TALKIN' 'BOUT AI GENERATION: COPYRIGHT AND THE GENERATIVE-AI SUPPLY CHAIN

by KATHERINE LEE,* A. FEDER COOPER** AND JAMES GRIMMELMANN***

"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, including conversational chatbots like ChatGPT, image generators like Midjourney and DALL·E, coding assistants like GitHub Copilot, music composition applications like Lyria, and video generation systems like Sora. Generative-AI models have different technical architectures and are

* Co-founder, The Center for Generative AI, Law, and Policy Research (The GenLaw Center); Ph.D. Candidate in Computer Science, Cornell University (on leave); Staff Research Scientist and GenAI Attack Team Lead, Google DeepMind. This article was written during Lee's time as a graduate student researcher at Cornell, in her capacity as an affiliate of The GenLaw Center and Cornell. All authors contributed equally to this work. We presented an earlier version of this work at the 2023 Privacy Law Scholars Conference and discussed the issues extensively with other participants in the Generative AI + Law Workshop at the 2023 International Conference on Machine Learning. Our thanks to the organizers and participants, and to Aislinn Black, Jack M. Balkin, Miles Brundage, Christopher Callison-Burch, Nicholas Carlini, Madiha Zahrah Choksi, Christopher A. Choquette-Choo, Christopher De Sa, Fernando Delgado, Jonathan Frankle, Deep Ganguli, Daphne Ippolito, Matthew Jagielski, Gautam Kamath, Kevin Klyman, Mark Lemley, David Mimno, Niloofar Mireshghallah, Milad Nasr, Pamela Samuelson, Ludwig Schubert, Andrew F. Sellars, Florian Tramèr, Kristen Vaccaro, and Luis Villa. We would additionally like to thank the editorial staff at the *Journal of the Copyright Society of the U.S.A.* for shepherding this Article through the publication process. The GenLaw Center, https://www.genlaw.org/, is an independent academic research organization. It has organized workshops at the International Conference on Machine Learning in July 2023 and July 2024, and in Washington, D.C. in April 2024. These events received financial and in-kind support from Schmidt Sciences, OpenAI, the ML Collective, the Georgetown Institute for Technology Law & Policy, the K&L Gates Initiative at Carnegie Mellon University, the Center for Democracy and Technology, Google, Microsoft, Schmidt Futures, Anthropic, and Cornell University.

** Co-Founder, The GenLaw Center; Assistant Professor of Computer Science, Yale University (to commence 2026). After completion of the initial version of this Article but prior to its official publication, Cooper started as a Postdoctoral Researcher at Microsoft Research and a Postdoctoral Affiliate at Stanford University.

*** Tessler Family Professor of Digital and Information Law, Cornell Law School and Cornell Tech; Director, Cornell Tech Research Lab for Applied Law and Technology (CTRL-ALT). CTRL-ALT has received funding from Microsoft for a project on Internet platform regulation. Grimmelmann previously was on the faculty of New York Law School and affiliated with its Institute for Information Law and Policy, which received funding from Microsoft for a project on the Google Books litigation.

trained on different types of data from different sources using different algorithms. Some take months and cost millions of dollars to train; other models can be spun up in a weekend. These models are made accessible to users in very different ways. Some are offered through paid online services; others are distributed as open-source artifacts, which let anyone download and modify them. Different generative-AI systems behave differently and raise different legal issues. We therefore need the right framework—one that digs deeper than the term "generative AI"—to reason precisely and clearly about the different legal issues at play.

The second problem is that 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. Whether the output of a generative-AI system is fair use can depend on how its training datasets were assembled. Whether the creator of a generative-AI system is secondarily liable can depend on the prompts that its users supply.

In this Article, we aim to bring order to the chaos. To do so, we make two contributions. First, we introduce the generative-AI supply chain: an interconnected set of stages that transform training data (millions of pictures of cats) into generations (new and hopefully never-seen-before pictures of cats that have never existed). Breaking down generative AI into these constituent stages reveals all the places at which companies and users make choices that may have legal consequences—for copyright and beyond. Second, we specifically apply the supply-chain framing to U.S. copyright law. This framing 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 do not give definitive answers as to who should and should not be held liable. Instead, we identify the key decisions that courts will need to make as they grapple with these issues, and we point out the consequences that would likely flow from different liability regimes.

INTRODUCTION

Generative artificial intelligence (i.e., "generative AI") systems like ChatGPT, Claude, Gemini, DALL·E, and Ideogram are capable of turning a user-supplied prompt like "give three arguments why Marbury v. Madison was wrongly decided" into a persuasive essay, or "a robot cowboy riding a rocket ship" into a work of digital art. The results can be striking. Lawyer Adam Unikowsky has described Claude as an "insane genius" for its ability to generate a "totally novel, weird" rule to decide an actual Supreme Court case. (Claude's suggestion was that the First Amendment treatment of government officials' social-media

posts should turn on how many “likes” those posts get.[1]) And here is what Midjourney came up with for "a weasel on the first day of high school":

Figure 1: Prompting Midjourney, a text-to-image generative-AI application, with "a weasel on the first day of high school.” (Produced by the authors.)

These systems’ unpredictability and complexity means that they break out of existing legal categories. In particular, the fact that generative-AI systems involve training on millions (or even billions) of examples of human creativity means that they raise serious copyright issues. These copyright issues have not gone unnoticed. Numerous groups of plaintiffs have sued leading generative-AI companies for copyright infringement, with potential damages reaching into the billions of dollars.[2] Legislators and regulators around the world are grappling with how to fit these technologies into copyright law’s framework—or how to modify copyright law to fit these technologies.[3]

[1] Adam Unikowsy, *In AI We Trust, Part II*, ADAM’S LEGAL NEWSLETTER (June 16, 2024), https://adamunikowsky.substack.com/p/in-ai-we-trust-part-ii.

[2] For a detailed tracker of the copyright lawsuits against generative-AI companies, including their claims and current statuses, *see* ChatGPT Is Eating the World, https://chatgptiseatingtheworld.com/, maintained by law professor Edward Lee.

[3] *See, e.g.*, United States Copyright Office, Artificial Intelligence Study, https://www.copyright.gov/policy/artificial-intelligence/.

This Article attempts to think carefully and systematically about how copyright applies to generative-AI systems. Our first contribution is to be precise about what "generative AI" is. It is not just one product from one company. Instead, "generative AI" is a catch-all name for a massive ecosystem of loosely related technologies, including conversational text chatbots like ChatGPT, image generators like Midjourney and DALL·E 3, 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 types and sources of data using different algorithms. Some take months and cost millions of dollars to train, while others can be spun up in a weekend. These models are also made accessible to users in very different ways.

For example, some are offered through paid online services while others are distributed open-source so that anyone can download and modify them.[4]
This Article takes the complexity and diversity of generative-AI systems seriously. To provide a clear framework for thinking about the different kinds of generative-AI systems and the different ways they are created and used, the Article introduces what we call the generative-AI supply chain: an interconnected set of eight stages[5] that transform training data (millions of pictures of cats) into

[4] The use of the term "open-source" in generative AI is quite complicated. Some models are truly open-source, in the sense that their parameters and information about the training data are publicly released. Others, which are often also called "open-source" models, only release the parameters, and do not release information about the training data. Some literature refers to this second case as "semi-closed." "Closed" models are those for which neither the model parameters nor information about the training data are available. For simplicity, we will elide this nuance; it is an important detail for understanding the generative-AI supply chain, but not for our purposes here concerning copyright. *See infra* Part I.A (regarding training data, model parameters, and models); *See infra* Part I.C (regarding the generative-AI supply chain); Milad Nasr, Nicholas Carlini, Jonathan Hayase et al., Scalable Extraction of Training Data from (Production) Language Models (Nov. 28, 2023) (unpublished manuscript), https://arxiv.org/abs/2311.17035 (for distinguishing closed, semi-closed, and open models); Stella Biderman, Hailey Schoelkopf, Quentin Anthony et al., *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*, *in* 2023 PROC. 40TH INT'L CONF. ON MACH. LEARNING 2397–2430 (2023); Dirk Groeneveld, Iz Beltagy, Pete Walsh et al., OLMo: Accelerating the Science of Language Models (Feb. 1, 2024) (unpublished manuscript), https://arxiv.org/abs/2402.00838 (for examples of open models); Hugo Touvron, Thibaut Lavril, Gautier Izacard et al., LLaMA: Open and Efficient Foundation Language Models (Feb. 27, 2023) (unpublished manuscript), https://arxiv.org/pdf/2302.13971.pdf; Hugo Touvron, Louis Martin, Kevin Stone et al., Llama 2: Open Foundation and Fine-Tuned Chat Models (July 19, 2023) (unpublished manuscript), https://arxiv.org/pdf/2307.09288.pdf (for examples of semi-closed models); OpenAI, *ChatGPT: Optimizing Language Models for Dialogue*, OPENAI (Nov. 30, 2022), https://web.archive.org/web/20221130180912/https://openai.com/blog/chatgpt/ (for an example of a system that embeds a closed model).

[5] *See infra* Part I.C (outlining, and then detailing, the eight different stages).

generations (new and hopefully never-seen-before pictures of cats that have never existed). Breaking down generative AI into these constituent stages reveals all of the places at which companies and users make choices that may have legal consequences—for copyright and beyond.

In our analysis, we specifically explore the copyright consequences. So next, the Article works systematically through the copyright analysis of the different stages of the supply chain. 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, to name a few. These issues cannot be analyzed in isolation because there are connections everywhere. Whether the output of a generative-AI system is fair use can depend on how its training datasets were assembled. Whether the creator of a generative-AI system is secondarily liable can depend on the prompts that its users supply. The Article traces the effects of upstream technical designs on downstream uses and assesses 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 can shed more light on the copyright questions. We do not give definitive answers as to who should and should not be held liable. Instead, 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.

The Article is organized in three Parts. It begins in Part I by describing the generative-AI supply chain in detail. To do so, this section leads with the necessary technical background on the broader field of machine learning (Part I.A) and then explains how generative AI both relates to and is distinct from more traditional machine learning (Part I.B). The heart of this section is a detailed, step-by-step guide to the generative-AI supply chain (Part I.C), describing what happens at each of the eight stages, the many different ways that each stage can be designed and implemented, and the design choices that the various actors must make to create, deploy, and use a generative-AI system.

Part II then provides the copyright analysis. Here, we follow the doctrinal stages of a typical copyright lawsuit: starting with authorship (Part II.A) and then covering infringement (Parts II.B through II.E), secondary liability (Part II.F), defenses, including fair use (Parts II.G through II.J), and remedies (Part II.K). We explore *what* might possibly be an infringing technical artifact, *who* might be an infringing actor, and *when* infringement might occur. We also include a discussion of three legal regimes that, while not strictly copyright, are close enough that they raise similar issues: removal of copyright management information (Part II.L), the right of publicity (Part II.M), and hot news misappropriation (Part II.N). This is where—we hope—our choice to detail the generative-AI supply chain proves its worth. Instead of asking discrete and insular questions like "Are generative-AI models fair use?" we consider how the fair-use analysis changes as one moves up and down the supply chain. We describe how the choices made by actors at one point in the supply chain affect the copyright risks faced by others; we show how

copyright compliance depends on coordinated action by parties upstream and downstream from each other.

Part III pulls back to provide broader lessons. First, we describe the options courts have, from no copyright liability at all to shutting down generative AI completely (Part III.A). We explain why courts may be drawn to various regimes and what the risks and instabilities of those regimes are. Then we offer some thoughts on how courts should conceptualize copyright and generative AI (Part III.B). 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 draw on more familiar technologies, which provide inadequate answers to the genuinely hard problems before them.

This Article is intended for two audiences: technologists and lawyers. We are two machine-learning researchers and a legal scholar, brought together by the goal of helping our communities understand each other's contributions and concerns around generative AI. Part I explains the basics of the technology for lawyers; many passages in Part II explain the basics of copyright law for technologists. We apologize if technologists find our discussion of how machine learning works to be tedious and elementary in places, or if lawyers feel the same way about our exposition of copyright law. But we think that having a shared foundation of knowledge will be useful to both groups and ultimately is the only way to be precise about the interplay between copyright and the generative-AI supply chain. If you find that a particular section is rehashing something you already know, it is fine to skim or skip it and flip ahead to the parts that are new. Not everything in this Article is for everyone, but we hope that it at least contains something for everyone.[6]

I. MACHINE LEARNING AND THE GENERATIVE-AI SUPPLY CHAIN

The terminology associated with generative AI is extensive, overloaded, and sometimes perplexing.[7] As a first step, we provide some background on data and machine learning, and we rely on these details to be precise about what is new (and not-so-new) in generative AI. Section A provides a general background on basic concepts in machine learning, and Section B on what makes "generative AI"

[6] The first draft of this Article was completed in July 2023. We have updated the references to technology and copyright court cases as of September 16, 2024. The contributions in this Article hold up in light of numerous new technological and product developments, deployment paradigms, prompting strategies, and much else. Given the speed and volume of updates in both machine learning and the courts, we have nevertheless opted to use this as a cutoff date for detailed examples and court cases that support our contributions.

[7] For a glossary of terms in machine learning and generative AI, *see* Appendix A, A. Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito et al., Report of the 1st Workshop on Generative AI and Law (Nov. 11, 2023) (unpublished manuscript), https://arxiv.org/abs/2311.06477.

distinctive. Our main contribution is in Section C, which describes the generative-AI supply chain in detail, and which we find to be an essential abstraction for legal analysis that contends with generative AI.

Before getting to the details, a note about the illustrative examples that we provide in this Part: Generative AI can produce many kinds (or modalities[8]) of outputs, including text, images, music and other audio, code, and video. We focus on two: text and images. It is important to consider multiple examples because there are significant differences between them, and we want to emphasize that not all generative-AI systems work the same way. As a result, there are important technical details for other modalities that are not covered in this Part.

A. Background on Machine Learning

Machine learning is the discipline of algorithmically identifying and applying patterns in data. We begin by discussing data, which are the fundamental inputs to all machine-learning algorithms. We then provide a brief primer on the aims of machine learning, with special attention paid to how techniques used for generation differ from methods used for more familiar tasks like prediction and classification.

1. Data

In the context of AI and machine learning, data are digital representations of information about things in the world: people, events, physical phenomena like the weather, creative works like stories or software, etc. For example, Census data reflect information about individual people and households in the United States during the survey period. This information is made up of specific features, such as age, zip code, and income. The Census Bureau decided which information to collect and then decided how to represent that information in digital form. In general, individual records (such as the Census data on a single household) are typically called data examples, and a collection of related examples is a dataset. Information comes in many forms, including raw numbers, text, audio, images, and video. All of these must be converted to numerical representations to be stored, processed, and interpreted by a computer and, subsequently, by machine-learning algorithms and models.[9] These representations are often in forms that are convenient for automatic processing, rather than convenient for human inspection.

---

[8] *See infra* Part I.B.2.

[9] For simple examples of different types of data formats used in machine learning, *see* YASER S.ABU-MOSTAFA, MALIK MAGDON-ISMAIL, & HSUAN-TIEN LIN, LEARNING FROM DATA: A SHORT COURSE 1–3 (2012); TREVOR HASTIE, ROBERT TIBSHIRANI, & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING: DATA MINING, INFERENCE, AND PREDICTION 1-6 (2nd ed. 2009). KEVIN P. MURPHY, PROBABILISTIC MACHINE LEARNING: AN INTRODUCTION (2022). *See infra* Part I.A.2 (defining models and algorithms in machine learning).

For example, text data are often represented as word embeddings. Instead of representing the word "cat" as the letters "c", "a", and "t", a machine-learning model might represent "cat" as a list of numbers. An advantage of this numerical representation is that, instead of being similar to "car" (because they have nearly the same letters), the embedding for "cat" can be similar (i.e., their numbers will not differ by much) to the embedding for "kitten" because they have similar meanings.[10] A useful intuition is that, for some word embeddings, you can do operations on them: you can take the word embedding for "king" (a list of numbers), subtract the word embedding representing "man," add the word embedding representing "woman," and get something very close to the word embedding for "queen."[11]

[10] More precisely, the embedding strategy described here attempts to capture semantic similarity, in which a distance metric on the vector space of embeddings of words reflects underlying semantic distance between the concepts words represent. *See* Murphy, *supra* note 12, at 26 (providing a short definition of word embeddings); *Id.* at 703–10 (providing a summary of different types of popular word embeddings); Vicki Boykis, What are embeddings? (June 2023) (unpublished manuscript), https://github.com/veekaybee/what_are_embeddings (for an accessible treatment of the history of embeddings and discussion in relation to modern-day generative-AI models); Tomas Mikolov, Kai Chen, Greg Corrado & Jeffrey Dean, *Efficient Estimation of Word Representations in Vector Space*, *in* 2013 INT'L CONF. ON LEARNING REPRESENTATIONS (2013) (discussing word2vec, a common neural-network-based approach for producing embeddings); Tomas Mikolov, Ilya Sutskever, Kai Chen, et al., *Distributed Representations of Words and Phrases and their Compositionality*, *in* 26 ADVANCES NEURAL INFO. PROCESSING SYS. (2013) (for influential follow-on work to word2vec).

