Section 2.8). Each stage gathers inputs from prior stage(s) and hands off outputs to subsequent stage(s), which we indicate with (sometimes bidirectional) arrows.

The connections between these supply-chain stages are complicated. In some cases, one stage clearly precedes another (e.g., model pre-training necessarily precedes model fine-tuning), but, for other cases, there are many different possible ways stages can interact, and they may involve different actors. We highlight some of this complexity in the following subsections.

2.1 The Creation of Expressive Works

Artists, writers, coders, and other creators produce expressive works. Generative-AI systems do, too;[2] but state-of-the-art systems are only able to do so because their models have been trained on data derived from pre-existing creative works.[3] It is worth remembering that, historically, the production of most creative works has had nothing to do with ML.[4] Painters have composed canvases, writers have penned articles, etc. without considering how their works might be taken up by automated processes. Nevertheless, these works can be transformed into quantified data objects that can serve as inputs for ML. They can be easily posted on the Internet and circulated widely, making them accessible for the development of generative-AI systems. Thus, authors and their works are a part of the generative-AI supply chain, whether they would like to be or not (Figure 1, stage 1).

2.2 Data Creation

Original expressive works are distinct from their datafied counterparts.[5] Data examples are constructed to be computer-readable, such as the JPEG encoding of a photograph. For the most part, the transformation of creative content to data formats predates generative AI (Figure 1, stage 2), but all state-of-the-art generative-AI systems depend on it. Text-to-text generation models are trained on digitized text, text-to-image models are trained on both text and images, text-to-music models are trained on text and audio files, and so on. This is an important point because works that have been transformed into data have been fixed in a tangible medium of expression, and hence are subject to copyright.[6] In turn, generative-AI systems are often are trained on data that include copyrighted expression. The GitHub Copilot system involves models trained on copyrighted code,[7] ChatGPT's underlying models are trained on text scraped from the web, Stability AI's Stable Diffusion is trained on text and images, and so on. For the most part, it is the copyright owners of these datafied individual works who are the potential plaintiffs in an infringement suit against actors at other stages of the supply chain (Section 3).

2.3 Dataset Collection and Curation

The training process for cutting-edge generative-AI models requires vast quantities of data. Dataset creators often meet this need by scraping the Internet.[8] This process involves numerous curatorial choices, including filtering out material that creators do not want to include, such as "toxic speech" [Lee et al. 2023].[9] Dataset creators are also necessarily curators.

With respect to the generative-AI supply chain, there are several points worth highlighting (Figure 1, stage 3). First, while dataset creation and curation can be carried out by the same entities that train generative-AI models, it is common for them to be split across different actors. The Stable Diffusion model, for example, is trained on images from datasets curated by the non-profit organization LAION.[10] It is necessary, therefore, to consider the liability of dataset creators separately from the liability of model trainers.

Second, dataset curation will frequently involve "the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship" [Copyright Law of the United States 2010a]. Thus, training datasets can themselves be copyrighted; copying of the dataset *as a whole* without permission could constitute infringement, separate and apart from infringement on the underlying works.[11]

Third, while a few datasets include metadata on the provenance of their data examples, many do not. Provenance makes it easier to answer questions about the sources a model was trained on, which can be relevant to infringement analysis. It also bears on the ease with which specific material can be located, and if necessary removed, from a dataset. However, the use of web-scraping to collect generative-AI training datasets makes provenance difficult to track [Lee et al. 2023]. Even if a dataset creator releases the dataset itself under a license, this does not guarantee that the works in the

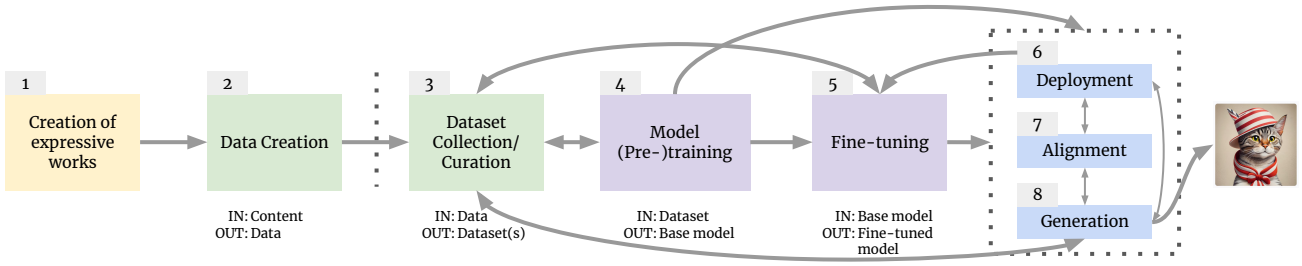

Figure 1: The generative-AI supply chain. We map out eight stages: 1) creation of expressive works, 2) data creation, 3) dataset collection/curation, 4) model (pre-)training, 5) model fine-tuning, 6) system deployment, 7) generation, and 8) model alignment. The creation of expressive works and data creation pre-date the advent of today's generative-AI systems (dotted line). There are many possible ways to connect the other six stages. Deployment, model alignment, and generation tend to happen in concert (dotted box). Generations can be used as training data (arrow from generation (7) to dataset collection/curation (3)). In this case, generation serves simultaneously as the creation of expressive works (1) and data creation (2). Curated data examples can be used for retrieval-augmented generation (arrow from dataset collection/curation (3) to generation (7)). APIs in deployed service can be used to do custom fine-tuning (arrow from deployment (6) to fine-tuning (5)).

dataset are appropriately licensed,[12] as is currently up for debate with the LAION-5B dataset [Anderson v. Stability AI, Ltd. 2023; Beaumont 2022; Schuhmann et al. 2022].[13]

2.4 Model (Pre-)Training

Following the collection and curation of training datasets, it is possible to train a generative-AI model. A model trainer[14] (Figure 1, stage 4) selects a training dataset, a model architecture (i.e., a set of initialized model parameters), a training algorithm, and a seed value for the random choices made during the training.[15] The process of transforming these inputs into a trained model is expensive. It requires a substantial investment of multiple resources: time, data storage, and compute. For example, BLOOM (a 176-billion-parameter open-source model from HuggingFace) was trained for 3.5 months, on 1.6 terabytes of text, using 384 GPUs [Bekman 2022; Workshop et al. 2023]; it cost an estimated $2-5 million.[16] As another point of reference, MosaicML has trained a GPT-3-quality model for less than $0.5 million.[17] Altogether, the dollar cost can range from six to eight figures.[18]

The output of the training process is typically called a **pre-trained model** or **base model**.[19] A base model has many possible futures. It could sit idly in memory, collecting figurative dust.[20] The model could be uploaded to a public server,[21] allowing others to download it and use it however they want.[22] The model could be integrated into a system and deployed as a public-facing application (Section 2.6), which others could use directly to produce generations (Section 2.7). Or, the model could be further modified by the initial model trainer, by another actor at the same organization, or, if made publicly available, a different actor from a different organization. That is, another actor could take the model parameters and use them as the input to do additional training with new or modified data. This possibility of future further training of a base model is why this stage of the supply chain is most often referred to as ***pre*-training**, and why a base model is similarly often called a **pre-trained model**. Such additional training of the base model is called **fine-tuning**.

2.5 Model Fine-Tuning

Base models trained on large-scale, web-scraped datasets are not typically optimized to apply specialized domains of knowledge. For example, an English text-to-text base model may be able to capture general English-language semantics, but not able to reliably apply detailed scientific information about molecular biology.

This is where fine-tuning comes in (Figure 1, stage 5). Fine-tuning is the process of modifying a preexisting model and making it better along some dimension of interest. This process often involves training on additional data that is more aligned with the specific goals.[23] If we think of training as transforming data into a model, fine-tuning transforms a model into another model. Fine-tuning essentially involves just running more training. However, fine-tuning and pre-training may use different inputs, which ultimately makes the trajectories and outputs of their respective training processes very different. [24] To add more precision to our previous statement: fine-tuning transforms a model into another model, while incorporating more data.

Forks in the supply chain. Two important observations follow from our description of fine-tuning as (effectively) just performing more training. For one, a model trainer does not have to fine-tune at all. Prior to fine-tuning, there is a fork in the generative-AI supply chain with respect to the possible futures of the base model after pre-training (stage 4): One could take the output base model from pre-training, and use this model directly as the input for system deployment (stage 6), generation (stage 7), or model alignment (stage 8). Alternatively, it is possible to perform multiple separate passes of fine-tuning — to take an already-fine-tuned model, and use it as the input for another run of fine-tuning on another dataset.[25]

For each possibility, there can be different actors involved. Sometimes, the creator of a model also fine-tunes it. Google's Codey models (for code generation) are fine-tuned versions of Google's PaLM 2 model [Google 2023]. In other cases, when a model's weights are publicly released (as Meta has done with its Llama family of models) [News 2023; Touvron et al. 2023a,b], others can take the

model and independently fine-tune them for particular applications. A Llama fine-tuner could release their model publicly, which in turn could be fine-tuned by another party.[26] To use a copyright analogy, a fine-tuned model is a derivative of the model from which it was fine-tuned; a repeatedly fine-tuned model is a derivative of the (chain of) fine-tuned model(s) from which it was fine-tuned.

It is helpful to make the base-/fine-tuned model distinction because different parties may have different knowledge of, control over, and intentions toward choices like which data is used for training and how the resulting trained model will, in turn, be put to use. A base-model creator, for example, may attempt to train the model to avoid generating copyright-infringing material. However, if that model is publicly released, someone else may attempt to fine-tune the model to remove these anti-infringement guardrails. A full copyright analysis may require treating them differently and analyzing their conduct in relation to each other (Section 3.4).

2.6 Model Release and System Deployment

It is possible to release a model or deploy it as part of a larger software system, use the model to produce generations (Section 2.7), or to take the model and further alter or refine it via model alignment techniques (Section 2.8). In brief, there is a complicated interrelationship between the deployment, generation, and alignment stages. They can happen in different orders, in different combinations, and at different times for different generative-AI systems. For purely expository purposes, we present them one at a time, starting with **model release** and **system deployment** (Figure 1, stage 6).

A model is open-source **released** when its model parameters are uploaded to a server or platform (like HuggingFace [HuggingFace 2023]), from which others can download it.[27] Released models, which include Meta's Llama family of models [News 2023; Touvron et al. 2023a,b] and Stable Diffusion [Rombach et al. 2022] give others direct access to their parameters. Developers can write their own code to produce generations, or alter the model through fine-tuning or model alignment (Section 2.8).

In contrast, closed-source models are not directly available to external users. They are typically embedded in large, complex software systems, which are **deployed** to both internal and external users through software services. For example, a model could be hosted by a company (e.g., OpenAI, Stability AI, or Google). It could be used internally to support various services (e.g., Google has integrated an internally-developed LLM into Google Search), or released as a hosted service that gives external users access to generative-AI functionality.

External-facing services can be deployed in a variety of forms, and *do not* typically include the ability to change the model's parameters. They can be browser-based user applications (e.g., ChatGPT, Midjourney, DreamStudio), or public (but not necessarily free) APIs for developers (e.g., GPT models, Cohere).[28] Some model trainers provide a combination of release and deployment options. For example, DreamStudio is a web-based user interface [DreamStudio 2023] built on top of services hosted by Stability AI [AI 2023b]; the DreamStudio application gives external users access to a generative-AI system that contains the open-source Stable Diffusion model [Rombach et al. 2022], which Stability AI also makes available for direct download.[29]

This is a familiar spectrum from Internet law, from cloud-hosted services at one end to fully open-source software at the other, with closed-source apps in between. These deployment methods offer varying degrees of customization and control on the part of the deployer and the user. For example, a generative-AI system deployed as a service will often modify the user-supplied prompt before inputting it to the model. Several applications (e.g., ChatGPT, Bard, and Sydney), add additional instructions ("application prompts") to the user's input to create a compound prompt [OpenAI 2023b; Zhang and Ippolito 2023].[30] The additional instructions change the behavior of the model's output on a user prompt.[31] For example, compare the following two application prompts: "I want you to act as an English translator, spelling corrector and improver…" and "I want you to act as a poet. You will create poems that evoke emotions and have the power to stir people's soul…" [Akın 2023].[32]

Typically, model trainers and owners maintain the most control over models deployed through hosted services and the least over models released as model parameters [Vincent 2023]. By embedding a model within a larger system,[33] they can imbue it with additional behaviors [Cooper et al. 2021a]. For example, APIs and web applications allow deployers to filters a model's inputs or outputs. For example, ChatGPT will often respond with some version of: "I'm really sorry, but I cannot assist you with that request," when its "safety" filters are tripped.[34] GitHub Copilot expressly states that it uses "filters to block offensive words in the prompts and avoid producing suggestions in sensitive contexts" [GitHub 2023a]. Additionally, some services include output filters to avoid generating anything that looks too similar to a training example [GitHub 2023b].[35] Unfortunately, output filtering is an imperfect process.(See Section 3.3).[36]

2.7 Generation

Generative-AI models produce output generations in response to input prompts.[37] While a few users produce generations from open-source models by writing code to interact with the model parameters to execute the generation process,[38] most users a interact with models only indirectly, through an API, web service, or application.

Users can affect generations in a few ways. First, there is the *prompt itself*. Some prompts, like `"a big dog"`, are simple and generic. Others, such as `"a big dog facing left wearing a spacesuit in a bleak lunar landscape with the earth rising as an oil painting in the style of Paul Cezanne"`, are more detailed. Second, there is the *choice* of which deployed system to use (which embeds an implicit choice of model). For example, a user that wants to perform text-to-image generation on a browser-based interface needs to select between Ideogram, DALL·E-2, Midjourney, and other publicly available text-to-image applications that could perform this task. A user typically selects an application with the outputs partially in mind, so that one choice or another can indicate an attitude towards the possibility of infringement. Users may also revise their prompt to attempt to create generations that more closely align with their goals. And, third, there is *randomness* in each generation.[39] It is typical, for example, for image applications to produce several candidate generations. DALL·E-2, Midjourney, and Ideogram all do this.

As we will see, characterizing the relationship between the user and the chosen deployed system is one of the critical choice points

in a copyright-infringement analysis. There are at least three ways the relationship could be described:[40]

- The user actively drives the generation through choice of prompt, and the system passively responds. In this view, the user is potentially a direct infringer, but the application is like a web host, ISP, or other neutral technological provider.
- The system is active and the user passive. In this view, the user is like a viewer of an infringing broadcast, or the unwitting buyer of a pirated copy of a book. Primary copyright responsibility lies with the deployed system, and possibly with others further upstream in the generative-AI supply chain.
- The user and system are active partners in generating infringing outputs. In this view, the user is like a patron who commissions a copy of a painting; the system is like the artist who executes it. They have a shared goal of creating an infringing work.

We will argue that there is no universally correct characterization. Which of these three is the best fit for a particular act of generation will depend on the system, the prompt, how the system is marketed, and how users can interact with the system's interfaces.[41]

Forks in the supply chain. There is a loop from generation back to the beginning of the supply chain. While not the most common contemporary practice, it is possible to use generations as training data for generative-AI models.[42] In this case, generation serves simultaneously as the creation of expressive works (i.e., stage 1) and data creation (i.e., stage 2) and generations can become inputs to dataset collection and curation processes (i.e., stage 3), which we indicate with an arrow in Figure 1. As we discuss in Section 3, this potential circularity also has implications for copyright.[43]

Alternatively, for the process of generation, some generative-AI systems interact with *external* deployed services, as is the case with ChatGPT plugins [OpenAI 2023a]. Such interactions between external services and generation further complicate the generative-AI supply chain that we depict in Figure 1. In particular, by potentially integrating with other systems, the generation stage could implicate an entirely separate, unspecified number of supply chains consisting of entirely different organizations and actors. This, too, raises important copyright implications (what if news articles or short stories are integrated by the plugin?).

2.8 Model Alignment

The generative-AI supply chain does not stop with generation. As discussed above, model trainers try to improve models during both pre-training and fine-tuning. For pre-training, they monitor evaluation metrics, and may pause or restart the process to alter the datasets and algorithm used (Section 2.4); for fine-tuning, they continue training the base model with data that is specifically relevant for a particular task (Section 2.5). Both of these base model modifications are coarse: They make adjustments to the dataset and algorithm, and do not explicitly incorporate information into the model about whether specific generations are "good" or "bad," according to user preferences.[44]

There is a whole area of research, called **model alignment**, that attempts to meet this need [Lowe and Leike 2023].[45] The overarching aim of model alignment is to *align* model outputs with specific generation preferences (see Figure 1, stage 8). Currently, the most popular alignment technique is called **reinforcement learning with human feedback (RLHF)** [Christiano et al. 2023; Ouyang et al. 2022]. As the name suggests, RLHF combines collected human feedback data with a (reinforcement learning) algorithm in order to update the model. Human feedback data can take a variety of forms, which include user ratings of generations. For example, such ratings can be collected by including thumbs-up and thumbs-down buttons in the application user interface, which are intended to query feedback about the system's output generation. In turn, the reinforcement learning algorithm uses these ratings to adjust the model — to encourage more "thumbs-up" generations and fewer "thumbs-down" ones.[46] Future training and alignment on the model may include both the inputted prompt and the generation in addition to the feedback provided. As discussed in the prior section, user-supplied prompts may include copyrighted content created by either the user themselves or by another party. Most generative-AI companies begin model alignment prior to deployment or release (Section 2.6). In this respect, model alignment complements other techniques, like input-prompt and output-generation filtering (Section 2.7)[47]

3 COPYRIGHT AND THE SUPPLY CHAIN

The hornbook statement of United States copyright doctrine is that original works of authorship are protected by copyright when they are fixed in a tangible medium of expression. A defendant directly infringes when they engage in conduct implicating one of several enumerated exclusive rights (reproducing, publicly distributing, etc.), with a work of their own that is substantially similar to a copyrighted work because it was copied from that work. Other parties may be held secondarily liable for conduct that bears a sufficiently close nexus to the infringement under one of several theories. Otherwise infringing conduct is legal when it is protected by one of several defenses, including the DMCA Section 512 safe harbors, fair use, or an express or implied license.

In this section, we first provide some brief background on what kinds of works copyright applies to (Section 3.1). We then apply aspects of the above orthodox, uncontested statement of copyright law to the generative-AI supply chain. We address issues of rights (Section 3.2), infringement (Sections 3.3 & 3.4), and fair use (Section 3.5). We defer discussion of safe harbors, licenses, paracopyright liability, and remedies to other work. Our goal is to be careful and systematic, not to say anything dramatically new.

3.1 What is copyrightable?

Copyright protects "(1) original works of authorship (2) fixed in any tangible medium of expression" [Copyright Law of the United States 1990].[48] "Original, as the term is used in copyright, means only that the work was independently created by the author (as opposed to copied from other works), and that it possesses at least some minimal degree of creativity" [Feist Publications v. Rural Telephone Service Company 1991, p. 345] Fixation is satisfied when the work is embodied in a tangible object in a way that is "sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration" [Copyright Law of the United States 2010a].[49]

We start with fixation. Unfixed works have no interaction with the generative-AI supply chain. A work must be fixed to be used as training data. Truly ephemeral creations, like unobserved dances

and songs that are never recorded, will never be captured in a way that can be used as an input to a training algorithm. Datasets, models, applications, prompts, and generations are all fixed in computers and storage devices. Once it is fixed, however, any kind of original expression can be used as an input for generative AI.

The originality requirement distinguishes material that was created by a human author from facts that "do not owe their origin to an act of authorship" [Feist Publications v. Rural Telephone Service Company 1991, p. 347]. In addition, some types of material are never copyrightable, including any "idea, procedure, process, system, method of operation, concept, [or] principle"[50] In practice, this means that the copyright in some works (e.g., product photographs) will be "thinner" and protect fewer aspects of the works than the "thicker" copyrights in others (e.g, abstract art), because the "range of creative choices that can be made in producing the works is narrow" [Rentmeester v. Nike, Inc. 2018, p. 1120]. In particular, any copyright in computer software — which is treated as a "literary work" for copyright purposes — typically excludes a great deal of functional material, such as coding conventions required by the choice of programming language [Samuelson 2016]. As a result, some individual training examples are uncopyrightable. (For example, birdsong-recognition AIs are trained on recordings of birds [Kahl et al. 2021; Naruto v. Slater 2018].[51]) But other items are copyrightable, and those copyrights will be held by a variety of authors.

Training datasets will include different amounts and proportions of copyrighted material. A dataset of birdsong recordings will be almost entirely copyright-free, but a dataset of illustrations will contain numerous copyrighted works. Further, datasets *themselves* may be copyrightable as **compilations** [Copyright Law of the United States 1976],[52] "formed by the collection and assembling of preexisting materials or of data" [Copyright Law of the United States 2010a].[53] A compilation is copyrightable as such when it features a sufficiently original "selection or arrangement" [Feist Publications v. Rural Telephone Service Company 1991, p. 348]. Originality in selection is choosing *what to include* in the dataset; originality in arrangement is choosing *how to organize* it.

Generations raise a doctrinal question that has been debated for decades: who, if anyone, owns the copyright in the output of a computer program [Samuelson 1985]? Although some have argued that the program itself should be regarded as the author, computer authorship is squarely foreclosed by U.S. copyright law [Grimmelmann 2016]. So far, the courts have held firm to this line for AI generations. *Thaler* [Thaler v. Perlmutter 2023] upheld the Copyright Office's refusal to register copyright in an image allegedly "autonomously created by a computer algorithm running on a machine." The Copyright Office had held that the image lacked human authorship, and the court agreed.[54] The author of a generation — if anyone — is some human connected to the generation. The four immediately relevant possibilities are (1) author(s) whose works the model was trained on, (2) some entity in the generative-AI supply chain (e.g., the model trainer or fine-tuner; application developer), (3) the user who prompted a service for the specific generation, or (4) no one. As between these four possibilities, there is no one-size-fits-all answer. All four arise in actual generative-AI applications.