[11] This example should not be taken too generally, as it does not always extend to other cases. *See* Ekaterina Vylomova, Laura Rimell, Trevor Cohn & Timothy Baldwin, *Take and Took, Gaggle and Goose, Book and Read: Evaluating the Utility of Vector Differences for Lexical Relation Learning* 1671, *in* 1 PROC. 54TH ANN. MEETING ASS'N FOR COMPUT. LINGUISTICS 1671 (2016). There are many ways to compute word embeddings. A common embedding strategy that quantifies word importance involves computing word frequency (term frequency, TF) for a particular document in a corpus, and scaling it by word rarity (inverse document frequency, or IDF) across documents in the corpus. For more on TD-IDF, *see generally* Karen Sparck Jones, *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, 1988 DOCUMENT RETRIEVAL SYS. 132; Gerard Salton & Christopher Buckley, *Term-weighting approaches in automatic text retrieval*, 24 INFO. PROCESSING & MGMT 513, 516 (1988). By relying strictly on frequencies, this type of embedding does not capture any semantic information in the encoded words. More sophisticated techniques involve learning word embeddings from data. For example, the BERT language model uses deep learning and a transformer architecture to encode word embeddings. *See generally* Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, *in* 1 PROC. 2019 CONF. N. AMERICAN CHAPTER ASS'N FOR COMPUT. LINGUISTICS: HUM. LANGUAGE TECHS. 4171 (2019). *See infra* Part I.B.3.a (discussing the transformer architecture).

Data are not identical to the things they reflect. A census record is not a household; the word embedding of "cat" is not a cat.[12] But data capture useful information about those things—and they enable computers to reason about them. For our purposes, a creative work like a painting or book is not itself data;[13] rather, it can be processed computationally to be converted into data to be used in machine-learning applications.

2. Machine Learning

An algorithm is a step-by-step procedure—often implemented in software—for performing a computation on data.[14] For example, algorithms include different procedures for sorting a list of numbers, searching for a word in a list of words, encrypting passwords, and much else. Machine learning is a subfield of computing that develops and applies algorithms to learn from data. These algorithms employ mathematical tools from probability and statistics to model (hopefully useful and interesting) patterns in the data. Machine-learning scientists and practitioners may use these algorithms (and their resulting learned patterns) for different aims.

Two overarching types of tasks that machine learning is commonly used for are discriminative [15] and generative [16] modeling. Discriminative modeling includes classification (is this image of a cat or a dog?) and regression (how many ice cream cones can I expect to sell if the weather is 80°F today?).[17] In contrast, generative modeling can produce new content, such as images or text.[18]

---

[12] For a detailed treatment of how data serve as a proxy for entities in the world, *see* DYLAN MULVIN, PROXIES: THE CULTURAL WORK OF STANDING IN 1–33 (1st ed. 2021).

[13] This is not the case in other fields, like archaeology, but is true for machine learning, and thus this Article.

[14] *See generally* THOMAS H. CORMEN, CHARLES E. LEISERSON, RONALD L. RIVEST & CLIFFORD STEIN, INTRODUCTION TO ALGORITHMS (4th ed. 2022).

[15] *See infra* Part I.A.2.a.

[16] *See infra* Part I.A.2.b.

[17] While the examples we provide in the Article concern classification of inputs into discrete output categories, regression tasks that involve real numbers, such as predicting housing price given a set of features, are also discriminative.

[18] This is a simplification that is sufficient for our purposes. Generative modeling does not necessarily produce new content; it estimates probability distributions from which such content can be (but does not have to be) sampled. These probabilities can be useful for applications other than content generation. For example, the BERT language model employs generative techniques and can be used to produce word embeddings, but not content intended to be consumed or enjoyed directly by a human user. *See generally* Devlin, Chang, Lee & Toutanova, *supra* note 14. The distinction between discriminative and generative modeling ultimately hinges on modeling choices regarding the underlying probabilities. *See generally* Dan Y. Rubinstein & Trevor Hastie, *Discriminative Versus Informative Learning*, *in* 1997 PROC. THIRD INT'L CONF. ON KNOWLEDGE DISCOVERY & DATA MINING (1997) (using the term "informative" instead of "generative").

a. Discriminative Modeling

Consider a machine-learning model that classifies images as either cats or dogs. A common analogy in legal literature is to describe a discriminative machine-learning model as a mathematical function that maps inputs to outputs.[19] Here, the model takes a computer-readable version of an image as input[20] and returns a label of either cat or dog as its output. Mathematically, we would say that the set of all inputs (images of pets) is $\mathcal{X}$, the set of possible outputs (the labels cat and dog) is $\mathcal{Y}$, and the function that maps an input to an output is $f : \mathcal{X} \mapsto \mathcal{Y}$. It is an underlying assumption of this analogy that the function f is deterministic: given the same input (a particular picture of a terrier), a given trained classification model will always produce the same output (the label dog).[21]

[19] For example, the function $f(x) = x + 1$ produces an output that is 1 greater than its input. If the input to this function is 5, the output is 6. If the input is 8, the output is 9. *See generally* A. Feder Cooper, Jonathan Frankle & Christopher De Sa, *Non-Determinism and the Lawlessness of Machine Learning Code*, *in* 2022 PROC. 2022 SYMPOSIUM ON COMPUT. SCI. & L. 1 (2022) (discussing the prevalence of the "function" view).

[20] *E.g.*, two-dimensional images can be saved as a set of numbers, which can serve as the input to a machine-learning algorithm. Often, images are formatted as a **matrix** representing pixels, where each pixel is a vector of numbers in the range 0-255 that represents combinations of *r*ed, *b*lue, and *g*reen (RGB) hues.

[21] *See generally* Cooper, Frankle & De Sa, *supra* note 22 (discussing this assumption in the legal literature on machine learning). Classifiers like these actually output probabilities (in this case, the probability that an input image should be labeled dog), which can then be converted to an output label. This typically involves using a classification rule, e.g., "if the probability of dog is greater than 0.5, then output dog, otherwise output cat." In this example of an image of a terrier, we assume that the model classifies the image correctly. It is also possible that the classifier does not, and instead always returns the label cat. Further, note that, in general, the input image could be of anything. Performing classification involves manipulating numbers under the hood—typically, linear algebra operations on vectors and matrices that contain the model parameters and the new input data example. So, one could provide, for example, an image of an airplane as input, and the model would still output a probability (that can be converted to a classification) of either cat or dog.

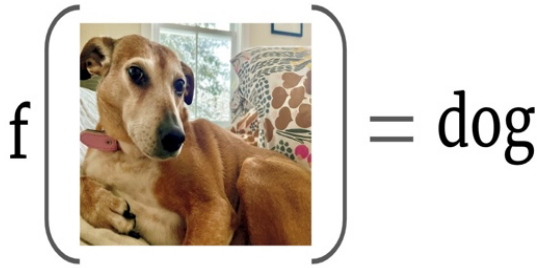


Figure 2: Depicting the analogy of a machine-learned model as a function, where a classifier f takes an image x as input and returns the class label y= dog. (Image: "Arabela, The Venus of Evanston." Source: Dr. Fernando Delgado, reprinted with permission.)

To produce such a model, one chooses a training algorithm that takes training data examples as input and produces a model as its output. It is important to emphasize that the training algorithm implements a *different* function than the classification function described above. The input to a training algorithm for the cat-dog classifier is a training dataset (numerous images of pets, each labeled cat or dog), not just a *single* data example (an image of a terrier). The output from this training algorithm is a model—the function that classifies an individual image as a cat or dog. The model, like its data inputs, is ultimately also some arrangement of numbers that are interpretable by a computer. For our purposes, there are two important concepts to understand about the numbers that comprise models: how they are organized and what they contain.

First, start with organization. There are many kinds of models (even for a simple cat-dog classification task), and each model has its own model architecture. Very roughly, a model architecture specifies *how* a model stores information about patterns in data: how much information, how it is arranged, and how it can be manipulated with code. For example, one could use the same model architecture to make a cat-dog classifier or a bird-airplane classifier, and they would use the same training algorithm. The difference is that the bird-airplane classifier would be trained on a training dataset of pictures of things in the sky, each labeled with bird or airplane.

An important class of model architectures is neural networks, which consist of interconnected nodes (also called neurons). There are many different neural-network architectures, which differ in things such as how many nodes there are (organized in how many layers), how long each list of nodes is (the size of each

layer), and how the nodes are connected to each other. And even for a given model architecture, there can be many different training algorithms.[22] It is a major area of machine-learning research to determine what model architectures, training algorithms, and training datasets are most effective for producing useful models.

Second, the contents. Noted above, the nodes in a specific model architecture are *interconnected*. Each of these connections has an associated parameter (also called weight). These parameters are lists of numbers that represent the strengths of connections between nodes in the network. The specific numbers that they contain come from training:[23] They are the particular patterns learned from running a specific training algorithm on a specific training dataset. This model training typically involves running an optimization-based routine, which iteratively processes the input data to update (i.e., train) the model parameters.[24] (A very loose analogy can be drawn to the way that neurons in the brain are connected to each other with different strengths to encode a human-learned representation of the world.)

Different model architectures vary widely in size and complexity and, in turn, have different capabilities for encoding learned relationships in the data. Simpler, more traditional statistical models like logistic regression have relatively few learned parameters, while modern-day deep neural networks can have *billions* (or

[22] Model architectures and training algorithms also include **hyperparameters**. Hyperparameters are parameters that traditionally are not learned; they are often set by a human. For the model, they can dictate the number of parameters and layers. For the training algorithm, they dictate properties of how training is run. For example, a hyperparameter called the "learning rate" determines how fast or slow model training should proceed. *See* A. Feder Cooper, Yucheng Lu, Jessica Zosa Forde & Christopher De Sa, *Hyperparameter Optimization Is Deceiving Us, and How to Stop It*, *in* 34 ADVANCES NEURAL INFO. PROCESSING SYS. (2021) (regarding the effects of hyperparameter choices on resulting learned models, and citations therein).

[23] There are other inputs that configure the training algorithm, and which affect the parameters that are learned. *See id*. and accompanying text (for an example of such configurations).

[24] There are many different optimization methods used in deep learning. *See generally* Robin M. Schmidt, Frank Schneider & Philipp Hennig, *Descending through a Crowded Valley - Benchmarking Deep Learning Optimizers*, *in* 139 PROC. 38TH INT'L CONF. ON MACH. LEARNING 9367–76 (2021). The most common is an optimization method called Adam (and variants thereof). *See generally* Diederik P. Kingma & Jimmy Lei Ba, *Adam: A Method for Stochastic Optimization*, *in* 2015 INT'L CONF. ON LEARNING REPRESENTATIONS (2015). However, optimization algorithms for machine learning, and for training generative-AI models, remain an active area of research. *See, e.g.*, Pierre Foret, Ariel Kleiner, Hossein Mobahi & Behnam Neyshabur, *Sharpness-aware Minimization for Efficiently Improving Generalization*, *in* 2021 INT'L CONF. ON LEARNING REPRESENTATIONS (2021); Dara Bahri, Hossein Mobahi & Yi Tay, *Sharpness-Aware Minimization Improves Language Model Generalization*, *in* 2022 PROC. 60TH ANN. MEETING ASS'N FOR COMPUT. LINGUISTICS (VOLUME 1: LONG PAPERS) 7360–71 (2022).

even *trillions*) of parameters.[25] At each step, the training algorithm adjusts the values of the model's parameters so that they are slightly better at describing the training data; with enough iterations on enough of the right kind of data, the model can describe patterns that are present in many training examples. After training is complete, we can evaluate the resulting model by running it on new (previously unseen) data examples and seeing how well it classifies them as either cat or dog.[26]

The above describes a sketch of machine learning that is familiar in legal scholarship. This work has scrutinized the implications of machine-learning-based decision-making in a variety of areas, such as whether or not to interview or hire a job candidate, grant an applicant a loan,[27] or, as in the case of the infamous Northpointe COMPAS system, to predict prison recidivism.[28] These types of yes/no decision-making tasks generally fall under the heading of discriminative machine learning, a type of machine learning that attempts to draw boundaries in available data, and that is often used for making predictions. As we stated at the beginning of this section, discriminative machine-learning tasks typically involve classification or regression.

---

[25] Consider three examples. First, PaLM, a language model built by Google, has 540 billion parameters. Aakanksha Chowdhery, Sharan Narang, Jacob Devlin et al., *PaLM: Scaling Language Modeling with Pathways*, 24 J. MACH. LEARNING RSCH. 1−113 (2023). Second, the largest Llama 2 model, a semi-closed model released by Meta, has 70 billion parameters. (Llama 2 is a **family** of models that come in different sizes. It is common today for open- and semi-closed models to come in a variety of differently sized architectures, with larger models tending to produce higher quality generations for a larger cost in computational resources (i.e., **compute**). Llama 2 is just one example of such a model family.) *See supra* note 7 and accompanying text (regarding Llama 2 and the distinction between open- and semi-closed models). *See infra* Part I.B.4 (discussing the importance of scale). Third, GLM-130B, a bilingual Chinese and English model, has 130 billion parameters. Aohan Zeng, Xiao Liu, Zhengxiao Du et al., GLM-130B: An Open Bilingual Pre-trained Model (Oct. 5, 2022) (unpublished manuscript), https://arxiv.org/abs/2210.02414. When discussing **model size**, it is typical to do so with respect to parameters (i.e., connections between nodes); there are typically many more connections between nodes than there are nodes.

[26] To do valid and reliable model evaluation, it is important to run the model on a **test dataset**. Test datasets are made up of reserved data examples that are not a part of training. They are ostensibly from the same *distribution* as the training data but, since the model has not seen them before being evaluated, these exact examples have not influenced the learned relationships in the model's parameters. *See* Abu-Mostafa, *supra* note 12, at 39–69.

[27] Danielle Keats Citron & Frank A. Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 655 (2014).

[28] *See generally* Jeff Larson, Surya Mattu, Lauren Kirchner & Julia Angwin, *How We Analyzed the COMPAS Recidivism Algorithm*, PROPUBLICA (May 16, 2016), https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm/ (for the original study indicating algorithmic bias in this system).

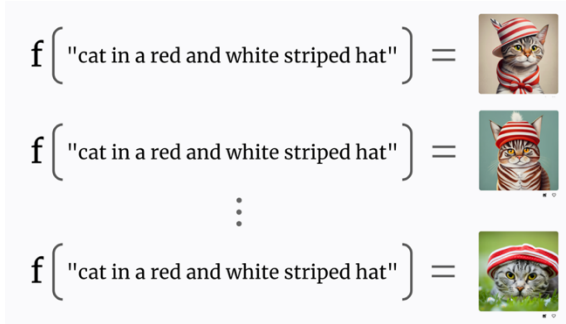


Figure 3: Images of "cat in a red and white striped hat" generated by the authors with Ideogram (Ideogram.AI, 2023, https://ideogram.ai/). Running the model (f) multiple times on the same input can generate different outputs.

b. Generative Modeling

Discriminative tasks are only one type of machine-learning modeling. Another paradigm is called generative machine learning. [29] Whereas discriminative machine-learning models return a *single*[30] output y from a set of

[29] Deep generative models, such as OpenAI's CLIP, Midjourney, or Stability AI's Stable Diffusion, are not the only form of generative machine learning. Generative machine learning is often subdivided into probabilistic graphical models and deep generative models. *See generally* OpenAI, *CLIP: Connecting text and images*, OpenAI (Jan. 5, 2021), https://openai.com/research/clip (regarding OpenAI's CLIP model). *See generally* Midjourney, https://midjourney.com/ (regarding Midjourney). *See generally Stable Diffusion XL*, Stability AI (2023), https://stability.ai/stablediffusion; Robin Rombach, Andreas Blattmann, Dominik Lorenz et al., *High-Resolution Image Synthesis with Latent Diffusion Models*, *in* 2022 IEEE Conf. on Comput. Vision & Pattern Recognition (2022) (regarding Stable Diffusion). *See generally* Daphne Koller & Nir Friedman, Probabilistic Graphical Models: Principles and Techniques (2009) (for a canonical textbook treatment on probabilistic graphical models). *See generally* Jakub M. Tomczak, Deep Generative Modeling (2022) (for details on different techniques for generative modeling in machine learning).