TODO: Cooper up to here

3.2 The Exclusive Rights

Copyright includes five relevant exclusive rights: reproduction, adaptation, public distribution, public performance, and public display.[55] Every stage in the generative-AI supply chain requires a reproduction and thus potentially implicates copyright. Because the remedies for infringement of a work are the same, regardless of whether the defendant violated one exclusive right or several, the precise dividing lines are often unimportant. We examine the adaptation right, and defer additional discussion to other work.

The adaptation right gives the copyright owner the exclusive right to "to prepare derivative works based upon the copyrighted work."[56] A derivative work combines the authorship in an existing (or "underlying") work with new authorship. In a compilation (Section 3.1), the underlying works are present in substantially unmodified form, whereas in a derivative work the underlying work is "recast, transformed, or adapted."[57]The adaptation right makes clear that copyright extends beyond literal similarity to incorporate changes of form, genre, and content such as translations, sequels, and film adaptations [Gervais 2013, 2022; Samuelson 2013].

A training dataset is probably not a derivative work of any of the works in it; it is more appropriately classified as a compilation "formed by the collection and assembling of preexisting materials" [Copyright Law of the United States 2010a]. To the extent that a model is similar to a work it was trained on, it is a derivative work because it is "based on" its training data. (Section 3.3). Similarly, a prompt could be a reproduction or derivative of an existing work (as when a diffusion model is prompted with an image to infill) [Anthropic 2023]. And generations are frequently derivative works of works in the training data or prompts, again subject to similarity.

3.3 Substantial Similarity

Substantial similarity is a qualitative, factual, and frustrating question. Two works are substantially similar when "the ordinary observer, unless he set out to detect the disparities, would be disposed to overlook them, and regard their aesthetic appeal as the same" [Peter Pan Fabrics, Inc. v. Martin Weiner Corp. 1960, p. 489]. A common test is a "holistic, subjective comparison of the works to determine whether they are substantially similar in total concept and feel" [Rentmeester v. Nike, Inc. 2018, p. 1118]. This is not a standard that can be reduced to a simple formula that can easily be applied across different works and genres.[58] We discuss base models and generations below, and defer discussion of data, datasets, fine-tuned models, aligned models, and deployed services to other work.

3.3.1 Pre-Trained/Base Models. A model is different in kind from the copyrightable works it was trained on. No viewer would say that the model has the same "total concept and feel" as a painting; no reader would say that it is substantially similar to a blog post; and so on. That said, the Copyright Act does not require that copies be directly human-intelligible to infringe. A Blu-Ray is not directly intelligible by humans, either, but it counts as a "copy" of the movie on it. Indeed, all digital copies are unintelligible. Instead, they are objects "from which the work can be perceived, reproduced, or otherwise communicated … *with the aid of a machine or device*" [Copyright Law of the United States 2010a]. Thus, even if a model is uninterpretable, it might still be possible to "perceive[]"

(a) `"an adventurous archaeologist with a whip and a fedora"` (b) `"ice princess"`

Figure 2: Generated by the authors using Midjourney.

or "reproduce[]" a copyrighted work embedded in its parameters through suitable prompting. Indeed, there is substantial evidence that many models have memorized copyrighted materials [Carlini et al. 2023a, 2021].[59] For example, Carlini et al. [2023a] shows how Stable Diffusion has memorized photographs.

A model might memorize more works or fewer [Carlini et al. 2023a,b]. But at least some models memorize at least some works closely enough to pass the substantial-similarity test. On this view, a model is a substantially similar copy of a work when the model is capable of generating the work.[60] Note that this is direct infringement, not secondary (Section 3.4). The theory is not that the generation is an infringing copy, and that the model is a tool in causing that infringement in the way that a tape-duplicating machine might be a tool in making infringing cassettes [A & M Records, Inc. v. Abdallah 1996]. Rather, the theory is that the model itself is an infringing copy, regardless of whether that particular generation is ever made.[61]

3.3.2 Generations. Some generations are nearly identical to a work in the model's training data (i.e., memorized). They are substantialy similar to that work. Other generations are very dissimilar from every work in the training data. There is no substantial similarity, because infringement is assessed on a work-by-work basis. Although it is in some sense based on all of the works in the training dataset, it does not infringe on any of them.[62] The hardest case is when an output is similar to a work in the training data in some ways, but dissimilar from it in other ways. This case is likely to arise in practice precisely because it lies in between the two extremes of memorized generations and original generations. Somewhere between them lies the murky frontier between infringing and non-infringing.

It is hard to make sweeping statements because of the factual intensity and aesthetic subjectivity of similarity judgments.[63] Whether a particular generation is substantially similar is ultimately a jury question requiring assessment of audiences' subjective responses to the works. Generative AI will produce cases requiring this lay assessment; it is impossible to anticipate in advance how lay juries will react to all of the possible variations. So, we will assume that lay audiences would say that some generations will infringe, but that it will not be possible to perfectly predict which ones.[64]

Even if complete answers are impossible, there are some interesting questions worth considering. As Matthew Sag observes [Sag 2023], certain characters are so common in training datasets that models have "a latent concept [of them] that is readily identifiable and easily extracted." For example, prompting Midjourney and Stable Diffusion with `"snoopy"` produces recognizable images of Snoopy the cartoon beagle. Characters are a special case in copyright; some cases relax the rule that infringement is measured on a work-by-work basis, instead measuring the similarity of the defendant's character to one who appears in multiple works owned by the plaintiff.[65] But the "Snoopy effect" is not confined to characters. Some works are simply so prevalent in training datasets that models memorize them. As an uncopyrighted example, Van Gogh's *Starry Night* is easy to replicate using Midjourney; Sag's paper includes a replication of Banksy's *Girl with Balloon*. This looks like substantial similarity.

A variation of the Snoopy effect arises when a model learns an artist's recognizable *style*. ChatGPT can be prompted to write rhyming technical directions in the style of Dr. Seuss; DALL·E-2 can be prompted to generate photorealistic portraits of nonexistent people in the style of Dorothea Lange [Casper et al. 2023]. As with characters, these outputs have similarities that span a body of source works, even if they are not close to any one source work. The proper doctrinal treatment of style is a difficult question [Sobel 2023]. The Snoopy effect can also be triggered without explicit prompting. The archaeologist generated in Figure 2a features a dark-haired male character with stubble, wearing a brown jacket and white shirt, with a pouch slung across his shoulder. These are features associated with Indiana Jones, but neither the features nor `"indiana jones"` appear in the prompt. Some caselaw holds such similarities are enough for infringement when the character is iconic enough [Metro-Goldwyn-Mayer v. American Honda Motor Co. 1995].[66]

Other copyright doctrines, however, may limit infringement in Snoopy-effect cases. One of them is *scènes à faire*: creative elements that are common in a genre cannot serve as the basis of infringement. For example, [Walker v. Time Life Films, Inc. 1986, p. 50] explains that "drunks, prostitutes, vermin and derelict cars would appear in any realistic work about the work of policemen in the South Bronx." Similarly, prompting Midjourney with `"ice princess"` produces portraits in shades of blue and white with flowing hair and ice crystals. (Figure 2b) Similarities to Elsa from *Frozen* arise simply because these are standard tropes of wintry glamour. Some of them may now be tropes *because* of the *Frozen* movies, but they are still uncopyrightable ideas, rather than protectable expression.[67]

To close this section, we note that not all similarity is infringing. Some similarities arise for innocent reasons. The defendant and the plaintiff might both have copied from a common predecessor work, and resemble each other because they both resemble the work they were based on. The similarities might consist entirely of accurate depictions of the same preexisting thing, like Grand Central Station at midday, and resemble each other because Grand Central Station resembles itself. The similarities might be purely coincidental. The plaintiff might even have copied from the defendant!

Copyright law therefore requires that the plaintiff prove that the defendant copied from their work, rather than basing it on some other source or creating it anew, an inquiry known as "copying in fact." This is a factual question. In some cases, there is direct evidence: e.g., the defendant admits copying or there is video of the defendant using tracing paper to copy a drawing. But in many cases, there are two kinds of indirect evidence: proof that the defendant

had *access* to the plaintiff's work, and examples of "probative" *similarities* in the works themselves. Access shows that copying was possible, and similarities can rebut alternative innocent theories.[68]

3.4 Direct Infringement

We next discuss direct infringement and generations. We defer other supply-chain stages and analysis of indirect infringement to other work. Direct copyright liability has no mental element: it is "strict liability." All that is required is that they intentionally made the infringing copy. George Harrison's 1970 "My Sweet Lord" has the same melody and harmonic structure as the Chiffon's 1962 "He's so Fine"; the court held that "his subconscious knew it already had worked in a song his conscious mind did not remember," and found him liable for infringement [ABKCO Music, Inc. v. Harrisongs Music, Ltd. 1983, p. 180].

But direct copyright does have an element of "volitional conduct" [CoStar Group, Inc. v. LoopNet, Inc. 2004]. Its purpose is to decide whether a defendant should be analyzed as a direct or indirect infringer.[69] Some courts have described the test in terms of causation: "who made this copy?"[70] The direct infringer is the party whose actions toward a specific item of content most proximately caused the infringing activity; anyone else is (potentially) an indirect infringer. Thus, for example, a service that can be used to upload and download infringing content that a user chooses does not engage in volitional conduct [Perfect 10, Inc. v. Giganews, Inc. 2017], but a service that curates a hand-picked selection of infringing content for users to download does [Capitol Records, Inc. v. MP3tunes, LLC 2014].

The simplest case is where the same actor supplies both the model and the prompt.[71] Here, the subconscious-copying doctrine is a surprisingly good fit for AI generation. The model's internals are like the contents of George Harrison's brain: creatively effective, but not fully amenable to inspection. If I prompt an image model with `"'ice princess'"`, I have set in motion a process that may draw on copyrighted works in the same way that George Harrison drew on works he had heard. If that process generates Elsa, the resulting infringement is on me the same way that the infringement of "He's So Fine" was on Harrison. I could have taken greater care to check whether the image I was generating resembled a copyrighted work – just as George Harrison could have thought harder or asked more people whether the tune sounded familiar.

Matters are more complicated for generation services. Here, the question is whether the user and/or the provider should be treated as a direct infringer. There are at least three plausible answers, depending on the facts. First, the *user of the service* might be a direct infringer. If a user enters a prompt for `"'elsa and anna from frozen'"`, the provider resembles a copy shop that provides a general-purpose tool and let users choose what to do with it [Perfect 10, Inc. v. Giganews, Inc. 2017]. Second, the *service provider* might be a direct infringer. If a user types in `"'heroic princesses'"` and the model generates a picture of Elsa and Anna, the user has acted innocently and it is the model that has narrowed down the space of possible outputs to one that happens to be infringing. Third, *both* the user of the service and service provider might be treated as direct infringers. Suppose the user inputs `"'frozen 3 screenplay'"` to a service that has been trained on thousands of Hollywood screenplays. Both the user and the service have the necessary volition to create a work that is substantially similar to the *Frozen* movies.

It seems unlikely, however, that a court would treat both service and user as indirect infringers. This would violate the doctrinal requirement that there be a direct infringer for indirect liability to attach, and it would leave both potentially responsible parties free of liability. The choice between the other three cases is partly factual, and partly policy-driven. It is factual because there are clear paradigm cases in which the user of the service makes the choice for infringement, the service provider makes the choice for infringement, and the two conspire together to infringe. But it is policy-driven because, between these three poles, the identification of the direct infringer depends on which analogies one finds persuasive, and what one thinks copyright's goals are.[72]

3.5 Fair Use

Many stages of the generative-AI supply chain involve *prima facie* infringing reproductions, so copyright's all-purpose defense, fair use, will play a major role in making generative AI possible at all [Copyright Law of the United States 1992] Others have discussed the fair-use issues in detail [Henderson et al. 2023b; Murray 2023; Sag 2023; Sobel 2017]. It is highly case-specific, so we will focus on only a few salient points. We discuss generations, taking each of the four fair-use factors in turn, and defer other stages to other work.

Factor One ("the purpose and character of the use …" [Copyright Law of the United States 1992][73]): A use is transformative when "the quoted matter is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings" [Leval 1990, p. 1111]. The modification, remixing, and abstraction of input works literally involves exactly this kind of transformation. Some AI skeptics might deny that AI-generated material can be expressive. But as long as audiences find "new information, new aesthetics, new insights and understandings" in these generations, the goals of transformative use will be served.[74] Other generations will not be transformative. When a model outputs a memorized work, here is no transformation in content (Section 3.3). Other changes can also be non-transformative, e.g., memorized examples that are noisier than the source image. The noise is not new expression conveying new aesthetics. It is just noise. The rest of the first factor does not point one direction or the other. Generations can be put to commercial use (e.g., backgrounds for a music video) and to noncommercial use (e.g., illustrating an academic article on generative AI). Some outputs will be put to favored purposes like education and news reporting, while other outputs will be put to run-of-the-mill entertainment purposes.[75]

Factor Two ("the nature of the copyrighted work" [Copyright Law of the United States 1992][76]): This factor depends on the model in question. Some training data will be informational; some will be expressive. Most training data will have been "published" within the meaning of copyright law; otherwise, it would not be available as training data at all. A very small fraction of training data may be "unpublished" within the meaning of copyright law — i.e., it has been shared "(1) …only to a select group (2) for a limited purpose and (3) with no right of further distribution by the recipients" [Patry

2023, S. 6.31] — and included through express breach of confidence. Here, this factor will favor the plaintiff.

Factor Three ("the amount and substantiality of the portion used …" [Copyright Law of the United States 1992][77]): This factor, like substantial similarity, will not systematically favor either side. Some generations will closely resemble the works they were copied from; others will copy only small portions of the works.[78] Even for workse that are transformative, it still matters whether the generation copies more than necessary. A "painting of a car driving in a snowstorm in the style of Frida Kahlo" might copy just Kahlo's brushwork or floral motifs, or it might also imitate the entire composition of one of her self-portraits.

Factor Four ("the effect of the use upon the potential market for …the copyrighted work."[79]): The outputs of a non-generative AI do not compete in the market for a copyrighted work. These outputs could *reduce the demand* for the copyrighted work. For example, an AI-powered recommendation system might analyze the frames of a movie and assign it a low rating for visual interest. But the rating does not substitute for the movie in the market for movies. Viewers consume the rating to learn about movies, not to enjoy the expression in the rating. Any harm to the copyright owner is not fourth-factor harm [Campbell v. Acuff-Rose Music 1994]. The outputs of a generative-AI system, however, can substitute for a copyrighted work under the fourth factor. Consider the following variations on a theme:

- Instead of paying to obtain a copy of 'The Old Sugarman Place" episode of *Bojack Horseman*, a user prompts a generative-AI system to generate `"'The Old Sugarman Place'"`. It generates a close duplicate — essentially a pirated edition at a lower price. This is a paradigmatic fourth-factor harm.
- The user prompts a generative-AI system to generate `"'The Old Sugarman Place'"`, and it generates a non-exact copy with significant changes to the dialogue and animation. This episode, "The New Sugarman Place," is also a direct competitor for this user's business. It might be a better or worse competitor, depending on how closely "The New Sugarman Place" matches "The Old Sugarman Place." But this is still factor-four harm.
- The user prompts a generative-AI system to generate a new episode of *Bojack Horseman*. The generation does not necessarily compete with "The Old Sugarman Place" itself.[80] Instead, it competes with commissioning the writers, animators, and voice cast to create new episodes, or with paying for a license to make new episodes.[81] This is also factor-four harm to the market for licenses and authorized derivatives. For example, in Sid & Marty Krofft Television v. McDonald's Corp. [1977] McDonald's created advertisements in the unsettling style of the children's show *H.R. Pufnstuff*.
- An individual prompts a generative-AI system to produce a generation in a broad style, e.g., `"animated sitcom about depression"`. The output is a video with dialogue and animation that do not look much like *Bojack*. The output does not directly compete with "The Old Sugarman Place," or with any particular work or particular author. Instead, it competes with animated television in general. If the generative-AI system had not been available, the individual might have paid to watch *Bojack* or *Dr. Katz* or some other show. Many authors might view this as undercutting the market for their work. Here, the fourth factor is *not even relevant*, because the new video is not substantially similar to any existing work. If a human creative team made a new animated sitcom about depression, they would be celebrated for their creativity not sued for infringement.
- An individual prompts a generative-AI system to produce a generation in a broad style, e.g. `"animated sitcom about depression"`. The output, however, is "The Old Sugarman Place." The difference between this and the first case is that the user does not know about the work that the generation substitutes for. This is a factor-four harm. The generative-AI system has diverted the individual from potentially learning about and paying to watch "The Old Sugarman Place."

To summarize, factors one, three, and four can point strongly in favor of fair use or strongly against, depending on the context, and factor two does not consistently point in either direction. We conclude that some generations will be fair uses and others will not.

4 WHICH WAY FROM HERE?

The generative-AI supply chain is extremely complex. So is copyright law. Putting the two of them together multiplies the intricacy. Two unsettling conclusions follow. First, because of the complexity of the *supply chain*, it is not possible to make accurate sweeping statements about the copyright legality of generative AI. Too much depends on the details of specific systems. All the pieces matter, from the curatorial choices in the training dataset, to the training algorithm, to the deployment environment, to the prompt supplied by the user. Courts will have to work through these details in numerous lawsuits and develop doctrines to distinguish among different systems and uses. Second, because of the complexity of *copyright law*, there is enormous play in the joints. Substantial similarity, fair use, and other doctrinal areas all have open-ended tests that can reach different results depending on the facts a court emphasizes and the conclusions it draws. This complexity gives courts the flexibility to deal with variations in the supply chain. Paradoxically, it also gives courts the freedom to reach any of several different plausible conclusions about a generative-AI system. We explore some of the ways that courts might use their discretion to apply copyright law to generative AI (Section 4.1), and then discuss some of the considerations that courts should keep in mind (Section 4.2).

4.1 Possible Outcomes

There are a few boxes that courts may find it appealing to sort generative-AI systems into.

4.1.1 No Liability. First, courts might hold that neither services nor users are liable for copyright infringement. Under a combination of no substantial similarity and fair use, anything produced by a generative-AI system would be categorically legal. Models and services would also be legal because intermediate nonexpressive fair use would shield them. Training datasets would also usually be legal as well (except perhaps in cases of blatant infringement like Books3) [Kadrey v. Meta Platforms, Inc. 2023; Knibbs 2023; Reisner 2023]. They would be fair-use inputs to noninfringing downstream stages of the supply chain.

This regime is clear and simple. It would also be unstable. While this outcome might make sense for some generative-AI systems,

it seems both unworkable for systems trained specifically to emulate the styles of particular creators, and retrieval systems that reproduce matching works exactly [Borgeaud et al. 2022]. If all generative AI were categorically legal, then developers might start adding generative components to other systems in order to launder copyrighted works through them. The endpoint could be the effective collapse of copyright. Assuming that this is not an outcome that courts would willingly preside over, then, a blanket no-liability regime seems unlikely. Instead, courts would be more likely to find at least some infringement — so the question becomes where to draw the line.

4.1.2 Liability for Generations Only. Second, courts could draw a line between services and users. In this regime, only generations would be treated as infringing.[82] In this world, generative-AI systems would be creative tools like Photoshop.[83] The user would be responsible for making sure that anything they create with the tools is noninfringing, but the tools would be shielded under something like a strong *Sony* rule, assembled out of a combination of no substantial similarity, no indirect infringement, and/or fair use. This result might be unfair to users whose infringements resulted from systems producing generations that reproduce material in the underlying model's training dataset, through no choice or fault of their own. But this is arguably the same kind of situation that some courts currently countenance when they hold that users can be liable for embedding images from Instagram even though Instagram is not liable for hosting those images [Sinclair v. Ziff Davis, LLC 2020].

The main difficulty with this regime would be policing against systems designed specifically for infringement. Something like the *Grokster* [Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd. 2005] rule, carefully followed, might suffice. The providers of a service that was geared to produce infringing outputs could be held liable. So could the publishers or deployers of a model that had been trained or fine-tuned to optimize its effectiveness at infringement. So could the curator of a dataset that included only infringing works, or was intentionally organized to meet the needs of a model known to be intentionally trained for infringement. At every stage, a party would be held responsible only for its own actions directed towards increasing the use of a system for infringement.

4.1.3 Notice and Removal. Courts could treat generative-AI services as generally legal, but require them to respond to knowledge of specific infringements under a *Napster*-like rule [A & M Records, Inc. v. Napster, Inc. 2001]. One plausible route to this regime would be to treat infringing generations as creating direct liability for users and only indirect liability for service providers. Another would use fair use to shield service providers as long as they took reasonable overall precautions, including responding when they had sufficient knowledge of infringement. And a third would be to find liability but craft an injunction that only required services to act against infringement they were aware of.[84]

If courts end up recreating a notice-and-takedown regime, they would likely settle on familiar elements from the DMCA notice-and-takedown provision of Section 512: a way for copyright owners to give notice of infringement, block infringing generations on notice, block infringing generations on actual knowledge, block infringing generations on red-flag knowledge, avoid having a business model that directly ties income to infringement, and terminate the abilities of repeat infringers to continue making generations.