[30] These single outputs can nevertheless have differing degrees of uncertainty associated with them. *See generally* A. Feder Cooper, Katherine Lee, Madiha Zahrah Choksi et al., *Arbitrariness and Prediction: The Confounding Role of Variance in Fair Classification*, *in* 2024 Proc. 38th Ann. AAAI Conf. on A.I. (2024) (for an accessible treatment of this topic).

possible outputs $\mathcal{Y}$ (Figure 2),[31] generative machine-learning models have *multiple possible reasonable outputs* for a given input. For example, many reasonable images match the caption: "cat in a red and white striped hat" (Figure 3). Similarly, a generative model for text could have many reasonable completions to the following sentence: "In the summer, I like to go to the [blank]", such as: "beach", "park", "pool", or "mountains" filling in the "[blank]".

This example shows how the analogy of machine learning as a function, which provides a useful intuition for discriminative modeling, does not cleanly extend to generative modeling. Instead of a single output y for a given input x, for generative modeling there are usually many reasonable outputs for a given input. For a model, choosing which among these possible outputs to produce involves some *randomness*, which means a model can generate different outputs even when run on the same input.

In more detail, generative models learn from their training data which outputs are more likely. In the sentence "In the summer, I like to go to the [blank]", the word "beach" is a more likely completion than "slopes". While the words "summer" and "beach" are often associated in writing (and thus also in the training data used to produce the model), this is not the case for "summer" and "slopes".[32] But "beach" and "pool" might be equally likely. So, the model's choice between "beach" and "pool" is made with some degree of randomness.[33]

B. Generative AI

Most commonly, the term "generative AI" is used to describe systems that can take in a variety of inputs—typically expressive content (e.g., a piece of text or an image)—and can produce expressive content as their outputs. The inputs are often user-generated, though they do not have to be.[34] This is why a user of an

---

[31] In the running classification example above, every input image must be labeled with either y = cat or y = dog.

[32] The model captures the *conditional probability* of the next word x *given* having already seen a prior sequence of words a. In the above example, we would consider the probability of the next word being x = "beach" given that a = "In the summer, I like to go to the".

[33] Discriminative and generative modeling can be related to each other mathematically. Under the hood, both approaches model conditional probabilities, but this observation gets abstracted away in typical discussions that analogize discriminative models to functions. *See generally* Rubinstein & Hastie, *supra* note 21. *See also* Andrew Y. Ng & Michael I. Jordan, *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes* 2, *in* 14 ADVANCES NEURAL INFO. PROCESSING SYS. (2001) (describing how the two approaches can be related to each other using Bayes' rule).

[34] Synthetic data, rather than user-generated data, can also be supplied as inputs, both for prompting and as training data. Producing and leveraging synthetic data (that has been produced by generative AI) for prompting and training is an active area of machine-learning research. *See, e.g.,* Aaron Gokaslan, A. Feder Cooper, Jasmine Collins et al.,

application like ChatGPT or DreamStudio is said to provide a prompt. The output that the application produces in response is called a generation. It is also helpful to distinguish training time (the process of training a model) from generation time (the process of using a model to generate an output).[35]

Generative AI builds on and extends traditional generative machine learning, but it adds several new elements that have contributed to its immense popularity and accompanying controversy. In the remainder of this section, we unpack four ways that generative AI is different and new. First, contemporary generative-AI *systems* often use *multiple models*, which have different architectures and use different approaches to perform different functions.[36] For this reason, it is often more helpful to think about the design and behavior of an overall system, rather than focusing entirely on a specific model within it. Second, these systems are often multimodal; they can work with data in different formats (such as a system

CommonCanvas: Open Diffusion Models Trained with Creative-Commons Images, *in* 2024 PROC. OF THE IEEE/CVF CONF. ON COMPUT. VISION PATTERN RECOGNITION (CVPR) 8250 (2024) (using generative-AI-produced text captions to train a text-to-image diffusion model); Liang Wang, Nan Yang, Xiaolong Huang et al., Improving Text Embeddings with Large Language Models (Dec. 31, 2023) (unpublished manuscript), https://arxiv.org/abs/2401.00368 (producing text embeddings using only synthetic data); Avi Singh, John D. Co-Reyes, Rishabh Agarwal et al., Beyond Human Data: Scaling Self-Training for Problem-Solving with Language Models (Dec. 11, 2023) (unpublished manuscript), https://arxiv.org/abs/2312.06585 (generating synthetic fine-tuning data); *See infra* Part I.C.7 (for more discussion on synthetic data).

[35] The amount of compute at training time has traditionally been the emphasis when discussing the scale of resources used to produce contemporary generative-AI models (particularly language models). Recently, research and products have begun to experiment with also using more compute at generation time—also called **test time**—to produce higher quality generations. Most notably, OpenAI's recent o1 model explores new training-time and test-time compute paradigms to improve "complex reasoning." OpenAI, *Learning to Reason with LLMs*, OPENAI (September 12, 2024), https://openai.com/index/learning-to-reason-with-llms/ ("Our large-scale reinforcement learning algorithm teaches the model how to think productively using its chain of thought in a highly data-efficient training process. We have found that the performance of o1 consistently improves with more reinforcement learning (train-time compute) and with more time spent thinking (test-time compute). The constraints on scaling this approach differ substantially from those of LLM pretraining, and we are continuing to investigate them."). *See infra* note 39 and accompanying text (discussing reinforcement learning); *infra* note 208 and accompanying text (discussing chain of thought).

[36] Historically, practitioners typically would have chosen to solve a particular problem with a particular modeling technique. For example, they would take either a discriminative or generative modeling approach, or use another modeling paradigm called **reinforcement learning**. Abu-Mostafa, *supra* note 12, at 11–14; Murphy, *supra* note 12, at 1–19 (for an intuition behind reinforcement learning). We introduce this concept in more detail when we discuss **model alignment** in the generative-AI supply chain. *See infra* Part I.C.8. Generative AI can involve all of these approaches. Reinforcement learning also plays an enormous role at both training time and test time in OpenAI's o1 model. *See id.*

that takes *text* as an input but produces *images* as outputs). Third, recent technological developments in the transformer architecture (often used for text) and diffusion models (often used for images) have significantly contributed to the power of contemporary generative systems. And fourth, there have been immense leaps in scale. Generative AI involves large-scale systems and the training of massive models on similarly massive datasets. Scale stands on its own as another reason why generative AI is different from more traditional generative modeling.

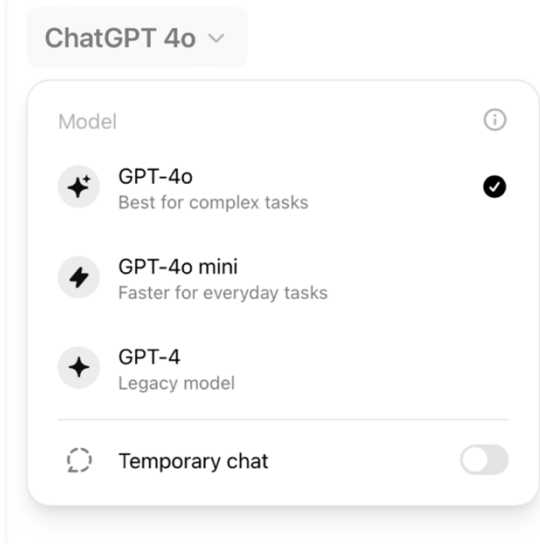


Figure 4: ChatGPT user interface showing a choice of accessing three underlying ChatGPT models: GPT-4o, GPT-4o mini, and GPT-4. (Screenshot produced by the authors, prior to the release of the o1 model.)

1. Generative-AI Systems

Most users of generative AI do not interact with a model directly. Instead, they use an interface to a system, in which the model is just one of several embedded, interoperating components.[37] For example, OpenAI hosts various ways to access its latest models. ChatGPT is a user interface (and the name of a system) for accessing several different underlying models (Figure 4). OpenAI also has a developer API, which serves as an interface for programmers to access

[37] *See generally* A. Feder Cooper & Karen Levy, *Fast or Accurate? Governing Conflicting Goals in Highly Autonomous Vehicles*, 20 COLO. TECH. L.J. 249 (2022). *See* A. Feder Cooper, Karen Levy & Christopher De Sa, *Accuracy-Efficiency Trade-Offs and Accountability in Distributed ML Systems*, *in* 2021 EQUITY & ACCESS ALGORITHMS MECHANISMS & OPTIMIZATION 1, 1–2 (2021) (discussing the importance of such a systems framing in contemporary computing applications). OpenAI also emphasizes this point in their policy research. For example, OpenAI produced a GPT-4 *system* card (emphasis added), and this point was made at the GenLaw 2023 workshop by Miles Brundage in his talk "Where and When Does the Law Fit into AI Development and Deployment?" OpenAI, GPT-4 System Card (Mar. 23, 2023) (unpublished manuscript), https://CDN.openai.com/papers/gpt-4-system-card.pdf (emphasizing systems, which contain models and other components).

OpenAI's different models using code. There are additional components behind each of these interfaces, including possibly (according to rumor) as many as sixteen GPT-4 models, to which different prompts are routed.[38] As another example, consider Stable Diffusion, an open-source model for producing image generations.[39] Most users do not typically interact directly with the Stable Diffusion model;[40] rather, they access a version that is embedded in a larger system operated by Stability AI,[41] which has multiple components, including a web-based application called DreamStudio.[42]

2. Generation Modalities

The input and output content types for generative-AI models are often referred to as modalities. For example, a chatbot that produces *text* generations when given a user-provided *text* prompt would use an underlying text-to-text model; this model operates in the text modality. Such a chatbot uses the same modality for the input and output, but this is not a requirement for generative AI more broadly. Many image-generation models (used in systems like Stable Diffusion,[43] DALL·E 2,[44] and ChatGPT Plus and Enterprise,[45] etc.) take a text description as input and produce an image generation as output. These models are multimodal, text-to-image models.

While we focus on generative-AI systems that involve text and image inputs and outputs, there are many other modalities that generative AI can be applied to, such as computer code, audio (music, voice), video, and molecular structures. Text-to-code models, which are designed specifically to take in natural language

[38] This rumor originated in a Twitter (now X) post. Maximilian Schreiner, *GPT-4 architecture, datasets, costs and more leaked*, THE DECODER (July 11, 2023), https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.
[39] Rombach, Blattmann, Lorenz et al., *supra* note 32.
[40] At a minimum, using the model directly would involve downloading the model parameters, writing code to run the model, and executing that code.
[41] *Stable Diffusion XL*, *supra* note 32.
[42] *See generally DreamStudio*, STABILITY AI (2023), https://stability.ai/stablediffusion.
[43] *See* Rombach, Blattmann, Lorenz et al., *supra* note 32 (describing the model); *Stable Diffusion XL*, *supra* note 32 (describing the product).
[44] *See* Aditya Ramesh, Prafulla Dhariwal, Alex Nichol et al., Hierarchical Text-Conditional Image Generation with CLIP Latents (Apr. 13, 2022) (unpublished manuscript), https://arxiv.org/abs/2204.06125 (describing the model); *DALL·E 2*, OPENAI (2022), https://openai.com/dall-e-2 (describing the product).
[45] *See DALL·E 3 is now available in ChatGPT Plus and Enterprise*, OPENAI (Oct. 19, 2023), https://openai.com/blog/dall-e-3-is-now-available-in-chatgpt-plus-and-enterprise (announcing the integration of DALL·E 3 text-to-image functionality into the paid versions of the ChatGPT chatbot system).

as input and generate code snippets as output,[46] include OpenAI Codex[47] and Code Llama[48] from Meta.[49] Notably, Codex is the generative-AI model embedded in the GitHub Copilot system,[50] which is named in active lawsuits regarding copyright infringement.[51] Google DeepMind's Lyria,[52] Google's MusicLM,[53] and OpenAI's Jukebox[54] are music-generation models embedded in larger systems.[55] OpenAI's website claims, "Provided with genre, artist, and

[46] ChatGPT, for example, can also produce code snippets, but is a chatbot system that also has other functionality. *See* OpenAI, *supra* note 7.

[47] Wojciech Zaremba, Greg Brockman & OpenAI, *OpenAI Codex*, OPENAI (Aug. 10, 2021), https://openai.com/blog/openai-codex (describing the Codex model). OpenAI, *Powering next generation applications with OpenAI Codex*, OPENAI (May 24, 2022), https://openai.com/blog/codex-apps (discussing applications using Codex). Mark Chen, Jerry Tworek, Heewoo Jun et al., Evaluating Large Language Models Trained on Code (July 7, 2021) (unpublished manuscript), https://arxiv.org/abs/2107.03374 (for the technical report detailing the original Codex model).

[48] Meta, *Introducing Code Llama, an AI Tool for Coding*, META NEWS (Aug. 24, 2023), https://about.fb.com/news/2023/08/code-llama-ai-for-coding/ (announcing Code Llama). Meta, *Introducing Code Llama, a state-of-the-art large language model for coding*, META RSCH. BLOG (Aug. 24, 2023), https://ai.meta.com/blog/code-llama-largelanguage-model-coding/ (describing Code Llama in a technical blog post). Baptiste Rozière, Jonas Gehring, Fabian Gloeckle et al., Code Llama: Open Foundation Models for Code (Aug. 24, 2023) (unpublished manuscript), https://arxiv.org/abs/2308.12950 (for the technical report detailing the Code Llama model).

[49] Both of these models use transformer-based architectures. *See infra* Part I.B.3a.

[50] *See generally GitHub Copilot documentation*, GITHUB (Aug. 28, 2023), https://docs.github.com/en/copilot.

[51] *See generally* Complaint, Doe 1 v. GitHub, Inc., No. 4:22-cv-06823 (N.D. Cal. Nov. 3, 2022). GitHub has since updated the Copilot model to go "beyond the previous OpenAI Codex model." However, the original Codex model is the one named in active lawsuits. *See generally* Shuyin Zhao, *Smarter, more efficient coding: GitHub Copilot goes beyond Codex with improved AI model*, GITHUB (July 28, 2023), https://github.blog/2023-07-28-smarter-more-efficient-coding-github-copilot-goes-beyond-codex-with-improved-ai-model/ (discussing Copilot's use of Codex).

[52] *Transforming the future of music creation*, GOOGLE DEEPMIND (Nov. 16, 2023), https://deepmind.google/discover/blog/transforming-the-future-of-music-creation/.

[53] *See* Andrea Agostinelli, Timo I. Denk, Zalán Borsos et al., MusicLM: Generating Music From Text (Jan. 26, 2023) (unpublished manuscript), https://arxiv.org/abs/2301.11325 (for the research paper on the model); Kristin Yim & Hema Manickavasagam, *Turn ideas into music with MusicLM*, GOOGLE (May 10, 2023), https://blog.google/technology/ai/musiclm-google-ai-test-kitchen/ (for the product announcement).

[54] Heewoo Jun, Christine Payne, Jong Wook Kim et al., Jukebox: A Generative Model for Music (Apr. 30, 2020) (unpublished manuscript), https://arxiv.org/abs/2005.00341.

[55] For example, Dream Track is a production system built using Lyria. *See* Google DeepMind, *supra* note 55.

lyrics as input, Jukebox outputs a new music sample produced from scratch."[56] Pika is an "idea-to-video platform" that provides tools to produce video generations using models that take in either text or image prompts.[57] Lastly, generative-AI models for molecular structure are intended for many different purposes, including to aid in the design of new drugs and to understand protein function. Examples of models in this domain include ProtGPT2[58] and DiffDock.[59] While these modalities also have important implications for copyright,[60] we limit our discussion and examples in the remainder of this Article to text and images.

a. Text Data and Generations

OpenAI's ChatGPT is a system that takes in text inputs and produces text outputs (among other modalities). ChatGPT is built on top of multiple models, including several different text-to-text model architectures trained on massive amounts of text data.[61] During training, each of these text-to-text models is shown text sequences, and for every sequence, it is trained to predict the next word given all the previous words in the sequence.

For example, if the sentence "In the summer, I like to go to the beach" were in the training data, then the model would first be shown "In" and trained to predict "the", then given "In the" and trained to predict "summer", and so on.[62]

---

[56] *See* OpenAI, *OpenAI JukeBox*, OPENAI (Apr. 30, 2020), https://openai.com/research/jukebox (describing the use of the transformer-based architecture in Jukebox).

[57] *See* Pika, *An idea-to-video platform that brings your creativity to motion*, PIKA (2023), https://pika.art/.