This is a very difficult technical problem. It would be much harder for a generative-AI system to implement than it is for a hosting platform to implement Section 512 compliance. The reason is that a notice directed to a hosting provider under Section 512(c) must include "Identification of the material that is claimed to be infringing . . . and information reasonably sufficient to permit the service provider to locate the material" [Copyright Law of the United States 2010b].[85] A valid notice is a roadmap; it tells the hosting provider exactly what to take down to comply. That material already exists, and the hosting provider can compare it to the copyrighted work to verify that they are substantially similar. But a notice to a generative-AI system is a notice against future generations, which may be different from each other and resemble the copyrighted work in different ways. Filtering for this kind of much more inexact match is much harder technically.[86] Further, there is no simple analogue for takedown in generative-AI models. Removing the influence of a particular example on a model is an active and unsolved area of research [Bourtoule et al. 2021; Meng et al. 2022].[87]

4.1.4 Infringing Models. A fourth possibility is that some or all generative-AI services are illegal because models themselves infringe. This outcome is an existential threat to model trainers and service providers; it makes their operations *per se* copyright infringement. It is also the outcome being sought by the class-action plaintiffs in high-profile lawsuits against OpenAI, Stability AI, and some of their partners. In this regime, the most important component of copyright law would become licensing. Models could only be trained on data that had been licensed from copyright owners; the terms under which those models and their generations could be used would have to be negotiated as part of the licensing agreement.[88]

4.2 Lessons

Having discussed what courts and policymakers could do, we now consider what they should do. In keeping with our bottom line — *the generative-AI supply chain is too complicated to make sweeping rules prematurely* — we offer a few general observations about the overall shape of copyright and generative AI that courts and policymakers should keep in mind as they proceed.

First, *copyright touches every part of the generative-AI supply chain.* Every stage from training data to alignment can make use of copyrighted works. Generative AI raises many other legal issues: Can a generative-AI system commit defamation [Bambauer and Surdeanu 2023; Brown 2023; Garon 2023; Henderson et al. 2023a; Volokh 2023]? Can a generative-AI system do legal work [Choi et al. 2023] and should they be allowed to [Mata v. Avianca 2023]? But these issues pertain to outputs of a generative-AI system–copyright pervades every step of the process.

Second, *copyright concerns cannot be localized* to a single link in the supply chain. Decisions made by one actor can affect the copyright liability of another actor far away in the supply chain. Whether an output looks like Snoopy or like a generic beagle depends on what images were collected in a dataset, which model architecture and training algorithms are used, how trained models are fine-tuned and aligned, how models are embedded in deployed

services, what the user prompts with, etc. Every single one of these steps could be under the control of a different person.

Third, *design choices matter*. There are obvious choices about copyright, like whether to train on unlicensed data (which can affect downstream risks), and how to respond to notices that a system is producing infringing outputs (which can affect upstream risks). But subtler architectural choices matter, too. Different settings on a training algorithm can affect how much the resulting model will memorize specific works. Different deployment environments can affect whether users have enough control over a prompt to steer a system towards infringing outputs. Copyright law will have to engage with these choices — as will AI policy.

Fourth, *fair use is not a silver bullet*. For a time, it seemed that training and using AI models would often constitute fair use. In such a world, AI development is generally a low-risk activity, at least from a copyright perspective. Yes, training datasets and models and systems may all include large quantities of copyrighted works — but they will never be shown to users. Generative AI scrambles this assumption. The serious possibility that some generations will infringe means that the fair-use analysis at every previous stage of the supply chain is up for grabs again.

Fifth, *the ordinary business of copyright law still matters*. Courts will need to make old-fashioned, retail judgments about individual works — e.g., how much does this image resemble Elsa in particular, rather than tropes of fantasy princesses? Courts *must* leave themselves room to continue making these retail judgments on a case-by-case basis, responding to the specific facts before them, just as they always have. Perhaps eventually as society comes to understand what uses generative AI can be put to and with what consequences, it will reconsider the very fundamentals of copyright law. But until that day, we must live with the copyright system we have. And that system cannot function unless courts are able to say that some generative-AI systems and generations infringe, and others do not.

Sixth, *analogies can be misleading*. There are plenty of analogies for generative AI ready to hand. A generative-AI model or system is like a search engine, or like a website, or like a library, or like an author, or like any number of other people and things that copyright has a well-developed framework for dealing with. These analogies are useful, but we wish to warn against treating any of them as definitive. As we have seen, generative AI is and can consist of many things. It is also literally a generative technology: it can be put to an amazingly wide variety of uses [Zittrain 2008]. And one of the things about generative technologies is that they cause convergence [Narechania 2022]. precisely because they can emulate many other technologies, they blur the boundaries between things that were formerly distinct. Generative AI can be like a search engine, and also like a website, a library, an author, and so on. Prematurely accepting one of these analogies to the exclusion of the others would mean ignoring numerous relevant similarities — precisely the opposite of what good analogical reasoning is supposed to do.

5 CONCLUSION

Our conclusion is simple. "Does generative AI infringe copyright?" is not a question that has a yes-or-no answer. There is currently no blanket rule that determines which participants in the generative-AI supply chain are copyright infringers. The underlying systems are too diverse to be treated identically, and copyright law has too many open decision points to provide clear answers. Copyright is not the only, or the best, or the most important way of confronting the policy challenges that generative AI poses. But copyright is here, and it is asking good questions about how generative-AI systems are created, how they work, how they are used, and how they are updated. These questions deserve good answers, or failing that, the best answers our copyright system is equipped to give.

REFERENCES

A & M Records, Inc. v. Abdallah 1996. 948 F. Supp. 1449 (C.D. Cal. 1996).

A & M Records, Inc. v. Napster, Inc. 2001. 239 F.3d 1004 (9th Cir. 2001).

ABKCO Music, Inc. v. Harrisongs Music, Ltd. 1983. 722 F.2d 988 (2d Cir. 1983).

Adobe. 2023. Experience the Future of Photoshop With Generative Fill. https://helpx.adobe.com/photoshop/using/generative-fill.html

Meta AI. 2023a. Use Policy. https://ai.meta.com/llama/use-policy/

Stability AI. 2023b. Stable Diffusion XL. https://stability.ai/stablediffusion

Fatih Kadir Akın. 2023. Awesome ChatGPT Prompts. *GitHub* (2023). https://github.com/f/awesome-chatgpt-prompts

American Broadcasting v. Aereo 2014. 134 S. Ct. 2498 (2014).

Anderson v. Stability AI, Ltd. 2023. No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023).

Anthropic. 2023. Introducing 100K Context Windows. https://www.anthropic.com/index/100k-context-windows/

Associated Press v. Meltwater U.S. Holdings, Inc. 2013. 931 F. Supp. 2d 537 (S.D.N.Y. 2013).

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, et al. 2022. Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073 [cs.CL]

Derek Bambauer and Mihai Surdeanu. 2023. Authorbots. *Journal of Free Speech Law* 3 (2023), 375.

Romain Beaumont. 2022. LAION-5B: A New Era of Large-Scale Multi-Modal Datasets. *LAION Blog* (31 March 2022). https://laion.ai/blog/laion-5b/

Stas Bekman. 2022. The Technology Behind BLOOM Training. *HuggingFace* (14 July 2022). https://huggingface.co/blog/bloom-megatron-deepspeed

Stella Biderman, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the Pile. arXiv:2201.07311 [cs.CL]

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, et al. 2022. Improving Language Models by Retrieving from Trillions of Tokens. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 2206–2240.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine Unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*. 141–159.

Nina Brown. 2023. Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation. *Journal of Free Speech Law* 3 (2023), 389.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]

Campbell v. Acuff-Rose Music 1994. 510 U.S. 569 (1994).

Capitol Records, Inc. v. MP3tunes, LLC 2014. 48 F.Supp.3d 703 (S.D.N.Y.2014).

Cariou v. Prince 2013. 714 F.3d 694 (2d Cir. 2013).

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023a. Extracting Training Data from Diffusion Models. arXiv:2301.13188 [cs.CR]

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. 2023b. Quantifying Memorization Across Neural Language Models. In *International Conference on Learning Representations*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2633–2650.

Cartoon Network LP, LLLP v. CSC Holdings, Inc. 2008. 536 F.3d 121 (2d Cir. 2008).

Stephen Casper, Zifan Guo, Shreya Mogulothu, Zachary Marinov, Chinmay Deshpande, Rui-Jie Yew, Zheng Dai, and Dylan Hadfield-Menell. 2023. Measuring the Success of Diffusion Models at Imitating Human Artists. arXiv:2307.04028 [cs.CV]

Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. https://arxiv.org/abs/2305.00118

Jonathan H. Choi, Kristen E. Hickman, Amy Monahan, and Daniel Schwarcz. 2023. ChatGPT Goes to Law School. *Journal of Legal Education* (2023). Forthcoming.

Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2023. Deep reinforcement learning from human preferences. arXiv:1706.03741 [stat.ML]

Samantha Cole. 2023. 'Life or Death:' AI-Generated Mushroom Foraging Books Are All Over Amazon. *404 Media* (29 Aug. 2023). https://www.404media.co/ai-generated-mushroom-foraging-books-amazon/

A. Feder Cooper, Karen Levy, and Christopher De Sa. 2021a. Accuracy-Efficiency Trade-Offs and Accountability in Distributed ML Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. Association for Computing Machinery, New York, NY, USA, Article 4, 11 pages. https://doi.org/10.1145/3465416.3483289

A. Feder Cooper, Yucheng Lu, Jessica Zosa Forde, and Christopher De Sa. 2021b. Hyperparameter Optimization Is Deceiving Us, and How to Stop It. In *Advances in Neural Information Processing Systems*, Vol. 34.

A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 864–876. https://doi.org/10.1145/3531146.3533150

Copyright Law of the United States 1976. https://www.law.cornell.edu/uscode/text/17/103 U.S.C. 17, 103.

Copyright Law of the United States 1990. https://www.law.cornell.edu/uscode/text/17/102 U.S.C. 17, 102.

Copyright Law of the United States 1992. https://www.law.cornell.edu/uscode/text/17/107 U.S.C. 17, 107.

Copyright Law of the United States 2010a. https://www.law.cornell.edu/uscode/text/17/101 U.S.C. 17, 101.

Copyright Law of the United States 2010b. https://www.law.cornell.edu/uscode/text/17/512 U.S.C. 17, 512.

CoStar Group, Inc. v. LoopNet, Inc. 2004. 373 F.3d 544 (4th Cir. 2004).

DAIR.AI 2023. General Tips for Designing Prompts. https://www.promptingguide.ai/introduction/tips

DC Comics v. Towle 2015. 802 F.3d 1012 (9th Cir. 2015).

DreamStudio 2023. https://dreamstudio.ai/

Feist Publications v. Rural Telephone Service Company 1991. 499 U.S. 340 (1991).

Deep Ganguli, Amanda Askell, Nicholas Schiefer, et al. 2023. The Capacity for Moral Self-Correction in Large Language Models. arXiv:2302.07459 [cs.CL]

Jon Garon. 2023. An AI's Picture Paints a Thousand Lies: Designating Responsibility for Visual Libel. *Journal of Free Speech Law* 3 (2023), 425.

Daniel Gervais. 2013. The Derivative Right, or Why Copyright Law Protects Foxes Better than Hedgehogs. *Vanderbilt Journal of Entertainment and Technology Law* 15 (2013), 785.

Daniel Gervais. 2022. AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines. *Seton Hall Law Review* 52 (2022), 1111.

GitHub. 2023a. About GitHub Copilot for Individuals, GitHub. https://docs.github.com/en/copilot/overview-of-github-copilot/about-github-copilot-for-individuals

GitHub. 2023b. Configuring GitHub Copilot in your environment. https://docs.github.com/en/copilot/configuring-github-copilot/configuring-github-copilot-in-your-environment

Aaron Gokaslan, A. Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. 2023. CommonCanvas: An Open Diffusion Model Trained with Creative-Commons Images. *arXiv preprint arXiv:2310.16825* (2023).

Google. August 17, 2023. Foundation Models. https://ai.google/discover/foundation-models/

James Grimmelmann. 2016. There's No Such Thing as a Computer-Authored Work – And It's a Good Thing, Too. *Columbia Journal of Law and the Arts* 39 (2016), 403.

Peter Henderson, Tatsunori Hashimoto, and Mark Lemley. 2023a. Where's the Liability in Harmful AI Speech? *Journal of Free Speech Law* 3 (2023), 589.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023b. Foundation Models and Fair Use. arXiv:2303.15715 [cs.CY]

Laura Heymann. 2008. Everything is Transformative: Fair Use and Reader Response. *Columbia Journal of Law and the Arts* 31 (2008), 445.

HuggingFace. 2023. Models. https://huggingface.co/models

Technology Innovation Institute. 2023. Falcon. https://falconllm.tii.ae/falcon.html

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. 2023. Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy. arXiv:2210.17546 [cs.LG]

Kadrey v. Meta Platforms, Inc. 2023. No. 3:23-cv-03417 (N.D. Cal. July 7, 2023).

Stefan Kahl, Connor M. Wood, Maximilian Eibl, and Holger Klinck. 2021. BirdNET: A Deep Learning Solution for Avian Diversity Monitoring. *Ecological Informatics* 61 (2021), 101236.

Kate Knibbs. 2023. The Battle Over Books3 Could Change AI Forever. *Wired* (4 Sept. 2023). https://www.wired.com/story/battle-over-books3/

Alan Latman. 1990. "Probative Similarity" as Proof of Copying: Toward Dispelling Some Myths in Copyright Infringement. *Columbia Law Review* 90 (1990), 1187.

Katherine Lee, A. Feder Cooper, James Grimmelmann, and Daphne Ippolito. 2023. AI and Law: The Next Generation.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 8424–8445.

Pierre N. Leval. 1990. Toward a Fair Use Standard. *Harvard Law Review* 103, 5 (1990), 1105.

Amanda Levendowski. 2018. How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem. *Washington Law Review* 93, 2 (2018), 579.

Joseph P. Liu. 2003. Copyright Law's Theory of the Consumer. *Boston College Law Review* 44 (2003), 397.

London-Sire Records, Inc. v. Doe 1 2008. 542 F. Supp. 2d 153 (D. Mass. 2008).

Ryan Lowe and Jan Leike. 2023. Aligning language models to follow instructions. *OpenAI* (2023). https://openai.com/research/instruction-following

James Manyika. August 17, 2023. An overview of Bard: an early experiment with generative AI. https://ai.google/static/documents/google-about-bard.pdf

Mata v. Avianca 2023. No. 22-cv-1461 (S.D.N.Y. June 22, 2023).

Peter S. Menell. 2011. In Search of Copyright's Lost Ark: Interpreting the Right to Distribute in the Internet Age. *Journal of the Copyright Society of the USA* 59 (2011).

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, Vol. 35.

Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd. 2005. 545 U.S. 913 (2005).

Metro-Goldwyn-Mayer v. American Honda Motor Co. 1995. 900 F.Supp. 1287 (C.D. Cal. 1995).

Sewon Min, Suchin Gururangan, Eric Wallace, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. 2023. SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. arXiv:2308.04430 [cs.CL]

Michael D. Murray. 2023. Generative AI Art: Copyright Infringement and Fair Use. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4483539

Tejas N. Narechania. 2022. Convergence and a Case for Broadband Rate Regulation. *Berkeley Technology Law Journal* 37 (2022), 339.

Naruto v. Slater 2018. 888 F.3d 418 (9th Cir. 2018).

Meta News. 2023. Introducing Code Llama, an AI Tool for Coding. https://about.fb.com/news/2023/08/code-llama-ai-for-coding/

Nichols v. Universal Pictures Corporation 1930. 45 F.2d 119 (2d Cir. 1930).

OpenAI. 2022. ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt

OpenAI. 2023a. ChatGPT plugins. https://openai.com/blog/chatgpt-plugins

OpenAI. 2023b. Custom instructions for ChatGPT. https://openai.com/blog/custom-instructions-for-chatgpt

OpenAI. 2023c. Our approach to AI safety. https://openai.com/blog/our-approach-to-ai-safety

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744.

Willilam F. Patry. 2023. *Patry on Copyright*.

Perfect 10, Inc. v. Giganews, Inc. 2017. 847 F.3d 657 (9th Cir 2017).

Peter Pan Fabrics, Inc. v. Martin Weiner Corp. 1960. 274 F.2d 487 (2d Cir. 1960).

Colin Raffel. 2023. Collaborative, Communal, & Continual Machine Learning. https://colinraffel.com/talks/faculty2023collaborative.pdf

Alex Reisner. 2023. Revealed: The Authors Whose Pirated Books are Powering Generative AI. *The Atlantic* (19 Aug. 2023). https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/

Rentmeester v. Nike, Inc. 2018. 883 F.3d 1111 (9th Cir. 2018).

Mark Riedl. 2023. A Very Gentle Introduction to Large Language Models without the Hype. https://rb.gy/tkfw5

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *2022 IEEE Conference on Computer Vision and Pattern Recognition*.

Matthew Sag. 2023. Copyright Safety for Generative AI. *Houston Law Review* (2023). Forthcoming.

Pamela Samuelson. 1985. Allocating Ownership Rights in Computer-Generated Works. *University of Pittsburgh Law Review* 47 (1985), 1185.

Pamela Samuelson. 2013. The Quest for a Sound Conception of Copyright's Derivative Work Right. *Georgetown Law Journal* 101 (2013), 1505.

Pamela Samuelson. 2016. Functionality and Expression in Computer Programs: Refining the Tests for Software Copyright Infringement. *Berkeley Technology Law Journal* 31 (2016), 1215.

Scale AI 2023. https://scale.com/

Sarah Scheffler, Eran Tromer, and Mayank Varia. 2022. Formalizing Human Ingenuity: A Quantitative Framework for Copyright Law's Substantial Similarity. In *Proceedings of the Symposium on Computer Science and Law*. 37–49.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open

large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 86––96.

Share-GPT 2023. https://sharegpt.com/

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The Curse of Recursion: Training on Generated Data Makes Models Forget. arXiv:2305.17493 [cs.LG]

Sid & Marty Krofft Television v. McDonald's Corp. 1977. 562 F.2d 1157 (1977).

Sinclair v. Ziff Davis, LLC 2020. 454 F.Supp.3d 342 (S.D.N.Y. 2020).

Skidmore v. Zeppelin 2020. 952 F.3d 1051 (9th Cir. 2020).

Benjamin L.W. Sobel. 2017. Artificial Intelligence's Fair Use Crisis. *Columbia Journal of Law and The Arts* 41 (2017), 45.

Benjamin L.W. Sobel. 2023. Elements of Style: A Grand Bargain for Generative AI. On file with the authors.

The Vicuna Team. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. *LMSYS Org* (30 March 2023). https://lmsys.org/blog/2023-03-30-vicuna/

TensorFlow 2023. TensorBoard: TensorFlow's visualization toolkit. https://www.tensorflow.org/tensorboard

Thaler v. Perlmutter 2023. No. 22-1564 (D.D.C. August 18, 2023).

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023a. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

Tremblay v. OpenAI, Inc. 2023. No. 3:23-cv-03223 (N.D. Cal. June 28, 2023).

Abhinav Venigalla and Linden Li. 2022. Mosaic LLMs (Part 2): GPT-3 quality for <$500k. *MosaicML* (29 Sept. 2022). https://www.mosaicml.com/blog/gpt-3-quality-for-500k

James Vincent. 2023. Meta's powerful AI language model has leaked online — what happens now? *The Verge* (2023). https://wandb.ai/site

Eugene Volokh. 2023. Large Libel Models? Liability for AI Outpu. *Journal of Free Speech Law* 3 (2023), 489.

Walker v. Time Life Films, Inc. 1986. 784 F.2d 44 (2d Cir. 1986).

Weights & Biases 2023. https://wandb.ai/site

BigScience Workshop et al. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. arXiv:2211.05100 [cs.CL] https://arxiv.org/abs/2211.05100

Rui-Jie Yew and Dylan Hadfield-Menell. 2023. Break It Till You Make It: Limitations of Copyright Liability Under a Pretraining Paradigm of AI Development. https://genlaw.github.io/CameraReady/30.pdf

Yiming Zhang and Daphne Ippolito. 2023. Prompts Should not be Seen as Secrets: Systematically Measuring Prompt Extraction Attack Success. arXiv:2307.06865 [cs.CL]

Jonathan Zittrain. 2008. *The Future of the Internet–And How to Stop It.* Yale University Press, USA.

NOTES

1 The generative-AI supply chain is a very good example of the "many hands" problem in computer systems. That is, there are many diffuse actors, at potentially many different organizations, that can each have a hand in the construction of generative-AI systems. It can be very challenging to identify responsible actors when these systems transgress broader societal expectations — in our case, the preservation of copyrights. See Cooper et al. [2022, pp. 867-869] (describing the problem of "many hands" in data-driven ML/AI systems); Yew and Hadfield-Menell [2023] (regarding an instantiation of this problem for generative AI and copyright).

2 We discuss this in more detail below with respect to generation (Section 2.7).

3 A data example is not the same as the expressive work. Additionally, some models are trained on synthetic data, typically generated by other generative-AI models [Gokaslan et al. 2023, e.g.]. However, training predominantly on synthetic data is not reflective of current common practices in today's generative-AI systems. Further, there are concerns that training on synthetic data can seriously compromise model quality. See generally Shumailov et al. [2023] (detailing "model collapse" in different generative models).