[58] *See generally* Noellia Ferruz, Steffen Schmidt & Birte Höcker, *ProtGPT2 is a deep unsupervised language model for protein design*, 13 NATURE COMMC'NS 4348 (2022). ProtGPT2 is based on GPT-2. *See generally* Radford, Wu, Child et al., *infra* note 70. (describing GPT-2, a language model with a transformer-based architecture).

[59] *See generally* Gabriele Corso, Hannes Stärk, Bowen Jing et al., *DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking*, *in* 2023 INT'L CONF. ON LEARNING REPRESENTATIONS (2023). DiffDock uses diffusion-based techniques. *See infra* Part I.B.3.b.

[60] And perhaps also patent law, for generative-AI systems that involve molecular structure.

[61] So far, the most notable of these models are OpenAI's GPT family of text-to-text models, including GPT-3.5, GPT-4, GPT-4o, and GPT-4o mini. Recent o-family models include o1 and o1-mini. OpenAI, *supra* note 7.

[62] Note that illustrating this prediction one word at a time is a simplification. It is common to train models on **tokens**, not words. Tokens are numbers that represent a word, sub-word, logogram, or punctuation mark. For instance, the word "hello" may be represented by the token-ID number 12. A more uncommon word like "credenza" may be divided into multiple sub-words, e.g., "cre", "den", "za"; each sub-word would be represented by a number, e.g., "cre" = 58, "den" = 29, "za" = 105), and so, altogether, the word "credenza" would be encoded as the vector [58,29,105]. Modeling data as tokens enables using transformers with non-text sequences, e.g., a token for a music model may be a musical

Text data are, in many ways, easier to collect than data in other modalities because text is so readily available on the Internet. Common data sources include data scraped from the web,[63] books (both copyrighted and in the public domain), and news articles,[64] as well as data obtained through user interactions.[65] Web data may include structured text like product reviews, and free-form social media posts and blogs.[66]

It is important to note that generative text models are used extensively beyond chatbot systems like ChatGPT.[67] For example, they also play an important role in

note or a specific pitch. In our example above, the sentence could be **tokenized** as the list ["In", "the", "summer", ",", "I", "like", "to", "go", "to", "the", "beach"], where each word (and the comma) corresponds to a token (that has an associated token ID). But a different tokenizer could, hypothetically, tokenize the word "summer" as two tokens, e.g., "sum" and "mer". The prediction of the next token would instead operate over a list that contains these tokens, rather than "summer".

[63] Web scraping involves using a computer to download and store the contents of a webpage. This is not a new phenomenon; businesses have been scraping data for decades. *See, e.g.*, hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180 (9th Cir. 2022) (discussing legality of scraping data from LinkedIn).

[64] *See* Katherine Lee, Daphne Ippolito & A. Feder Cooper, The Devil is in the Training Data (2023) (unpublished manuscript), *in* Katherine Lee, A. Feder Cooper, James Grimmelmann & Daphne Ippolito, AI and Law: The Next Generation 5 (2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4580739 (discussing training data sources). *See generally* Tom B. Brown, Benjamin Mann, Nick Ryder et al., Language Models are Few-Shot Learners (2020) (unpublished manuscript), https://arxiv.org/abs/2005.14165; Leo Gao, Stella Biderman, Sid Black et al., The Pile: An 800GB Dataset of Diverse Text for Language Modeling (Dec. 31, 2020) (unpublished manuscript), https://arxiv.org/abs/2101.00027; Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee et al., *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, 21 J. MACH. LEARNING RSCH. 1 (2020).

[65] For example, it is widely believed (though unconfirmed) that user data ingested by the ChatGPT interface are used to train the underlying model(s). *See New Ways to Manage Your Data in ChatGPT*, OPENAI (2023), https://openai.com/blog/new-ways-tomanage-your-data-in-chatgpt (describing only the cases in which user data are *not* used to train the ChatGPT system).

[66] *See* Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the secret list of websites that make AI like ChatGPT sound smart*, WASHINGTON POST (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/. *See generally* Jesse Dodge, Maarten Sap, Ana Marasović et al., *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*, *in* 2021 PROC. 2021 CONF. ON EMPIRICAL METHODS NAT. LANGUAGE Processing 1286 (2021).

[67] *See* Alec Radford, Karthik Narasimhan, Tim Salimans & Ilya Sutskever, Improving Language Understanding by Generative Pre-training (2018) (unpublished manuscript), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf; Alec Radford, Jeffrey Wu, Rewon Child et al., Language Models are Unsupervised Multitask Learners (2019) (unpublished

translation systems.[68] The training data for these different types of applications tend to differ according to use case, e.g., translation-model training datasets include text from multiple languages and chat-model training datasets include dialog.[69]

b. Image Data and Generations

Multimodal text-to-image[70] systems include DALL·E,[71] DALL·E 2,[72] DALL·E 3[73] (which is embedded within ChatGPT[74]), Ideogram,[75] Midjourney,[76] and Stability AI's DreamStudio.[77] The generative-AI models in these systems are

manuscript), https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf; Raffel, Shazeer, Roberts, Lee et al., *supra* note 67; Devlin, Chang, Lee & Toutanova, *supra* note 14 (which all use generative text models to perform a variety of text-based tasks including translation, question answering, summarization, and text classification).

[68] Google Translate uses generative AI to produce translated text given an input in another language. *See* Isaac Caswell & Bowen Liang, *Recent Advances in Google Translate*, GOOGLE RSCH. (June 8, 2020), https://research.google/blog/recent-advances-in-google-translate/ (describing the Google Translate system in 2020, which uses a transformer model in conjunction with another type of model called a Recurrent Neural Network). *See infra* Part I.B.3.a (discussing transformer models).

[69] *See generally* Romal Thoppilan, Daniel De Freitas, Jamie Hall et al., LaMDA: Language Models for Dialog Applications (2022) (unpublished manuscript), https://arxiv.org/pdf/2201.08239.pdf (discussing the inclusion of dialogue in the training of a chat model).

[70] There are also unimodal image-to-image models and systems, like the one owned and operated by Runway. *See* Runway, *Image to Image*, RUNWAY (2023), https://runwayml.com/aimagic-tools/image-to-image/.

[71] *See generally* Aditya Ramesh, Mikhail Pavlov, Gabriel Goh et al., *Zero-Shot Text-to-Image Generation*, *in* 2021 PROC. 38TH INT'L CONF. ON MACH. LEARNING 8821 (2021) (the original DALL·E model paper); Alec Radford, Jong Wook Kim, Chris Hallacy et al., *Learning Transferable Visual Models From Natural Language Supervision*, *in* 2021 PROC. 38TH INT'L CONF. ON MACH. LEARNING 8748 (2021) (the critic model used to rank DALL·E-generated outputs for a given prompt). Both components are part of the OpenAI DALL·E system. *See generally* OpenAI, *DALL·E: Creating images from text*, OPENAI (Jan. 5, 2021), https://openai.com/research/dall-e.

[72] *See generally* Ramesh, Dhariwal, Nichol et al., *supra* note 47 (the original DALL·E 2 research paper); *DALL·E 2*, *supra* note 47 (the DALL·E 2 OpenAI system).

[73] *See generally* James Betker, Gabriel Goh, Li Jing et al., Improving Image Generation with Better Captions (2023) (unpublished manuscript), https://cdn.openai.com/papers/dall-e-3.pdf (the original DALL·E 3 research paper); OpenAI, *DALL·E 3*, OPENAI (2023), https://openai.com/dall-e-3 (describing the functionality of DALL·E 3 in OpenAI products).

[74] *DALL·E 3 is now available in ChatGPT Plus and Enterprise*, *supra* note 48.

[75] *See generally Ideogram.AI*, IDEOGRAM.AI (2023), https://ideogram.ai/.

[76] MIDJOURNEY, *supra* note 32.

[77] *DreamStudio*, *supra* note 45.

trained on huge amounts of image-text pairs, in which each image is paired with a caption describing it (e.g., a picture of a cat lounging in a sunbeam with the caption "a black cat underneath a window").[78] These datasets are also often scraped from the Internet, and can include both copyrighted and public-domain images and captions.[79] In some cases, only the images are scraped from the Internet, and the corresponding captions are produced using machine learning—for example, using an image-to-text generative-AI model or system to produce synthetic captions.[80]

Trained models leverage these learned relationships at generation time: when supplied with text prompts as inputs, they generate image outputs to match the prompt.[81] Today's text-to-image models can produce generations that span a variety of artistic styles—from cartoons to photorealistic images—and can

[78] Until recently, one common source of training data was LAION-5B, a dataset constructed from images and alt-text from the Common Crawl corpus. *See generally* Romain Beaumont, *LAION-5B: A New Era of Large-Scale Multi-Modal Datasets*, LAION (Mar. 31, 2022), https://laion.ai/blog/laion-5b (describing the LAION-5B dataset). *See generally* Christoph Schuhmann, Romain Beaumont, Richard Vencu, et al., *LAION-5B: An open large-scale dataset for training next generation image-text models*, *in* 2022 THIRTY-SIXTH CONF. ON NEURAL INFO. PROCESSING SYS. DATASETS & BENCHMARKS TRACK (2022) (for the *NeurIPS* conference Datasets Track paper on LAION-5B). *See generally* Dodge, Sap, Marasović et al., *supra* note 69 (regarding the Common Crawl corpus; also see citations therein). LAION-5B has since been removed from HuggingFace and other public hosting services due to identification of CSAM in images at its linked URLs. *See generally* Abeba Birhane, Vinay Uday Prabhu & Emmanuel Kahembwe, Multimodal datasets: misogyny, pornography, and malignant stereotypes (2021) (unpublished manuscript), https://arxiv.org/abs/2110.01963 (for one of the first studies documenting pornography in LAION-linked images). *See generally* Emilia David, *AI image training dataset found to include child sexual abuse imagery*, THE VERGE (Dec. 20, 2023), https://www.theverge.com/2023/12/20/24009418/generativeAI-image-laion-csam-googlestability-stanford (regarding the Stanford study that prompted LAION's removal).

[79] It is possible for one item in the pair to be copyrighted and the other to be in the public domain, such as a copyrighted image with a public-domain (or uncopyrightable) caption. *See infra* Part II.B.

[80] One reason to use synthetic captions is that not every image comes with a caption already. Another is that a computer-generated caption of "a black cat underneath a window" may be more usefully descriptive of the image's contents than the human-generated "Mr. Tickles loves his nappies!" *See generally* Junnan Li, Dongxu Li, Silvio Savarese & Steven Hoi, BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models (2023) (unpublished manuscript), https://arxiv.org/abs/2301.12597 (for the BLIP-2 model, which can be used for synthetic captioning of images); Gokaslan, Cooper, Collins et al., *supra* note 37 (for an example application of caption generation using BLIP-2).

[81] Of course, many such generations can match the prompt; there are multiple reasonable outputs for the same input. *See supra* Part I.A.2.b. Some generative-AI systems include models that rank match quality. *See generally* Radford, Kim, Hallacy et al., *supra* note 74 (discussing the CLIP-model-based ranking methodology used in DALL·E).

incorporate different abstract concepts and concrete elements. For an example of such a generation, consider Figure 3, which shows different versions of "cat in a red and white striped hat"—different cats, different hats, different compositions, and different artistic styles.

3. Machine-Learning Techniques in Generative AI

While "generative AI" might be a relatively new term-of-art, a lot of the technology that powers today's generative-AI systems has a long history. Many familiar concepts—training algorithms, optimization, neural networks, etc.—all play important roles.[82] In this respect, there is no magic behind generative AI. However, there have been a few especially important technological developments in machine learning over the past decade that have helped usher in this new phase of generative-AI applications with seemingly magical capabilities. In this section, we address two: the transformer architecture[83] and diffusion-based models.[84]

[82] *See supra* Part I.A.2. It is also true that generative models, as an overarching type of machine learning, are also not completely new. *See supra* Part I.A.2.b. Automatic text and music generation date back to the middle of the 20th century. *See generally* Claude E. Shannon, *A Mathematical Theory of Communication*, 27 BELL SYS. TECH. J. (1948). (describing Markov-chain-based language models). *See generally* Darrell Conklin, *Music Generation from Statistical Models*, 45(2) J. NEW MUSIC RSCH. 160 (2003) (describing prior techniques in statistical music generation). Google published the first transformer architecture in 2017. *See generally* Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin, *Attention Is All You Need*, *in* 30 ADVANCES NEURAL INFO. PROCESSING SYS. 15 (2017). Prior to 2017, generative model architectures powered products like older versions of Apple's Siri voice assistant and of Google Translate. *See generally* Siri Team, *Deep Learning for Siri's Voice: On-device Deep Mixture Density Networks for Hybrid Unit Selection Synthesis*, APPLE MACH. LEARNING RSCH. (Aug. 2017), https://machinelearning.apple.com/research/siri-voices (describing Apple's Siri technology circa 2017). *See generally* Quoc V. Le & Mike Schuster, *A Neural Network for Machine Translation, at Production Scale*, GOOGLE BRAIN TEAM (Sept. 27, 2016), https://research.google/blog/a-neural-network-for-machine-translation-at-production-scale/ (describing the transition from phased-based translation systems to neural-network-based translation systems, before the release of transformers in 2017). Another example of earlier generative architectures is Generative Adversarial Networks (GANs), which have also had a place in popular discourse for around a decade with respect to deep fakes. *See generally* Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza et al., *Generative Adversarial Nets*, *in* 27 ADVANCES NEURAL INFO. PROCESSING SYS. 9 (2014).

[83] *See infra* Part I.B.3.a.

[84] *See infra* Part I.B.3.b.

a. Transformer Architecture

Transformers are a type of model architecture: a set of configurations for a neural network.[85] They are particularly good at capturing context in sequential information by modeling how elements in a sequence relate to each other. Consider our example sentence from above: "In the summer, I like to go to the [blank]". The word that should fill in the "[blank]" will have relationships to many of the other words in the sequence (such as "summer", "I", and "go"); these relationships mean that "beach" is a more likely candidate than "slopes". Given their effectiveness, since their release in 2017,[86] transformers have become the *de facto* way to model sequence-formatted data, including modalities as diverse as text, code, music, and protein structure. In recent years, some image-generation models also incorporate transformer-architecture variants.[87]

The transformer architecture can be used to train a generative model,[88] and today, almost all generative text models are transformer-based, including ChatGPT. Indeed, the "T" in "GPT" stands for Transformer.[89] This architecture

[85] *See supra* Part I.A.2 (introducing neural networks).

[86] Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, *supra* note 85. We do not address the technical details of transformers in the Article, but we nevertheless choose to mention them because they are a common term that repeatedly comes up in the context of generative AI. *See generally* Mark Riedl, *A Very Gentle Introduction to Large Language Models without the Hype*, MEDIUM (Apr. 13, 2023), https://markriedl.medium.com/a-very-gentle-introduction-to-large-language-models-withoutthe-hype-5f67941fa59e; Timothy B. Lee & Sean Trott, *A jargon-free explanation of how AI large language models work*, ARS TECHNICA (July 31, 2023), https://arstechnica.com/science/2023/07/a-jargon-free-explanation-of-how-ai-large-language-models-work/ (providing more in-depth, yet still accessible, treatments on transformers).

[87] *See* Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, *in* 2021 INT'L CONF. ON LEARNING REPRESENTATIONS (2021) (regarding vision transformers, or ViTs). *See* William Peebles & Saining Xie, Scalable Diffusion Models with Transformers (2023) (unpublished manuscript), https://arxiv.org/abs/2212.09748 (for diffusion-model transformers, or DiTs).

[88] *See supra* Part I.A.2.b (describing generative models). Not all transformer-based models generate content, for example, BERT (Bidirectional Encoder Representations from Transformers). *See generally* Devlin, Chang, Lee & Toutanova, *supra* note 14. Such models are not trained to predict (and then generate) the next word in a sequence. Instead, they are useful for other tasks: producing word embeddings, filling in missing data (e.g., blanks in provided text like "[blank] re-recorded her old studio albums after her masters were sold."), or performing a question-and-answering task. *See supra* Part I.A.1 (defining word embeddings).

[89] GPT is an acronym for Generative Pre-trained Transformer. "Generative" you already know, and we will discuss "Pre-trained" later. *See infra* Part I.C. Other transformer-based language models include LaMDA and the family of Llama models. *See generally*

consists of two parts: a neural network[90] and something called the attention mechanism. The attention mechanism, the key innovation in the original paper introducing the transformer architecture,[91] works particularly well to model contextual information in sequences. A transformer-based model has computational elements that enable it to take in an input sequence (e.g., a sequence of tokens[92] of text) and to weigh the importance of each part (i.e., each token) of the sequence in relation to the others. Based on this context of the input sequence,[93] the model can generate the next token as the output,[94] in a way that preserves the sequential relationships and makes it straightforward to add the generated token to the context and generate *the next token after that*, followed by the *next token after that*, and so on.[95] While, at this level of abstraction, this process may not seem particularly "intelligent," at large scale—i.e., training enormous transformer architectures on enormous datasets—it is responsible for the demonstrated capabilities of many current, transformer-based generative-AI systems.[96] For these systems, "intelligence" comes from the underlying

Thoppilan, De Freitas, Hall et al., *supra* note 72. (describing LaMDA). *See generally* Touvron, Lavril, Izacard et al., *supra* note 7; Touvron, Martin, Stone et al., *supra* note 7; Meta, *supra* note 51 (describing the Llama, Llama 2, and Code Llama models).

[90] *See supra* Part I.A.2.a (discussing neural-network model architectures).

[91] Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, *supra* note 85.

[92] *See supra* note 65 and accompanying text (defining token).

[93] This is where the term **context window** (or **context length**) originates; it refers to the size of the input sequence. *See generally* Anthropic, *Introducing 100K Context Windows*, ANTHROPIC (May 11, 2023), https://www.anthropic.com/index/100k-contextwindows/ (describing the context window in Anthropic's Claude chatbot system).

[94] Modeling data as tokens enables using transformers with non-text sequences, e.g., a token for a music model may be a musical note or a specific pitch. *See supra* note 65 and accompanying text.

[95] More precisely, it is not the transformer architecture on its own that takes the next-generated token, appends it to the previous tokens, and then passes it back through the model to produce the next token (and so on). Instead, this technique is a hallmark of an **autoregressive model**. It is common to combine transformer architectural blocks (which implement the attention mechanism, described above) with autoregressive elements to produce outputs at generation time (which predict one token at a time according to the previously generated tokens). Sometimes, this combination is called an **autoregressive transformer**, but we will use the shorthand of "transformer" in this Article for simplicity. This shorthand is not uncommon: GPT models are autoregressive transformers, even though the "autoregressive" part does not make it into the acronym. *See supra* note 92 (spelling out the GPT acronym).

[96] OpenAI's o1 model series involves additional mechanisms—reinforcement learning, different amounts of test-time compute, and more—that contribute to the models' capabilities. In a recent technical report, OpenAI claims that this training recipe is responsible for improved abilities to "reason" about challenging problems. *See supra* note 38 and accompanying text (discussing o1).

transformer models' abilities to mimic the sequential relationships of human-written text.

Lastly, it is also important to note that the transformer architecture can be implemented on an enormous scale. Just as deep neural networks contain numerous layers and connections between them, transformers can contain numerous layers and connections to construct models with billions of model parameters,[97] where (generally speaking, though with exceptions) larger models yield higher-quality generations. It is this large-scale stacking of transformers that gives us the term large language models (LLMs), which is commonly used to refer to transformer-based models like Llama, as well as systems that incorporate such models, like ChatGPT and Claude.

b. Diffusion-Based Models

Diffusion-based models[98] are popular in image generation; for example, Midjourney's underlying text-to-image model and (as the name suggests) the Stable Diffusion text-to-image model.[99] For text-to-image diffusion-based model training, the training data consist of pairs of images and corresponding text captions describing them.[100] Training occurs in two passes. First, for each

[97] For language models, this scale reflects the current state-of-the-art. *See infra* Part I.B.4.

[98] Such models are commonly called "diffusion models" in the literature. However, as we note below, "diffusion" is a training mechanism that involves sampling, and, for the purposes of this Article, should be understood as a training algorithm, not a model. This algorithm typically trains a neural-network architecture. We choose to disambiguate these subtleties with the term "diffusion-based model," even though it is not the term commonly used in the scientific field. *See generally* Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathany & Surya Ganguli, *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*, *in* 2015 PROC. 32ND INT'L CONF. ON MACH. LEARNING 2256 (2015) (regarding early work diffusion probabilistic models).

[99] Stable Diffusion is a text-to-image model that combines a transformer architecture for modeling text with diffusion for modeling images. DALL·E 2 uses a mix of transformers and diffusion, in a two-step process. *See generally* Rombach, Blattmann, Lorenz et al., *supra* note 32. (regarding the Stable Diffusion model). *See generally* MIDJOURNEY, *supra* note 32 (regarding the Midjourney text-to-image system). *See generally* Ramesh, Dhariwal, Nichol et al., *supra* note 47; Aditya Ramesh, *How DALL·E 2 Works*, ADITYA RAMESH (2022), http://adityaramesh.com/posts/dalle2/dalle2.html (detailing the DALL·E 2 system). To be precise, "diffusion" refers to a class of training algorithms; the resulting models are "diffusion models" in the same looser sense that the creature and not the doctor is "Frankenstein." The model itself is yet another neural network. A common neural-network architecture for diffusion models is called U-Net. *See generally* Olaf Ronneberger, Philipp Fischer & Thomas Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, 2015 MED. IMAGE COMPUT. & COMPUT. ASSISTED INTERVENTION 234–41.

[100] Diffusion is built on concepts from Bayesian inference—namely, Markov chains and variational methods. Early work on diffusion probabilistic models (DPMs) shows the

Figure 5: Several screenshots of the generation process using the Midjourney system, which uses text-to-image, diffusion-based models (Midjourney, Midjourney (2023), https://midjourney.com). We prompt with "an adventurous archaeologist with a whip and a fedora", and the Midjourney user interface shows the iterative de-noising process to produce the generations.

training-data example (the image and its caption), noise is incrementally added to the image until it effectively looks like static. This process intentionally corrupts the image, degrading its quality. Second, a neural network is trained to incrementally reverse this corruption process, removing noise and restoring the image to its original form. (Both of these passes are iterative; each has multiple steps that happen over time.) The first pass involves the repeated addition of noise, and the second involves denoising the fully noised image, a little bit at a time.[101] During the de-noising pass, the neural network is trained by evaluating

relationship between diffusion and concepts from variational autoencoders, another type of deep generative model. Starting in around 2019, DPMs started to become competitive with GANs, with respect to image generation. *See generally* Sohl-Dickstein, Weiss, Maheswaranathany & Ganguli, *supra* note 101 (regarding early work diffusion probabilistic models). *See generally* Dirk P. Kingma & Max Welling, *Auto-Encoding Variational Bayes*, *in* 2014 INT'L CONF. ON LEARNING REPRESENTATIONS 14 (2014); Danilo Jimenez Rezende, Shakir Mohamed & Daan Wierstra, *Stochastic Backpropagation and Approximate Inference in Deep Generative Models*, *in* 2014 PROC. 31ST INT'L CONF. ON MACH. LEARNING 1278 (2014) (regarding early work on variational autoencoders). *See generally* Goodfellow, Pouget-Abadie, Mirza et al., *supra* note 85 (describing GANs). *See generally* Yang Song & Stefano Ermon, *Generative Modeling by Estimating Gradients of the Data Distribution*, *in* 32 ADVANCES NEURAL INFO. PROCESSING SYS. 6840 (2019); Jonathan Ho, Ajay Jain & Pieter Abbeel, *Denoising Diffusion Probabilistic Models*, *in* 33 ADVANCES NEURAL INFO. PROCESSING SYS. 6840 (2020) (detailing the first methods that were competitive with GANs on image generation tasks).

[101] In a bit more detail, diffusion uses simulation techniques from the physical sciences to approach the machine-learning problem. Such simulations treat dynamical systems as a series of states; a given system can transition from one state to another over time. This

how well the de-noised image matches the original, noise-free image in the training data, and this evaluation is associated with the original text caption in the training data.[102]

To generate an output from a trained model, a system treats a user's text prompt like a caption for an unknown image. Starting from an image that consists only of noise, the system repeatedly applies the model to remove some noise, iteratively producing a series of images that are intended to increasingly align with the text prompt. We can therefore think of the production of a generated output as a sequence of images unfolding over time, starting from the completely noisy image and ending with the final generation, with every iteratively de-noised image between the two (for example, see Figure 5). It is possible to string these images together into an animation, as the Midjourney system does when producing generations in its user interface.[103]

4. The Role of Scale

Last, we turn explicitly to an important theme that has cropped up repeatedly throughout this section: scale. Above, we discussed how generative-AI systems are large-scale and have many components. Generative-AI models built using transformers or diffusion represent just one subset of these components, and they also tend to be massive. For example, state-of-the-art transformer-based LLMs currently have billions (or even trillions) of parameters.

The massive scale of these models is intended to capture the richness and complexity of equally massive datasets. As we mentioned briefly above, these datasets are often scraped from the Internet. This is a relatively new practice for training in machine learning. Prior to the publication of the first article describing the transformer architecture in 2017,[104] much of machine-learning research involved training models on smaller datasets. As points of comparison, both the MNIST[105] and CIFAR-10[106] datasets, two common benchmarks in discriminative

modeling approach has many applications besides image generation, including simulating the thermodynamics of molecules, the spread of a disease, and price movements in the stock market. For diffusion, the states are the intermediate images between the noise-free and completely noisy, static-resembling image.

[102] *See* Complaint at p. 12, Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135 (D. Del. Feb. 3, 2023) (giving an intuitive description of this process in the context of Stability AI's model).

[103] MIDJOURNEY, *supra* note 32.

[104] Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, *supra* note 85.

[105] Yann LeCun & Corinna Cortes, *MNIST handwritten digit database* (1999), https://www.lri.fr/~marc/Master2/MNIST_doc.pdf.

[106] Alex Krizhevsky, Vinod Nair & Geoffrey Hinton, *CIFAR-10 (Canadian Institute for Advanced Research)* (2009), http://www.cs.toronto.edu/~kriz/cifar.html.

deep learning tasks, each contain 60,000 labeled images. Even ImageNet,[107] a more challenging benchmark, has only 15 million labeled images. In contrast, datasets to train generative-AI models, such as LAION-5B,[108] have billions of training data examples.

In fact, today's generative-AI training data sets are so large, machine-learning practitioners do not have effective or efficient ways to fully know their exact contents.[109] This is one of the important impacts of scale. Earlier datasets like CIFAR-10, and even ImageNet, are small enough that they can be manually curated.[110] In the case of MNIST, the origin (i.e., provenance) of every data example is known and documented.[111] For large-scale datasets scraped from the web, provenance is much trickier, which will have implications for copyright,[112] discussed in Part II.

Nevertheless, despite such novel challenges, scale also confers new capabilities.[113] Today's generative-AI models are able to produce incredible content, in large part because of their large scale,[114] though it is not well understood exactly how or why.[115] Although it is possible to break down

---

[107] Jia Deng, Wei Dong, Richard Socher et al., *ImageNet: A large-scale hierarchical image database*, *in* 2009 IEEE CONF. ON COMPUT. VISION & PATTERN RECOGNITION 248—255 (2009).

[108] Beaumont, *supra* note 81; Schuhmann, Beaumont, Vencu et al., *supra* note 81. *See supra* Part I.B.2.

[109] Lee, Ippolito & Cooper, *supra* note 67, at 5 (discussing the challenges of provenance in generative-AI training datasets).

[110] *Id.*

[111] *Id.*

[112] This is also true because provenance is often not well-documented on the web. *See generally id.* Missing provenance in these datasets can also implicate other legal issues, not just copyright. Notably, LAION-5B contains CSAM. *See* David, *supra* note 81; Birhane, Prabhu & Kahembwe, *supra* note 81.

[113] *See generally* Brown, Mann, Ryder et al., *supra* note 67. (discussing new capabilities made possible with GPT-3). *See generally* Jared Kaplan, Sam McCandlish, Tom Henighan et al., Scaling Laws for Neural Language Models (Jan. 23, 2020) (unpublished manuscript), https://arxiv.org/abs/2001.08361 (discussing how model training scales with model size, dataset size, and available computing power).

[114] *See generally* Jensen Huang & Ilya Sutskever, *Fireside Chat with Ilya Sutskever and Jensen Huang: AI Today and Vision of the Future*, NVIDIA ON-DEMAND (2023), https://www.nvidia.com/en-us/on-demand/session/gtcspring23-s52092/ (regarding OpenAI's co-founder and chief scientist, Ilya Stuskever, crediting the importance of scale). *See generally* Jason Wei, Yi Tay, Rishi Bommasani et al., Emergent Abilities of Large Language Models (2022) (unpublished manuscript), https://arxiv.org/abs/2206.07682 (for a computer science paper on the same topic).

[115] This is an active area of current research. Recent work has investigated whether these new capabilities may be attributed to factors other than scale, but this area is still in its infancy. *See generally* Rylan Schaeffer, Brando Miranda & Sanmi Koyejo, Are Emergent

generative-AI systems into different known aspects and components (as we have done throughout this section), much remains unknown about how these systems actually work in detail.[116] Some machine-learning researchers are focused on applying models and systems to new tasks, without worrying too much about *why* they work, while others are focused on developing insights about the *why*. For the time being, policymakers, like everyone else, will have to work from a position of imperfect technical knowledge.

C. The Generative-AI Supply Chain

The previous section provided only a working definition of generative-AI models and a brief overview of the generative-AI systems they are embedded in. But even this overview showed the remarkable complexity and diversity of these models and systems. In turn, this complexity and diversity are important for how we reason about copyright implications. They affect *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.[117]

Abilities of Large Language Models a Mirage? (2023) (unpublished manuscript), https://arxiv.org/abs/2304.15004 (discussing how choices of evaluation metrics can affect perceptions of model capabilities).

[116] Understanding the inner workings of large-scale, machine-learning models has been an active area of research over the last decade. *See generally* David Baehrens, Timon Schroeter, Stefan Harmeling et al., *How to explain individual classification decisions*, 11 J. MACH. LEARNING RSCH. 1803 (2010); Chris Olah, Arvind Satyanarayan, Ian Johnson et al., *The Building Blocks of Interpretability*, DISTILL (Mar. 6, 2018), https://distill.pub/2018/building-blocks/; Nelson Elhage, Neel Nanda, Catherine Olsson et al., A Mathematical Framework for Transformer Circuits (Dec. 22, 2021) (unpublished manuscript), https://transformer-circuits.pub/2021/framework/index.html (discussing interpretability, explainability, and mechanistic interpretability). *See generally* Pang Wei Koh & Percy Liang, *Understanding Black-box Predictions via Influence Functions*, 70 PROC. MACH. LEARNING RSCH. 1885 (2017); Ekin Akyurek, Tolga Bolukbasi, Frederick Liu et al., *Towards Tracing Knowledge in Language Models Back to the Training Data*, *in* 2022 FINDINGS ASS'N FOR COMPUT. LINGUISTICS: EMNLP 2022 2429 (2022); Roger Grosse, Juhan Bae, Cem Anil et al., Studying Large Language Model Generalization with Influence Functions (Aug. 7, 2023) (unpublished manuscript), https://arxiv.org/abs/2308.03296 (discussing influence functions). While these fields of work have provided insights, many believe that there lacks sufficient evidence to depend on models to make consequential decisions. *See generally* Zachary Lipton, *The Mythos of Model Interpretability: In Machine Learning, the concept of interpretability is both important and slippery*, 16 QUEUE 31 (2018).

[117] The generative-AI supply chain is a 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

To provide a framework for reasoning about this diversity and complexity, it is helpful to conceive of a generative-AI supply chain that has eight stages (Figure 6):

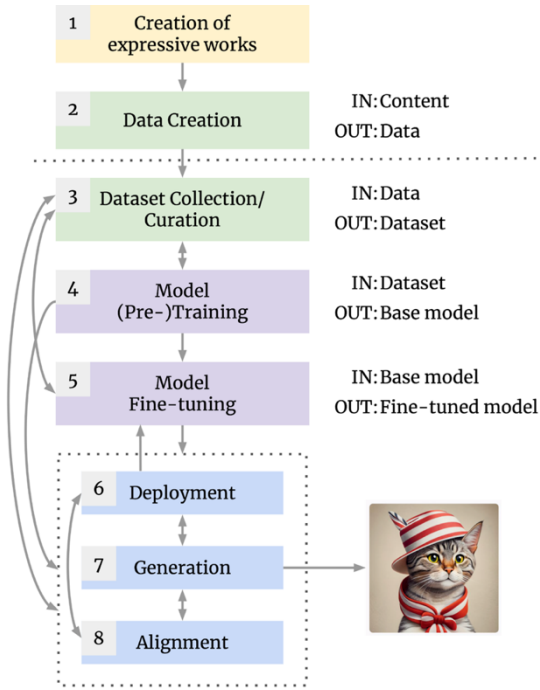


Figure 6: The generative-AI supply chain.

broader societal expectations—in our case, the preservation of copyrights. *See* A. Feder Cooper, Emanuel Moss, Benjamin Laufer & Helen Nissenbaum, *Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning* 867–69, *in* 2022 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 864 (2022) (describing the problem of "many hands" in data-driven machine learning/AI systems in relation to complex "ML pipelines"). *See* Rui-Jie Yew & Dylan Hadfield-Menell, Break It Till You Make It: Limitations of Copyright Liability Under a Pretraining Paradigm of AI Development (2023) (unpublished manuscript), https://genlaw.github.io/CameraReady/30.pdf (regarding an instantiation of this problem for generative AI and copyright).