4 It appears increasingly likely that some content will be created specifically for model training. For example, hiring photographers to take photographs specifically for model training. Companies like Scale AI already create content (in the form of labels and feedback) specifically for the purpose of training models [Scale AI 2023].

5 Of course, data examples can still be copies of original works, and thus still infringe intellectual property rights.

6 We discuss fixation in Section 3.1. An exception is training data produced by generative-AI systems, as such data currently have been found to not be copyrightable. See Thaler v. Perlmutter [2023]. We discuss using generations as training data in Section 2.7.

7 Until recently, Copilot was built on top of OpenAI's Codex model.

8 This is not the only way to collect large amounts of data. See Lee et al. [2023] (discussing other ways datasets may come to be).

9 See generally Lee et al. [2023] (discussing dataset creation and curation choices, including toxic content filtering).

10 Technically, LAION presents the dataset as a collection the URLs of the images. Model trainers visit each URL to collect images for training.

11 In practice, however, it appears that most uses of training datasets are licensed — either through a bilateral negotiation or by means of an open-source license offered to the world by the dataset compiler.

12 Indeed, the creators would have to check that they have abided by each data example's respective license. Some example pairs could potentially have multiple licenses – e.g., an image and its associated caption could have their own copyrights and licenses.

13 LAION-5B, a large image-caption dataset, was released as under Creative Commons CC-BY 4.0. LAION-5B released a dataset of text captions and URLs to images, instead of the images themselves [Beaumont 2022; Schuhmann et al. 2022]. It is unclear if the LAION team had the rights to license the images within. Notably,the website introducing the LAION dataset provides a feature called "pwatermark," which is a prediction of how likely the image is to contain a watermark. The LAION team estimates that the 6.1% of the dataset Laion2B-en contains watermarked images. Another example comes from the complaint in Tremblay v. OpenAI, Inc. [2023], which alleges that ChatGPT's underlying model(s) were trained on datasets that do not license the books data that they contain. The complaint alleges that the training data included books from infringing "shadow libraries" like Library Genesis. See Complaint at Tremblay v. OpenAI, Inc. [2023, p. 34] But this claim is based on circumstantial evidence, because the datasets it was trained on have not been made public. Text from books have been a key player in other dataset-related complaints. For example, The Pile data was originally released under the MIT license [Biderman et al. 2022]. The Pile was core to the complaint in Kadrey v. Meta Platforms, Inc. [2023], since the Pile claimed to contain 108GB of the dataset Books3 (which itself contains content from Bibliotek, a popular torrent interface). The original download URL for The Pile (https://the-eye.eu/public/AI/pile/) is no longer resolving (as of September 2023). LAION has also been taken down from popular hosting services, following a report documenting the presence of CSAM at associated image URLs.

14 We distinguish between the person or organization that trains from those that create the model architecture, as they may not be the same.

15 ML uses tools from probability and statistics, which reason about randomness. However, computers are not able to produce truly random numbers. Instead, algorithms exist for producing a sequence of *pseudo*-random numbers. A random seed is an input to a pseudo-random number generator, which enables the reproduction of such a sequence. The trainer also selects hyperparameters [Cooper et al. 2021b], which we elide for simplicity.

16 Training costs are often not reported. Even when training cost is reported, development costs (including labor) are often omitted, despite being a critical (and often most expensive) part of overall model development.

17 The original cost to train GPT-3 is unpublished, though, based on its size, is likely higher than $0.5 million. MosaicML reports to have trained a GPT-3-*quality* model. This means the model performs to a similar standard as GPT-3 does. Nevertheless, MosaicML's model is substantively different from GPT-3. For one, MosaicML's model is much smaller — 30 billion parameters compared with the original GPT-3 model's 175 billion. Additionally, MosaicML trained their model on more data, shifting some of the development cost toward data collection and away from model training. It is worth noting that GPT-3 was originally released two years before MosaicML's model was trained, and thus the MosaicML training process likely incorporated additional technological improvements. See generally Venigalla and Li [2022] (regarding MosaicML's model). See generally Brown et al. [2020] (for the size of GPT-3).

18 Further, the training process is not completely automated; training often requires people to monitor and tweak the model. For example, model trainers typically run evaluation metrics on the model while it is being trained, in order to assess the progress of training. Google's TensorBoard [TensorFlow 2023] and software from Weights & Biases [Weights & Biases 2023] are two tools for running evaluation metrics and monitoring during training. Depending on these metrics (which attempt to elicit how "useful" or "good" the model is, but are not comprehensive [Lee et al. 2023]) model trainers may pause the training process to manually revise the training algorithm (e.g., change the hyperparameters.) or the dataset, which we indicate with bidirectional arrows at Figure 1, stages 3-4. Human intervention in response to metrics necessarily makes model training an iterative process.

19 Others use the term "foundation model." The term "foundation" can be easily misunderstood. It should not be interpreted to connote that "foundation models" contain technical developments that make them fundamentally different from models produced in the nearly-a-decade of related prior work. The term itself has been met with controversy within the ML community, which can be seen

expressed on programming forums and in conversations, e.g., we refer to a Twitter thread (and its associated offshoots) that involves renowned researchers and some of the Stanford authors that coined the term "foundation models." (*See* https://twitter.com/tdietterich/status/1558256704696905728).

20 This reveals the murky line between what exactly is a program and what exactly is data in ML, more generally. The set of parameters can be viewed as a *data structure* containing vectors of numbers that, on its own, does not *do* anything. However, we could load that data structure into memory and apply some relatively lightweight linear algebra operations to produce a generation. In this respect, we could also consider the model to be a program (and, indeed, an algorithm). The model, if given a prompt input, can also be executed like a program. Note that the term "model" is overloaded; it can be used to refer to the model parameters (vectors of numbers) or to the model as a combination of software and the model parameters, which together can be executed like a program.

21 For example, HuggingFace hosts a repository of over 300,000 open-sourced models [HuggingFace 2023].

22 They could fine-tune the model (Section 2.5), embed the model in a system that they deploy for others to use (Section 2.6), produce generations (Section 2.7), align the model (Section 2.8), or do some subset of these other stages of the supply chain. From this example, we can see how the supply chain is in fact iterative, which we illustrate in Figure 1.

23 And thus the reason for the bidirectional arrow between stages 3 and 5 in Figure 1. Similar to pre-training, monitoring metrics during fine-tuning may lead to further dataset curation (Section 2.4).

24 There are other relevant factors in training, including choice of hyperparameters and choice of hardware. These, too, can change between pre-training and fine-tuning. We again elide these details for simplicity.

25 In this respect, it is important to note that a model is a "base" or "fine-tuned" model *only in relation to other models*. These terms do not capture inherent technical features of a model; instead, they describe different processes by which a model can be created.

26 To give a concrete example of the many actors in the generative-AI supply chain, consider Vicuna. LMSYS Org fine-tuned Meta's Llama model on the crowd-sourced ShareGPT dataset to produce Vicuna [Share-GPT 2023; Team 2023]. ShareGPT is a crowd-sourced dataset composed of conversational logs of user interactions with ChatGPT. It contains both content created by users and by the generative-AI model embedded in ChatGPT (either GPT-3.5 or GPT-4, depending on the user) [Share-GPT 2023]. Vicuna has also released their model publicly, affording a potentially infinite host of actors the ability to fine-tune the model on additional data. See Raffel [2023, slide 15] (for a figure showing many fine-tuned models building on one base model).

27 Meta first asked interested parties to request Llama's model parameters, rather than uploading them publicly on the web. However, Llama's model parameters were quickly leaked on the website 4chan [Vincent 2023]. This incident shows how challenging it can be to control access to models once released. Llama also includes a use policy in the Llama 2 Community License that outlines prohibited uses of the model. Of course, it is impossible to enforce prohibited uses when releasing model parameters. This is also why many model trainers choose to release models through hosted services. See AI [2023a] (for the Llama 2 Community License).

28 Another deployment option is a command-line interface (CLI), which takes a user-supplied prompt as input (via a code terminal) and directly returns the resulting generation as output. https://ollama.ai/ (the download link of the Ollama CLI, which is a wrapper program around various Llama-family LLMs).

29 It is possible that models released and deployed in multiple ways might not all be exactly the same; they could have different versions of model parameters. This may be made explicit to users, as with ChatGPT, or may not be communicated to them, and thus unclear or unknown. See generally OpenAI [2022] (regarding both GPT-3.5 and GPT-4 model integration into the ChatGPT web application).

30 See generally Zhang and Ippolito [2023] (which discovers proprietary system prompts); OpenAI [2023b] (announcing a ChatGPT feature that allows users to provide their own additional prompts, which get appended to their future inputs to create compound prompts).

31 This kind of prompt transformation is another technique for steering the behavior of a model.

32 See Akın [2023] (These prompts and more can be found on this site); DAIR.AI [2023] (This handbook provides an introduction to creating prompts for large language models); OpenAI [2023b].

33 By analogy, the function f that contains the model is not directly available to users; instead, f is made accessible indirectly via a hosted service.

34 These filters may detect undesired inputs and prevent the model from generating an output, or detect undesired outputs and prevent the system from displaying the generation. In both cases, the model parameters would not be changed. This need not be the case, the model parameters may also be directly modified through alignment to respond to undesired inputs in a more desirable way. Of course, though, for ChatGPT, we do not know exactly how filters are implemented.

35 See https://news.ycombinator.com/item?id=33226515 (for related discussion on the Hacker News forum).

36 Each mechanism for making model functionality widely available has different pricing structures that can ultimately impact the quality of the model. While the open-source community works hard to create and release models that compete with the best closed-source models, current open-source models are mostly trained on open-sourced data and are often lower quality. The best open-sourced models are very good, but still not as good as closed-source proprietary models. For example, Technology Innovation Institute in Abu Dhabi recently released the model, Falcon 180B (a 180 billion parameter model), which they claim is better than Meta's Llama 2 but still behind GPT-4 [Institute 2023]. Additionally, differences between open- and closed-source datasets can lead resulting trained models to vary in quality. For example, Min et al. [2023] uses public domain and permissively licensed text to train a language model, and demonstrates a degradation in quality in domains that are not well represented in the data. Additionally, data in the public domain can be unrepresentative of certain demographic groups [Levendowski 2018].

37 See Section 2.4 (noting, however, that models do not *have to* be used to produce generations).

38 See Section 2.4 (discussing how the term "model" is overloaded, and can refer to model parameters being embedded in a program that executes (typically linear algebra) operations to to perform generation.)

39 For generative models, there are many reasonable outputs for the input. There are also other sources of randomness in generation that are implementation-specific, such as the choice of decoding strategy for language models. See Riedl [2023] (for an accessible discussion of decoding).

40 We focus on deployed systems — and their API and web-based interfaces — because there are more opportunities for the deployer to control the model. But, of course, the user could have written some code to produce generations using released open-source model parameters.