1. Creation of expressive works.[118]

The supply chain necessarily starts with creative works: all of the books, artwork, software, and other products of human creativity that generative AI seeks to learn from and emulate.

2. Data creation.[119]

Creative works and other information must then be converted into data: digitally encoded files in standard formats.

3. Dataset collection and curation.[120]

Individual pieces of data are useless for AI training by themselves. Instead, they must be compiled into training datasets: vast and carefully structured collections of related data. This process requires both extensive automation and thoughtful, human-curated decision-making.

4. Model (pre-)training.[121]

To create a generative-AI model, its creator picks a technical architecture, assembles training datasets, and then runs a training algorithm to encode features of the training data in the model. Model training is both a science and an art, and it involves massive investments of time, money, computing resources, and (often) human monitoring and intervention. The model that results from the initial training process is called a "base" or "pre-trained model" because it is often just a starting point.

5. Model fine-tuning.[122]

A base model can also be fine-tuned to improve its performance or adapt it to a specific problem domain. This process, too, involves extensive choices—and it need not be carried out by the same entity that did the initial training.

6. Model release and system deployment.[123]

---

[118] *See infra* Part I.C.1.
[119] *See infra* Part I.C.2.
[120] *See infra* Part I.C.3.
[121] *See infra* Part I.C.4.
[122] *See infra* Part I.C.5.
[123] *See infra* Part I.C.6.

A model by itself is an inert artifact. It can be used only by technical experts with access to (often substantial) computing resources. To make a model usable by a wider user base, a model must be released—its parameters uploaded to the Internet, where developers and researchers can download and integrate them into their own applications and projects—or deployed as a service. Deployment involves embedding the model in some larger software system that provides a convenient interface (e.g., ChatGPT's web-based UI, Midjourney's Discord bot).

7. Generation.[124]

A deployed service can be used to generate outputs (generations): new creative works that are based on statistical patterns in the training dataset but typically combine them in new ways. A generation is based on the particular system, as well as a prompt supplied by the user: an input that describes the particular features they want the output to have. Generation is the only part of the supply chain that most end users see directly.

8. Model alignment.[125]

The supply chain does not end with generation. Both before and after deployment, the developers of a generative-AI system can perform alignment by rating prompts and generations: further adjusting the model and the system it is embedded in to better achieve users' (and their own) needs. Those needs can include safety, helpfulness, and legal compliance.

Each stage gathers inputs from the prior stage(s) and hands off outputs to the subsequent stage(s), which we indicate with (sometimes bidirectional) arrows. In this way—as in many others—the supply chain feeds back into itself. It is not a simple cascade from data to generations. Instead, each stage is regularly adjusted to better meet the needs of the others. This framing is broadly useful for reasoning about different legal considerations for generative AI; we employ it specifically for copyright analysis in Part II.

The first two stages, the creation of expressive works and data creation, predate the advent of generative-AI systems. Nevertheless, they are indispensable parts of the production of generative-AI content, which is why we begin our discussion of the supply chain with these processes. The next six stages reflect processes that are new for generative-AI systems.

The connections between these supply-chain stages are complicated. While in some cases, one stage clearly precedes another, in other cases, there are many

[124] *See infra* Part I.C.7.
[125] *See infra* Part I.C.8.

different possible ways that stages can interact.[126] We highlight some of this complexity in the following subsections; we call attention to different possible timelines for when supply-chain stages can be invoked, and which actors can be involved at each stage.

1. Creation of Expressive Works

Artists, writers, coders, and other creators produce expressive works. Generative-AI systems do too,[127] but state-of-the-art systems are only able to produce expressive works because their models have been trained on data derived from pre-existing works by artists, writers, coders and other creators.[128] While perhaps obvious, it is nevertheless important to emphasize that the processes of producing most creative works have (thus far) had nothing to do with machine learning.[129] Historically, painters have composed canvases, writers have penned articles, coders have developed software, etc. without consideration of how their works might be taken up by automated processes. Nonetheless, once a work exists, it can be turned into data readable by a computer and used as training data. As a result, content creators and their original works are a part of the generative-AI supply chain, whether they would like to be or not (see Figure 6, stage 1).

2. Data Creation

Expressive *works* are distinct from their *representations* as data. In copyright terms, digital representations like the JPEG encoding of a photograph on an SSD

[126] *E.g.*, model pre-training necessarily precedes model fine-tuning. Fine-tuning and alignment involve additional training, and thus both necessarily follow pre-training (It is increasingly common to refer to all types of additional training, such as fine-tuning and alignment, as *post*-training.) *See* Figure 6.

[127] We discuss this in more detail below with respect to generation. *See infra* Part I.C.7. We also discuss this when we delve into copyright and authorship. *See infra* Part II.A.

[128] As we address below, a data example is not the same as its associated expressive work. Additionally, some models are trained on synthetic data, typically generated by other generative-AI models (which, in turn, were trained on pre-existing works). Training predominantly on synthetic data is a growing practice. *See* Gokaslan, Cooper, Collins et al., *supra* note 37 (for an example of training a model partially on synthetic data). There are some concerns that training on synthetic data can seriously compromise model quality. *See generally* Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao et al., The Curse of Recursion: Training on Generated Data Makes Models Forget (May 27, 2023) (unpublished manuscript), https://arxiv.org/abs/2305.17493 (detailing "model collapse" in different generative models). However, recent work shows that reduction in model quality can be avoided with extensive data curation. *See* Suriya Gunasekar, Yi Zhang, Jyoti Aneja et al., Textbooks Are All You Need (2023), https://arxiv.org/abs/2306.11644 (unpublished manuscript).

[129] 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. *See* SCALE AI, https://scale.com/.

storage device are "copies" of the works they encode. What makes these digital representations different from other copies, like paint on a canvas or an engraved musical score, is that they are *data*: they are in computer-readable formats.[130] For the most part, the transformation of creative content to digital data formats predates generative AI[131] (see Figure 6, stage 2). It is a process that has grown in tandem with the proliferation of the modern Internet.[132] Regardless, all state-of-the-art generative-AI systems depend on access to digitized data, and so they depend on this process. Text-to-text generation models are trained on 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 for our purposes because many of the works that have been transformed into data include copyrightable expression. For example, the GitHub Copilot system uses models trained on copyrighted code, ChatGPT's underlying model(s) are trained on copyrighted text data scraped from the web, Stability AI's Stable Diffusion is trained on copyrighted text and images, and so on. For the most part, it is the copyright owners of these individual works who are the plaintiffs in copyright-infringement suits against actors all along the supply chain.

3. Dataset Collection and Curation

Models are not trained on individual, isolated data examples; instead, data examples are grouped together into larger datasets for training.[133] The training process for cutting-edge generative-AI models requires particularly vast quantities of data; dataset creators often meet this need by scraping data from the Internet.[134] Conceptually, this process requires three distinct stages: collection (gathering data), curation (deciding which gathered data to use), and arrangement (putting that data in a recurring standard structure). In practice, these stages are profoundly intermingled: curatorial choices shape the choices of what data to collect, certain collection choices limit the design space of possible arrangements, and arrangement can take place during collection and curation.[135] For example,

---

[130] *See supra* Part I.A.1 (defining data).
[131] *See, e.g.,* NANNA BONDE THYLSTRUP, THE POLITICS OF MASS DIGITIZATION (2024).
[132] *Id.*
[133] *See supra* Part I.A.2; *supra* Part I.B. This is not to say individual training examples are unimportant. Specific pieces of training data can have an out-sized influence on generations, compared with other pieces of training data. *See generally* Koh & Liang, *supra* note 119; Akyurek, Bolukbasi, Liu et al., *supra* note 119; Grosse, Bae, Anil et al., *supra* note 119. (discussing influence functions).
[134] This is not the only way to collect large amounts of data. *See* Lee, Ippolito & Cooper, *supra* note 67, at 5 (discussing other ways datasets may come to be).
[135] Note, however, that collection, curation, and arrangement do not *always* have to happen together, and may involve different sets of actors. It is also possible for curation to happen

scraping data from the Internet can simultaneously involve curatorial choices, e.g., filtering out unwanted types of data, such as "toxic speech."[136] Such curatorial choices muddle the line between dataset creation and curation, as both processes can effectively happen in tandem.

With respect to the generative-AI supply chain, there are several points worth highlighting in dataset collection and curation processes (see Figure 6, stage 3). First, while dataset creation and curation can be carried out by the same entities that train generative-AI models,[137] it is common for these processes 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.[138] It is necessary, therefore, to consider the potential liability of dataset creators and curators separately from the potential liability of model trainers.[139]

Second, training datasets are important objects in their own right. Note that dataset curation, as described above, 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."[140] As such, training datasets can themselves potentially be copyrighted.

Third, while a few training datasets include metadata on the provenance of their constitutive data examples, many datasets do not. Provenance makes it easier to answer questions about the data sources a model was trained on, which can be relevant to an infringement analysis.[141] It also bears on the ease with which specific material can be located, and if necessary, removed, from a dataset.[142] However, the use of web scraping to collect generative-AI training datasets is directly in tension with maintaining information about provenance. As we discussed above, relying on Internet sources and the scale of scraped datasets makes determining the origins of individual data examples very challenging.[143]

---

after the start of model training, in response to metrics that are observed during the training process. That is, curation could follow (and then also precede further) model (pre-)training (Figure 6, stage 4; *see infra* Part I.C.4), or model fine-tuning (Figure 6, stage 5; *see infra* Part I.C.5). These complex interactions are the reason for the bidirectional arrows between stages in Figure 6.

[136] *See generally* Lee, Ippolito & Cooper, *supra* note 67 (discussing dataset creation and curation choices, including toxic content filtering).

[137] *See infra* Part I.C.4.

[138] Technically, LAION presents the dataset as a collection of the URLs of the images. Stable Diffusion then visits each URL to collect images for training. *See supra* Part I.B.2.b; *supra* I.B.4 and citations therein.

[139] *See infra* Part II.E.

[140] 17 U.S.C. § 101.

[141] *See infra* Part II.B–II.F.

[142] *See infra* Part II.G; *infra* Part III.A.3 (discussing challenges of removal).

[143] *See supra* Part I.B.4 and references therein. *See generally* Lee, Ippolito & Cooper, *supra* note 67, at 5 (discussing provenance challenges for generative AI).

Notably, and as we will discuss below, even if particular dataset creators and curators release a training dataset with a chosen license, this does not guarantee that the works within the dataset are appropriately licensed.[144] For example, the complaint in *Tremblay v. OpenAI, Inc.* alleges that ChatGPT's underlying model(s) were trained on datasets that do not properly license the books data that they contain.[145]

Differences in the licensing status of training data can have consequences for the characteristics and quality of the resulting models., For example, Sewon Min and colleagues used public-domain and permissively-licensed text to train a language model, and demonstrated a degradation in quality in domains that are not well represented in the data.[146] Similarly, Aaron Gokaslan and colleagues found a degradation in quality when training diffusion-based text-to-image models using only Creative-Commons licensed images.[147] Additionally, data in the public domain can be unrepresentative, which can introduce biases into models trained on it.[148]

4. Model (Pre-)Training

Following the collection and curation of training datasets (Figure 6, stage 3), it is possible to train a generative-AI model (Figure 6, stage 4). The model

[144] Indeed, the creators and curators would have to check that they have abided by each data example's respective license(s). *See infra* Part II.A (regarding authorship and training datasets).

[145] In particular, the complaint in *Tremblay v. OpenAI* alleges that the training data included books from infringing "shadow libraries" like Library Genesis. Complaint at p. 34, Tremblay v. OpenAI, Inc., No. 3:23-cv-03223 (N.D. Cal. June 28, 2023). Note, however, this claim is based on circumstantial evidence, because the datasets the GPT models were trained on have not been made public. Text data from books have been a key player in other dataset-related complaints. For example, The Pile dataset was originally released under the MIT license. Stella Biderman, Kieran Bicheno & Leo Gao, Datasheet for the Pile (Jan. 13, 2022) (unpublished manuscript), https://arxiv.org/abs/2201.07311. The Pile was core to the complaint in *Kadrey*, since the Pile contains 108GB of the dataset Books3 (which itself contains content from Bibliotek, a popular torrent interface). *See generally* Complaint, Kadrey v. Meta Platforms, Inc., No. 3:23-cv-03417 (N.D. Cal. July 7, 2023). The original download URL for The Pile (https://the-eye.eu/public/AI/pile/) is no longer resolving (as of September 2023).

[146] Sewon Min, Suchin Gururangan, Eric Wallace et al., SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore (Aug. 8, 2023) (unpublished manuscript), https://arxiv.org/abs/2308.04430.

[147] Gokaslan, Cooper, Collins et al., *supra* note 37.

[148] Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018).

trainer[149] selects a training dataset, a model architecture (i.e., a set of initialized model parameters), and a training algorithm.[150]

As mentioned above, the process of training—from transforming these inputs into a trained model—is expensive. It requires a substantial investment of multiple resources: time, data storage, computing power, and human labor. For example, BLOOM, a 176-billion-parameter open-source model from HuggingFace was trained for 3.5 months on 1.6 terabytes of text and using 384 GPUs.[151] It cost an estimated $2–5 million in computing resources for both the development and ultimate training of BLOOM.[152] As another point of reference, MosaicML, a company (acquired by Databricks[153]) that develops solutions for training as cheaply and efficiently as possible, trained a GPT-3-quality model for less than $0.5 million.[154] Altogether, the dollar cost for pre-training can generally range

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

[150] This description omits additional choices that we elide for simplicity. First, as noted above, the trainer selects hyperparameters to calibrate and customize the training algorithm. *See supra* Part I.B.1. Second, the trainer must select a **seed value** for the random choices made during the training. Machine learning 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.

[151] *See generally* Stas Bekman, *The Technology Behind BLOOM Training*, HUGGINGFACE (July 14, 2022), https://huggingface.co/blog/bloom-megatron-deepspeed (for training details). *See* BigScience Workshop, Teven Le Scao, Angela Fan et al., BLOOM: A 176B-Parameter Open-Access Multilingual Language Model (June 27, 2023) (unpublished manuscript), https://arxiv.org/abs/2211.05100 (for the model details).

[152] 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 the most expensive) part of overall model development.

[153] Databricks, *Databricks Completes Acquisition of MosaicML*, DATABRICKS (July 19, 2023), https://www.databricks.com/company/newsroom/press-releases/databricks-completesacquisition-mosaicml. MosaicML (recently rebranded as Databricks Mosaic) offers hosted services for model training, model fine-tuning, and generation. So, too, does Together AI. Together AI, *Together Fine-tuning*, Together AI (November 25, 2024), https://www.together.ai. ("Fine-tune leading open-source models with your data to achieve greater accuracy for your tasks"). *See infra* Part I.C.5 (discussing fine-tuning); Part I.C.6 (discussing deployed hosted services).

[154] 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. MosaicML's model is substantively different from GPT-3. For one, MosaicML's model is a much smaller 30 billion parameters, compared to the original GPT-3 model's 175 billion. Additionally, MosaicML trained on more data, shifting some of the development cost towards data

from six to eight figures (or even more), depending on the size of the model, the size of the training dataset, the particular training algorithms, the length of the training process, the efficiency of the software and hardware used, and other choices.

Further, the training process is not completely automated; training often requires people to monitor and make tweaks to the process. For example, model trainers typically run evaluation metrics on the model while it is being trained, in order to assess the progress of training.[155] Depending on these metrics,[156] model trainers may pause the training process to manually revise the training algorithm[157] or the dataset, which we indicate with bidirectional arrows at Figure 6, stages 3-4. Human intervention in response to metrics necessarily makes model training an iterative process.

The output of this process is typically called a pre-trained model or base model.[158] At this point, the base model has many possible futures. It could just sit idly in memory, collecting figurative dust, never to be used to produce

collection and curation, 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* Abhinav Venigalla & Linden Li, *Mosaic LLMs (Part 2): GPT-3 quality for <$500k*, MOSAICML (Sept. 29, 2022), https://www.mosaicml.com/blog/gpt-3-qualityfor-500k (regarding MosaicML's model). *See generally* Brown, Mann, Ryder et al., *supra* note 67 (for the size of GPT-3).

[155] Google's TensorBoard and software from Weights & Biases are two tools for running evaluation metrics and monitoring during training. *See generally* TensorFlow, *TensorBoard: TensorFlow's visualization toolkit*, TENSORFLOW (2023), https://www.tensorflow.org/tensorboard. (regarding Tensorboard). *See generally Weights & Biases*, WEIGHTS & BIASES, https://wandb.ai/site (regarding Weights & Biases).

[156] Evaluation metrics attempt to elicit how "useful" or "good" the model is. These metrics are not comprehensive, since there is no single way to capture "usefulness" or "goodness" in math. *See generally* Lee, Ippolito & Cooper, *supra* note 67, at 5 (for a discussion of evaluation metrics and the impossibility of exactly and comprehensively defining "useful" and "good").

[157] *E.g.*, change the hyperparameters.

[158] Others use the term "foundation model." The term "foundation" can be (and has been) 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, rather that they serve as a foundation to build other models on. The term itself is controversial in the machine-learning community. *See* Thomas G. Dietterich (@tdietterich), TWITTER (Aug. 12, 2022, 7:58 PM), https://twitter.com/tdietterich/status/1558256704696905728) (including the replies and offshoots).

generations.[159] The model parameters could be uploaded to a public server,[160] from which others could download it and use it however they want.[161] The model could be integrated into a system and deployed as a public-facing application,[162] which others could use directly to produce generations.[163] Or, the model could be further modified by the initial model trainer, by another actor at the same organization, or, if made publicly available, by 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 (and a chosen training algorithm, etc., as at the beginning of this section).[164]

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. Additional training of a base model is called fine-tuning, to which we now turn.

5. Model Fine-Tuning

As described above, models reflect their training data.[165] Base models trained on large-scale, web-scraped datasets reflect a lot of general information sourced from different parts of the Internet. They are not typically trained to reflect specialized domains of knowledge. For example, a text-to-text base model trained on large quantities of English-language data may be able to capture general English-language semantics and information; however, such a model may not be

[159] This point shows that the line distinguishing "programs" from "data" in machine learning is murky. The set of parameters in a model can be viewed as a **data structure** containing vectors of numbers that, on their own, do not *do* anything. However, one could load that data structure into memory and apply some relatively lightweight linear algebra operations to produce a generation. *See supra* Part I.B. In this respect, we could also consider the model to be a program (and an algorithm). This is why we talk about the model being *within* the function f in our analogical discussion of machine-learning-as-a-function. (*See supra* Part I.A.2.a.) 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 (just the vectors of numbers) or to the model as a combination of software and the model parameters, which together can be executed like a program.

[160] For example, HuggingFace hosts a repository of over 300,000 open- and semi-closed models. *See generally Models*, HuggingFace (Sept. 2, 2023), https://huggingface.co/models.

[161] They could fine-tune the model (*see infra* Part I.C.5), embed the model in a system that they deploy for others to use (*see infra* Part I.C.6), produce generations (*see infra* Part I.C.7), align the model (*see infra* Part I.C.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 6.

[162] *See infra* Part I.C.6.

[163] *See supra* Part I.B; *infra* Part I.C.7.

[164] Cooper, Moss, Laufer & Nissenbaum, *supra* note 120 (discussing the "many hands" problem).

[165] *See supra* Part I.A.2; *supra* Part I.B

able to, for example, reliably reflect detailed scientific information about molecular biology (e.g., answering the question "what is mitosis?").

This is where fine-tuning comes in (Figure 6, stage 5): the process of modifying a preexisting, already-trained model, with the general goal of making this model better along some dimension of interest. As the name suggests, most fine-tuning aims to retain the general strengths of what a model has already learned while optimizing its specific details. This process often involves training on additional data that is more aligned with the specific goals.[166] Training transforms data into a model, and fine-tuning transforms a model into another model.

Fine-tuning essentially involves just running more training. In this respect, the overall process of fine-tuning is similar to pre-training: both execute a training process. However, fine-tuning and pre-training run with different inputs, which ultimately makes the trajectories and outputs of their respective training processes very different. That is, even though fine-tuning and pre-training often employ the same training algorithm, they typically use different input training data and different input model parameters.[167] To add more precision to our previous statement, fine-tuning transforms a model into another model while incorporating more data.

Whereas pre-training data tend to be more general, fine-tuning data are typically sourced from a specific problem domain of interest; whereas the input model architecture to pre-training is an initialized, untrained model,[168] for fine-tuning, the input model parameters have already undergone some training and are no longer in their initialized state. Continuing the example above, a base language model could be fine-tuned on scientific papers to improve its ability to summarize scientific content; the fine-tuning stage takes the learned parameters of the more general base model and updates them by training further on scientific text data.

a. Forks in the Supply Chain

Two important observations follow from our description of fine-tuning as (effectively) just performing more training. On the one hand, a model trainer does not have to fine-tune at all. A base model produced during pre-training could be

[166] And thus the reason for the bidirectional arrow between stages 3 and 5 in Figure 6. Similar to pre-training, monitoring metrics during fine-tuning may lead to further dataset curation. *See supra* Part I.C.4. This is also sometimes why fine-tuning is considered a part of alignment, as the fine-tuning dataset can contain examples that direct the model to possess additional behaviors, such as instruction following. *See infra* Part I.C.8 (discussing model alignment).

[167] As discussed above, 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. *See supra* Part I.A.1; *supra* Part I.B.4.

[168] *I.e.*, the vectors of numbers that constitute the model parameters have not "learned" anything yet. *See supra* Part I.A.1; *supra* Part I.C.4.

used as-is for later stages in the supply chain. On the other hand, just as it is possible to fine-tune *less* than once, it is possible to fine-tune *more* than once, taking an already-fine-tuned model, and using it as the input for another fine-tuning pass. Thus, a model is a "base" or "fine-tuned" model *only in relation to other models*. A model is a "base" model when it is the input to a fine-tuning pass; it is a "fine-tuned" model when it is the output from a fine-tuning pass.[169] 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.

Pre-training and fine-tuning can be split among different actors. Sometimes, the creator of a model also fine-tunes it. Google's Codey models (for software code generation) are fine-tuned versions of Google's PaLM 2 model.[170] In other cases, another party does the fine-tuning. A particularly common case arises when a model's parameters are publicly released (as Meta has done with its several versions of Llama models), [171] allowing others to take the model and independently fine-tune it for particular applications. For example, LMSYS Org fine-tuned Meta's publicly released Llama model on the crowdsourced ShareGPT dataset to produce the Vicuna model.[172] The creators of Vicuna, in turn, have also released their model publicly, allowing anyone else with the requisite resources and know-how the ability to fine-tune the model on additional data.[173] For another example, when a model is deployed within a hosted service (Figure 6, step 6), developers of that service may expose APIs to end-users that enable them to fine-tune the model.[174]

---

[169] These terms do not capture the intrinsic technical features of a model; instead, they describe different processes by which a model can be created.

[170] Google, *Foundation Models,* GOOGLEAI (Aug. 17, 2023), https://ai.google/discover/foundationmodels/ (describing Codey).

[171] Touvron, Lavril, Izacard et al., *supra* note 7; Meta, *supra* note 51.

[172] *See generally* The Vicuna Team, *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*, LMSYS ORG (Mar. 30, 2023), https://lmsys.org/blog/2023-03-30-vicuna/ (regarding the Vicuna model). 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). *See generally* SHAREGPT, https://sharegpt.com/ (regarding the ShareGPT dataset).

[173] *See* Colin Raffel, *Collaborative, Communal, & Continual Machine Learning* 15 (2023), https://colinraffel.com/talks/faculty2023collaborative.pdf (for a figure showing many fine-tuned models building on one base model).

[174] For example, ChatGPT has a fine-tuning API, which enables end-users to fine-tune some of the underlying system's models through the hosted service's API on a custom dataset. Andrew Peng, Michael Wu, John Allard et al., *GPT-3.5 Turbo fine-tuning and API updates* (Aug. 22, 2023), https://openai.com/blog/gpt-3-5turbo-fine-tuning-and-api-updates. This is why we place an arrow from deployment (Figure 6, stage 6) to fine-tuning

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 are used for training and how the resulting trained model will 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-intended guardrails.[175] A full copyright analysis may require treating these actors differently, and indeed, may require analyzing their conduct in relation to each other.[176]

6. Model Release and System Deployment

At this point in the supply chain, we have a trained generative-AI model—either a base model[177] or a fine-tuned model.[178] As noted above with respect to base models, trained models have a variety of possible futures, of which fine-tuning is just one option. The next three stages address other futures for base and fine-tuned models: it is possible to release a model or deploy it as part of a larger software system (Figure 6, stage 6), use the trained model parameters directly to produce generations (Figure 6, stage 7),[179] or to take the trained model and further alter or refine it via model alignment techniques (Figure 6, stage 8).[180] In brief, there is a complicated orchestration between the deployment, generation, and alignment stages, which can happen in different orders, in different combinations, and at different times for different generative-AI systems. For ease of exposition, we still present these stages of the generative-AI supply chain one at a time, and we begin here with model release and system deployment (Figure 6, stage 6).

(Figure 6, stage 5). There are also generative-AI companies that offer fine-tuning services of (typically open-source) models to individual and corporate clients, either through APIs or bespoke business agreements, e.g., Databricks Mosaic (previously MosaicML) and Together AI. *See infra* Part I.C.6 (discussing deployed services). *See supra* note 156 and accompanying text (discussing Databricks Mosaic and Together AI).

[175] This process could involve fine-tuning the model on additional copyrighted works so that it is better at reproducing those works at generation time. But it could also involve training the model on user prompts using RLHF, *see infra* Part I.C.8, to undo the work that the base-model trainer has done to make the model refuse requests to generate copyrighted works. Instead of responding "I'm sorry, I can't provide the text of works under copyright", the model would respond with a generation that is a copy of or substantially similar to a copyrighted work in the model's training data.

[176] *See infra* Part II.E.

[177] *See supra* Part I.C.4.

[178] *See supra* Part I.C.5.

[179] *See infra* Part I.C.7.

[180] *See infra* Part I.C.8.

A model is released when its parameters are uploaded to a server or platform (like HuggingFace[181]), from which others can download it.[182] Released models, which include Meta's (multiple-versioned) Llama family of models[183] and Stable Diffusion,[184] give downloaders direct access to their parameters. This enables developers and practitioners to directly embed the model in their own code to produce generations or to alter the model (and thus potentially its behavior) through fine-tuning or model alignment techniques.[185]

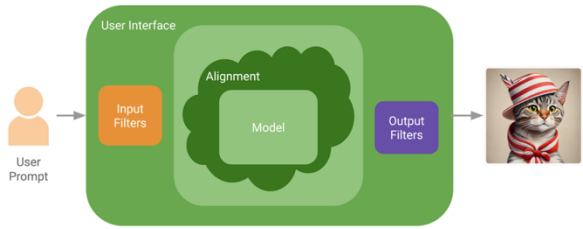


Figure 7. A high-level illustration of how, for deployed services, a model is embedded within a generative-AI system

In contrast, closed-source models are not directly available to users external to their model trainers and owners. Such models are typically embedded in large, complex software systems (Figure 7),[186] which can be deployed to both internal and external users through software services. For example, a model could be hosted by a company like OpenAI, Microsoft, Pika, Stability AI, or Google. It could be used internally at those companies for a variety of software-based services (e.g., an internally developed Google LLM integrated into Google

[181] *Models*, *supra* note 163.
[182] Meta first asked interested parties to request Llama's model parameters, rather than uploading them for anyone to download. However, Llama's model parameters were quickly leaked on the website 4chan. James Vincent, *Meta's powerful AI language model has leaked online—what happens now?*, WEIGHTS & BIASES (2023), https://wandb.ai/site. 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. *Use Policy*, META AI (2023), https://ai.meta.com/llama/use-policy/ (for the Llama 2 Community License). Meta has more recently released its third generation of Llama-family models: Llama 3.
[183] Touvron, Lavril, Izacard et al., *supra* note 7; Touvron, Martin, Stone et al., *supra* note 7; Meta, *supra* note 51.
[184] Rombach, Blattmann, Lorenz et al., *supra* note 32.
[185] *See infra* Part I.C.8.
[186] *See supra* Part I.B.1.

Search[187]), or released as a hosted service that gives external users access to generative-AI functionality. External-facing services could be deployed in a variety of forms. Most consumer-facing services are for generation-based products, and these services do not typically include the ability to change the model's parameters. In developer- and business-focused products, hosted pre-training and fine-tuning services are now also not uncommon.[188] They can be browser-based user applications (e.g., ChatGPT, Claude, Midjourney, DreamStudio, Perplexity), or public (but not necessarily free) APIs for developers (e.g., GPT and o-series models, Cohere, Anthropic).[189]

Of course, release and deployment are not mutually exclusive. For example, DreamStudio is a web-based user interface[190] built on top of services hosted by Stability AI.[191] The DreamStudio application gives external users access to a generative-AI system that contains the open-source Stable Diffusion model,[192] which Stability AI also makes available for direct download.[193]

This is a familiar spectrum from Internet law: cloud-hosted services at one end and 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 user and the deployer. Typically, model trainers and owners maintain more control over models deployed through hosted services and less control over models whose parameters they have released.[194] When trainers and owners embed models within systems, rather than release them directly,[195] they can imbue models with additional behaviors prior to giving users access to model functionality.

For instance, a generative-AI system deployed as a web-based application or as an API will often modify the user-supplied prompt before supplying it as input to the model. Several applications (ChatGPT, Gemini, and Sydney, just to name

---

[187] Liz Reid, *Generative AI in Search: Let Google do the searching for you*, GOOGLE (May 14, 2024), https://blog.google/products/search/generative-ai-google-search-may-2024/.

[188] *See supra* note 177 and accompanying text (discussing fine-tuning APIs and services).

[189] 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).

[190] *DreamStudio*, *supra* note 45.

[191] *Stable Diffusion XL*, *supra* note 32.

[192] Rombach, Blattmann, Lorenz et al., *supra* note 32.

[193] It is possible that models that are 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, *supra* note 7 (regarding both GPT-3.5 and GPT-4 model integration into the ChatGPT web application). *See* Figure 4.

[194] *See generally* Vincent, *supra* note 185.

[195] 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. *See supra* Part I.A.2.

a few) add additional instructions (i.e., application or system prompts) to the user's input to create a compound prompt.[196] The additional instructions change the model's outputs.[197] For example, providing each of the following system prompts to a language model directs the model to behave differently: "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 . . .".[198]

APIs and web applications also allow model deployers to include software that filters inputs or outputs (Figure 7). Concretely, 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.[199] GitHub Copilot expressly states that it uses "filters to block offensive words in the prompts and avoid producing suggestions in sensitive contexts."[200] Additionally, some APIs and web applications include output filters to avoid generating anything that looks too similar to a training example.[201] Unfortunately, using output filters to find generations that are similar

[196] *See generally* Yiming Zhang & Daphne Ippolito, Prompts Should not be Seen as Secrets: Systematically Measuring Prompt Extraction Attack Success (July 13, 2023) (unpublished manuscript), https://arxiv.org/abs/2307.06865 (which discovers proprietary system prompts). *See generally Custom instructions for ChatGPT*, OPENAI (Aug. 17, 2023), https://openai.com/blog/custom-instructions-for-chatgpt (announcing a ChatGPT feature that allows users to provide their own additional prompts, which get appended to their future inputs to create compound prompts).

[197] This kind of prompt transformation is another technique for steering the behavior of a model, which is sometimes described as a type of alignment. *See infra* Part I.C.8.

[198] Fatih Kadir Akın, *Awesome ChatGPT Prompts*, GITHUB (Aug. 17, 2023), https://github.com/f/awesome-chatgpt-prompts. (These prompts and more can be found on this site). DAIR.AI, *General Tips for Designing Prompts*, PROMPT ENGINEERING GUIDE (Aug. 17, 2023), https://www.promptingguide.ai/introduction/tips. (This handbook provides an introduction to creating prompts for large language models). *Custom instructions for ChatGPT*, *supra* note 199. Strictly speaking, these are not system prompts, but prompts that users provide that alter the behavior of the system in subsequent in-context generations. The general idea is the same, however.