41 These three options highlight additional observations about prompts. Thus far, we have primarily discussed generations as expressive works, but prompts could also be expressive works. The expressive example we gave above was: `"a big dog facing left wearing a spacesuit in a bleak lunar landscape with the earth rising in the background as an oil painting in the style of Paul Cezanne high-resolution aesthetic trending on artstation"`. Sufficiently expressive prompts written by the direct user of a service could be subject to copyright. Context windows are so large, it is even possible for the user to prompt with an entire expressive work. As we discuss below in our copyright analysis, it is of course possible for this expressive work to have also been authored by another individual. Prompts could also be produced by generative AI, but this does not have the same authorship considerations. For example, Anthropic's team discussed using the entire text of *The Great Gatsby* as a prompt to demonstrate the long context window of their language model, Claude [Anthropic 2023]. While *The Great Gatsby* is now in the public domain, it is easy to imagine another book entered as the prompt, or a copyrighted image as the prompt in an image-to-image system. Or copyrighted audio as input to an audito-to-audio model, etc. User-supplied prompts may be stored on system-deployers' servers for non-transient periods of time, and may even serve training data for a future model. Such prompts may also be used in model alignment (Section 2.8).

42 Using model outputs as training data for future models has been a common practice in other settings. For instance, back-translation, the process of using a machine-translation model to generate additional training data (by translating data from one language to another) is a common technique [Sennrich et al. 2016].

43 There are also concerns that this practice can have negative effects on model quality [Shumailov et al. 2023].

44 Of course, words like "good" and "bad" can have multiple valences, and resist the kind of quantification on which ML depends. See Lee et al. [2023] (discussing the challenges of defining "good" and "bad" in the context of model behavior).

45 See Lowe and Leike [2023] (for an introduction to InstructGPT, a model that is aligned with human feedback).

46 In the reinforcement learning setting, data is not labeled as explicitly as it is in discriminative setting, e.g., our example of an image classifier, where each training data image has a label of either `cat` or `dog`. (Instead, generations may be labeled "good" or "bad" based on human feedback, and the reinforcement learning algorithm updates the model in response to that feedback. In RLHF, feedback is generated by a person interacting with the system; however, RL can also use feedback automatically generated by an algorithm specification [Bai et al. 2022].

47 Before making models publicly available, these companies contract with firms, like Scale AI [Scale AI 2023], that simulate the user feedback process. These firms typically employ people to label generations as "good" or "bad," according to guidance from the generative-AI company. In general, the process of model alignment is a critical part of the supply chain. It serves as a mechanism for steering models away from generating potentially harmful outputs (See Cole [2023], describing a book on mushroom foraging built from generations, which

mistakenly indicate that toxic mushrooms are safe to eat) and toward the policies of the company or organization that deployed the model. See Ganguli et al. [2023]; Manyika [2023]; OpenAI [2023c] (documenting safety considerations, alignment, and RLHF at Google, OpenAI, and Anthropic).

48 17 U.S.C. § 102(a) (numbering added).

49 17 U.S.C. § 101 (definition of "fixed").)

50 17 U.S.C. § 102(b).

51 See Kahl et al. [2021]. Animals are not recognized as "authors" for copyright purposes. See Naruto v. Slater [2018].

52 17 U.S.C. § 103(a).

53 17 U.S.C. § 101 (definition of "compilation").

54 That is, programs, like animals, are not "authors" within the meaning of the Copyright Act.

55 17 U.S.C. § 106

56 17 U.S.C. § 106(2)

57 17 U.S.C. § 101 (definition of "derivative work).

58 But see Scheffler et al. [2022] (describing a principled computational basis for comparing works)

59 See Carlini et al. [2021] (GPT-2 memorizes training data); Carlini et al. [2023a] (Stable Diffusion and Imagen memorize images); Chang et al. [2023] (suggestive evidence that GPT-4 memorizes training data).

60 This is a sticky technical problem. Research has shown that memorization is not easily identifiable, and thus the amount of memorization in a model is not always or easily quantifiable. In particular, the choice of memorization identification technique and available information (e.g., knowledge of the training dataset, context window, etc.) affect the amount of memorization that can be identified. See, e.g., Carlini et al. [2023b].

61 Alert readers will note the similarity to the debate over whether the mere act of making a work available without a download infringes the distribution right. See London-Sire Records, Inc. v. Doe 1 [2008]; see generally Menell [2011].

62 While it may be straightforward to pose the question: "is the given generation substantially similar to work 1," it is not at all straightforward to answer. Training datasets are massive. Manually comparing the generation to every single work in the dataset is infeasible; it would simply take too long. While automated methods could help identify works in the training set that are *likely to be* similar to the generation, there is no automated metric that can definitively say if two works are substantially similar. See generally Scheffler et al. [2022] (which proposes one possibility for a metric for identifying substantial similarity)). Even with automated methods, checking *every* generation that a system produces against every other work in the training dataset to evaluate similarity is extremely computationally expensive.

63 To quote Learned Hand on the idea-expression dichotomy, "Nobody has ever been able to fix that boundary, and nobody ever can" [Nichols v. Universal Pictures Corporation 1930, p. 121].

64 Notably, providing guarantees that any given generated work might not potentially infringe copyright is impossible if the training data contains copyrighted data. This is simply because provable guarantees require formal definitions, and there are no widely accepted formal definitions of substantial similarity. But see Scheffler et al. [2022] (providing a possible starting point). Instead, current ML techniques focus on reducing the likelihood that generations from a model will closely resemble any of the model's training data.

65 E.g., DC Comics v. Towle [2015]; see generally Sag [2023] (discussing caselaw and scholarship)

66 See Metro-Goldwyn-Mayer v. American Honda Motor Co. [1995] (car commercial featuring "a handsome hero who, along with a beautiful woman, lead a grotesque villain on a high-speed chase, the male appears calm and unruffled, there are hints of romance between the male and female, and the protagonists escape with the aid of intelligence and gadgetry" infringes on James Bond character).

67 See Nichols v. Universal Pictures Corporation [1930, p. 121] ("Though the plaintiff discovered the vein, she could not keep it to herself; so defined, the theme was too generalized an abstraction from what she wrote. It was only a part of her 'ideas.' ")

68 See generally Skidmore v. Zeppelin [2020] (discussing proof of copying in fact); Latman [1990] (distinguishing "probative" similarities that prove copying in fact from substantive similarities that constitute improper appropriation).

69 See American Broadcasting v. Aereo [2014, 2512-13] at 2512-13 (Scalia, J., dissenting)

70 See Cartoon Network LP, LLLP v. CSC Holdings, Inc. [2008, p. 130]; see also Perfect 10, Inc. v. Giganews, Inc. [2017].

71 Such as a text-to-image model developer using the model to create example prompt/generation pairs to display on their website.

72 It is worth briefly noting that plugins could additionally pull in content from external sources, such as a news website, that gets included in a generation. Recall that this data is *not* included in training the model; instead, it is fed into the model at generation time to try to improve the quality of generations with more up-to-date information [OpenAI 2023a] Hypothetically, this content could get included verbatim in generations, leading to infringement issues in generation separate from those discussed above.

73 17 U.S.C. § 107(1).

74 See Cariou v. Prince [2013, p. 707] (focusing audience perceptions of works rather than author's intentions in assessing transformative use); see generally Heymann [2008] (assessing transformative use from audience perspective); Liu [2003] (discussing audience interests in copyright).

75 See 17 U.S.C. § 107 [Copyright Law of the United States 1992] (favoring "purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research")

76 17 U.S.C. § 107(2)

77 17 U.S.C. § 107(3)

78 See Associated Press v. Meltwater U.S. Holdings, Inc. [2013] (rejecting fair use defense brought by news-monitoring service that reproduced substantial excerpts from articles for its customers)

79 17 U.S.C. § 107(4)

80 Perhaps the user has already watched all of the existing episodes.

81 For another example, imagine that the user of a service prompts a text-to-image system to create a portrait of them in the style of a particular living artist; the generation is a substitute for commissioning the artist to paint one.

82 Here, we use the term "user" broadly. A user could be a customer using a web application to produce a generation, a developer using an API to produce a generation in their own code, a developer using an API to produce a generation for a company, etc.

83 Sometimes literally so. See Adobe [2023].

84 Regardless of which of these doctrinal routes a court took, there would be an inevitable gravitational force pulling the provider's duties towards the duties of a service provider under Section 512(c) or (d). This is not because Section 512 applies to generative-AI services. It largely does not — analysis that we defer to other work. Instead, the Section 512 doctrines may be a convergence point because courts have now had two decades of experience — which means two decades of precedents — with the Section 512 safe harbors. These precedents have come to set expectations — among copyright owners, in the technology industry, in the copyright bar, and in the judiciary — for what legally "responsible" behavior by an online intermediary looks like. A generative-AI service operator that does not appear to be making a good-faith effort to achieve something like this system may strike a court as intending to induce infringement, not making a good-faith effort to comply with an injunction, etc.

85 U.S.C. § 512(c)(3)(A)(iii).

86 That said, matching material against a catalog of copyrighted works is a problem that has been very approximately solved by major social networks, which use perceptual hashing to prevent the upload of various kinds of identified content. Generative-AI companies could at least add similar perceptual-hash-driven filtering to the outputs of their models, but clearly this would only solve part of the problem [Ippolito et al. 2023; Lee et al. 2022].The challenges of implementing removal for models are even harder. A service can add filters on the input and output sides — monitoring prompts and scanning outputs. It can also fine-tune or align the model, or provide it with an overall prompt that instructs the model to respond in ways that reduce its propensity to infringe. Further, a model by itself does not implement these controls. The model cannot control how it is prompted or what the user does with the output. The model cannot stop anyone from fine-tuning it to remove its guardrails.

87 Absent the ability to do so, the safest bet is to retrain the model from scratch. Due to the time and expense required to retrain a model, it will often be infeasible to retrain it simply to remove infringing works, and completely unworkable to retrain on each new notice. We defer further discussion of how courts could deal with this difficulty to other work. A notice-and-removal regime also has implications for training datasets. A dataset provider cannot pull back these works for which it receives a notice from others who have already used those works for training. But it can delete the works from the dataset it makes available to others going forward. For an open-source dataset, or one that has been leaked, this second option may be futile, as others will still have copies of the dataset that they can share. Compared with a model, it is much easier to remove a work from a training dataset; one searches for the work and removes it. Indeed, one could use exact hashing rather than perceptual hashing and still get substantial efficacy in removing a large number of identified works from the dataset — or, for datasets compiled from web crawls or other sources, remove works by tracing their provenance through into the part of the dataset they have ended up in. This makes datasets more attractive as removal targets. They are upstream from many models. Also, it is easier to define and enforce enforceable removal obligations.

88 Each model would have a fully licensed training dataset, and the question of infringement would not arise except in cases where there were infringing works

in the dataset itself or some other failure of quality control somewhere along the supply chain.