[199] These filters may detect undesired inputs and prevent the model from generating an output, or they may detect undesired outputs and prevent the system from displaying the generation. Sometimes, such filters are also described as an alignment mechanism. *See infra* Part I.C.8. 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 training-based alignment methods to respond to undesired inputs in a more desirable way. Of course, though, for ChatGPT, we do not know exactly how filters are implemented.

[200] GitHub, *About GitHub Copilot for Individuals*, GITHUB (Aug. 17, 2023), https://docs.github.com/en/copilot/overview-of-github-copilot/about-github-copilot-for-individuals.

[201] *Configuring GitHub Copilot in your environment*, GITHUB (Aug. 17, 2023), https://docs.github.com/en/copilot/configuring-github-copilot/configuring-github-copilot-in-your-environment. *See* https://news.ycombinator.com/item?id=33226515 (for related discussion on the Hacker News forum).

or exact copies of training data is an imperfect process, as we discuss in more detail below.[202]

Finally, each mechanism for making model functionality widely available has its own pricing structures that can ultimately impact the quality of the model and its generations. 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.[203] The developers of commercially released models can typically invest more resources in acquiring and licensing training data.

7. Generation

In general, the output of a generative-AI system—the generation—will depend on three things. First, there is the *choice of deployed system* (which, of course, embeds an implicit choice of model or models). For example, a user that wants to perform text-to-image generation on a browser-based interface needs to select between Ideogram, DreamStudio, ChatGPT, Midjourney, and other publicly available text-to-image applications that could perform this task. There are many reasons to choose one system over another, including cost, speed, convenience, overall quality, quality for specific tasks, and aesthetic preference. A business consultant might choose a text-generation system that produces dry but straightforward outputs, while a creative writer might use a system that produces looser and more surprising outputs as a way to spark their own creativity.

Second, 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 in the background as an oil painting in the style of Paul Cezanne high-resolution aesthetic trending on artstation", are more detailed. Further, users may revise their initial prompt to attempt to create generations that more closely align with their goals. A user who starts with "a big dog" might use "a big dog in the style of Henri Matisse" for a second prompt and "a big dog facing left wearing a spacesuit in the style of Paul Cezanne" for a third as they refine their concept of what they are looking for.

[202] *See infra* Part II.C.

[203] The best open-sourced models are very good, but often still not as good as state-of-the-art closed-source proprietary models. For example, Technology Innovation Institute in Abu Dhabi released the model, Falcon 180B (a 180 billion parameter model), which they claimed was better than Meta's Llama 2 but still behind GPT 4. *Falcon*, TECH. INNOVATION Inst., https://falconllm.tii.ae/falcon.html.

And third, there is *randomness* in each generation.[204] It is typical, for example, for image applications to produce four candidate generations. DALL·E 2, Midjourney, and Ideogram (Figure 5) all do this. Supplying ChatGPT with the same text prompt yields different results (Figure 8).

Although we have been describing primarily short text prompts, it is also possible for users to give generative-AI systems quite large prompts, or to chain together multiple prompts through multiple interactions to iteratively refine the generated outputs.[205] Typically, for a transformer-based LLM, the effective upper limit on the length of a prompt is the model's context window: the number of tokens that the model can keep track of at once.[206] Modern systems can have context windows in the hundreds of thousands of tokens, which corresponds to hundreds of pages of text. With such a large context window, a user could prompt a model with the complete text of a moderately long book—and the output would potentially depend on every word in the book.

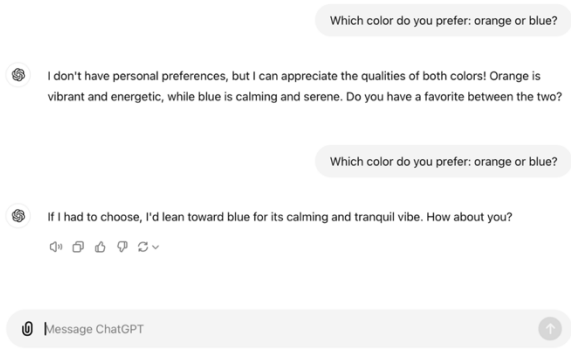


Figure 8. Supplying ChatGPT with the same exact prompt yields different generations. Output produced by the authors.

[204] Recall that, for generative models, there are many reasonable outputs for the same input. *See supra* Part I.A.2.b. 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, *supra* note 89 (for an accessible discussion of decoding).

[205] A user may chain together multiple prompts, or provide examples, in order to guide the generation process. Jason Wei, Xuezhi Wang, Dale Schuurmans et al., Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (2023) (unpublished manuscript), https://arxiv.org/abs/2201.11903. The system may also include software to include context from prior generations supplied by the user. OpenAI says it has implemented automated techniques similar to chain of thought for its o1 model. *See supra* note 38 (discussing chain of thought).

[206] *See supra* note 96 and accompanying text (discussing context windows).

a. Forks in the Supply Chain

Even though it produces outputs, the generation process is not the end of the supply chain. First, there is a loop from generation back to the beginning, because generated outputs can themselves be used as "synthetic" training data for generative-AI models.[207] In this case, generation (stage 7) creates new (potentially) expressive works (stage 1) that are already in a digital format (stage 2), which can be used to start the pipeline all over again.[208]

Second, some generative-AI systems use a technique called retrieval-augmented generation (RAG). RAG involves selecting specific data examples from a dataset and appending them to the user-supplied prompt, in order to guide and constrain the generation process.[209] This is why we also draw an arrow from dataset collection and curation (stage 3)[210] to generation (stage 7) in Figure 6 (shown as the arrow to the dotted box around deployment, alignment, and generation): collected or curated RAG datasets can impact generation. For example, a legal search engine could use a RAG model to identify and summarize the most relevant cases to a user's query. As new cases are decided and added to the caselaw database, they become available as search results to be summarized;

[207] 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. *See generally* Rico Sennrich, Barry Haddow & Alexandra Birch, *Improving Neural Machine Translation Models with Monolingual Data*, *in* 2016 PROC. 54TH ANN. MEETING ASS'N FOR COMPUT. LINGUISTICS (VOLUME 1: LONG PAPERS) 86–96 (2016).

[208] It is an increasingly common practice to create and curate datasets that consist partially or entirely of synthetic data. *See* Gokaslan, Cooper, Collins et al., *supra* note 37. As a result, the boundaries between generation (Figure 6, stage 7) the creation of expressive works, (Figure 6, stage 1), and data creation (Figure 6, stage 2) can be quite porous, and can overlap with both dataset collection and curation (Figure 6, stage 3). There are concerns that this practice can have negative effects on model quality. *See generally* Shumailov, Shumaylov, Zhao et al., *supra* note 131. However, researchers have also found that careful curation can mitigate at least some of these concerns. *See* Gunasekar, Zhang, Aneja et al., *supra* note 131 (regarding the use of synthetic textbooks to train high-quality text-to-text models).

[209] Generative-AI systems using retrieval augmented generation identify relevant examples from a database and append those retrieved examples to the user-supplied prompt. In practice, retrieved examples may be from a dataset (either the training dataset or a separate, retrieval data set) or from a service (such as incorporating data from the output of a plugin that performs an Internet search). *See generally* Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann et al., *Improving Language Models by Retrieving from Trillions of Tokens*, 162 PROC. MACH. LEARNING RSCH. 2206–40 (2022) (for an introduction to retrieval-based models). *See generally* Min, Gururangan, Wallace et al., *supra* note 149 (for an example of a retrieval model with a separate retrieval database).

[210] *See supra* Part I.C.3.

for such a model, these cases are not used to retrain the model, but they are reflected in the output—skipping directly from stage 3 to stage 7.

Third, some generative-AI systems interact with *external* deployed services as part of the generation process. We discussed above some deployed generative-AI systems that have developer APIs, which give users the ability to integrate generative-AI functionality into their own applications. It is similarly possible for generative-AI system deployers to integrate their code with other services on the web.

To make this concrete, consider OpenAI's ChatGPT plugins. Plugins enable ChatGPT to integrate with other products and services, including "Expedia, FiscalNote, Instacart, KAYAK, Klarna, Milo, OpenTable, Shopify, Slack, Speak, Wolfram, and Zapier"[211] in order to shape output generations. Since several of ChatGPT's underlying GPT model(s) were trained in 2021,[212] some of the information it has learned is out-of-date. One stated purpose of plugins is to address delays in training updates—to give ChatGPT access to more recent data acquired from other web-hosted services to improve the quality of generations.[213] For example, one of the use cases on the OpenAI website involves a user querying for information about 2023 Oscar winners. To produce the corresponding generation, ChatGPT is illustrated as performing a web search, retrieving the recent winners list, and appearing to summarize (in user-requested poetic format) the 2023 winners.[214] As this example shows, interactions with external services couple a generative-AI service to other services that have their own complex supply chains.

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. During pre-training, they monitor evaluation metrics and may pause or

[211] OpenAI, *ChatGPT plugins,* OPENAI (Mar. 23, 2023), https://openai.com/index/chatgpt-plugins.

[212] According to generations produced by the authors in August 2023, when we prompted with queries whose answers depended on more recent information. While it is public information that the pre-training cutoff date for the (now deprecated) GPT-3.5 is September, 2021, explicit cutoff dates for GPT-4 are not public information. *See, e.g.*, Winne Nwanne, *Comparing GPT-3.5 & GPT-4: A Thought Framework on When To Use Each Model*, AZURE AI SERVICES BLOG (March 18, 2024), https://techcommunity.microsoft.com/blog/azure-ai-services-blog/comparing-gpt-3-5--gpt-4-a-thought-framework-on-when-to-use-each-model/4088645/ (discussing training cutoff dates).

[213] "By integrating explicit access to external data—such as up-to-date information online, code-based calculations, or custom plugin-retrieved information—language models can strengthen their responses with evidence-based references." OpenAI, *supra* note 214.

[214] *Id.*

restart the process to alter the datasets and algorithm being used;[215] during fine-tuning, they continue training the base model with data that is specifically relevant for a particular task.[216] But both of these types of modifications are relatively coarse. While they adjust the dataset and algorithm, they do not explicitly incorporate information into the model about whether specific generations are "good" or "bad" according to user preferences.[217]

A major active area of machine-learning research, called model alignment, attempts to provide more granular improvements to models to bring them closer to developers' and users' goals for specific prompts and generations.[218] The overarching aim is to *align* model outputs with specific generation preferences (Figure 6, stage 8). One of the most popular alignment techniques is reinforcement learning with human feedback (RLHF).[219] As the name suggests, RLHF updates the model by combining collected human feedback data with a (reinforcement-learning-based) algorithm. Human feedback data can take a variety of forms, including user ratings of generations. For example, such ratings can be collected by including thumbs-up and thumbs-down buttons in an application's user interface, letting users vote on whether they were satisfied with the generation they received. In turn, the reinforcement-learning algorithm uses these ratings to adjust the model in ways that are intended to make more "thumbs-up" generations and fewer "thumbs-down" ones.[220]

---

[215] *See supra* Part I.C.4.

[216] *See supra* Part I.C.5.

[217] Of course, words like "good" and "bad" can have multiple valences and resist the kind of quantification on which machine learning depends. *See* Lee, Ippolito & Cooper, *supra* note 67, at 5 (discussing the challenges of defining "good" and "bad" in the context of model behavior).

[218] *See* Ryan Lowe & Jan Leike, *Aligning language models to follow instructions*, OPENAI (Sept. 2, 2023), https://openai.com/research/instruction-following (for an introduction to InstructGPT, a model that is aligned with human feedback).

[219] Paul Christiano, Jan Leike, Tom B. Brown et al., Deep reinforcement learning from human preferences (June 12, 2017) (unpublished manuscript), https://arxiv.org/abs/1706.03741v1; Long Ouyang, Jeff Wu, Xu Jiang et al., Training language models to follow instructions with human feedback (2017) (unpublished manuscript), https://arxiv.org/pdf/2203.02155.pdf.

[220] In reinforcement learning, data are not labeled as explicitly as in the discriminative setting, e.g., our example of an image classifier, where each training data image has a label of either cat or dog. *See supra* Part I.A.2.a. 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. *See* Yuntao Bai, Saurav Kadavath, Sandipan Kundu et al., Constitutional AI: Harmlessness from AI Feedback (Dec. 15, 2022) (unpublished manuscript), https://arxiv.org/abs/2212.08073 (using reinforcement learning with AI-generated feedback).

While this example includes feedback from end users, most generative-AI companies begin model alignment prior to deployment or release.[221] Before making models publicly available, these companies contract with firms like Scale AI,[222] which 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.[223] 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[224] and toward the policies of the company or organization that deployed the model.[225] In this respect, model alignment complements other techniques, like input-prompt and output-generation filtering,[226] in generative-AI systems.

II. TRACING COPYRIGHT THROUGH THE SUPPLY CHAIN

Part I provided our first main contribution: the generative-AI supply chain.[227] The supply chain is our framework for reasoning about the diversity and complexity of generative AI—the technical artifacts, timelines, and actors involved in the production, deployment, and use of generative-AI systems. To frame this contribution, we first needed a broad introduction to machine learning and generative AI. That introductory material was intended primarily for copyright lawyers, although we hope that AI developers also find useful the zoomed-out overview of the parts of their field that are relevant to copyright law.

Now it is time to flip the script. This Part will provide our second main contribution: describing precisely the copyright implications for generative AI by analyzing United States statutes and caselaw in relation to the stages of the

---

[221] *See supra* Part I.C.6.

[222] SCALE AI, *supra* note 132.

[223] This type of feedback is hardly unique to generative AI. Search engines, for example, extensively use human raters to obtain feedback on whether search results are useful and responsive to user queries. However, the way the feedback is used for generative AI is, of course, different.

[224] Samantha Cole, *'Life or Death:' AI-Generated Mushroom Foraging Books Are All Over Amazon*, 404 MEDIA (Aug. 29, 2023), https://www.404media.co/ai-generatedmushroom-foraging-books-amazon/ (for an example of harmful outputs: a book on mushroom foraging built from generations that mistakenly indicate that toxic mushrooms are safe to eat).

[225] *See* James Manyika, *An overview of Bard: an early experiment with generative AI*, GOOGLE (Aug. 17, 2023), https://ai.google/static/documents/google-about-bard.pdf; OpenAI, *Our approach to AI safety*, OPENAI (Apr. 5, 2023), https://openai.com/blog/our-approach-toai-safety; Deep Ganguli, Amanda Askell, Nicholas Schiefer et al., The Capacity for Moral Self-Correction in Large Language Models (Feb. 15, 2023) (unpublished manuscript), https://arxiv.org/abs/2302.07459 (documenting safety considerations, alignment, and RLHF at Google, OpenAI, and Anthropic).

[226] *See supra* Part I.C.6. Sometimes filters are described as an alignment mechanism.

[227] *See supra* Part I.C.

generative-AI supply chain. To this end, similarly to Part I, we provide a broad survey of introductory material on copyright law. This material is intended primarily for generative-AI developers, although we hope that copyright lawyers will find useful the zoomed-out overview of the parts of their field that are relevant to generative AI.

This Part is organized a bit differently than Part I. Whereas in Part I we began with two integrated sections on machine learning and generative AI,[228] this Part applies copyright law to the generative-AI supply chain, one *issue* at a time. We take up these issues in the logical order that they typically arise in a copyright lawsuit. In each section, we first present the relevant introductory material on copyright law.[229] Our goal is to be careful and systematic, not to say anything dramatically new. We then rely on this material to analyze the copyright implications of each link in the generative-AI supply chain. Copyright lawyers and others who are familiar with copyright law should feel free to skip the introductory material as they see fit. Those for whom this material is new may find it most helpful to read straight through.

This is a long Part, so a brief outline is in order. Copyright protects original works of authorship fixed in a tangible medium of expression.[230] A defendant directly infringes the original work of an author when they engage in conduct implicating one of several enumerated exclusive rights (reproducing, publicly distributing, etc.)[231] with a work of their own that is substantially similar to a copyrighted work[232] because it was copied from that work.[233] Other parties may be held secondarily liable for conduct that bears a sufficiently close nexus to the infringement under one of several theories.[234] Otherwise-infringing conduct is legal when it is protected by one of several defenses, including the DMCA Section 512 safe harbors,[235] fair use,[236] or an express[237] or implied[238] license.[239] In

[228] *See supra* Part I.A; *supra* Part I.B.

[229] With some exceptions: in some cases, for simplicity of presentation, we deviate slightly from this format. *See, e.g., infra* Part II.B (which is organized in relation to each exclusive right).

[230] *See infra* Part II.A.

[231] *See infra* Part II.B.

[232] *See infra* Part II.C.

[233] *See infra* Part II.D.

[234] *See infra* Part II.E (direct infringement); *infra* Part II.F (indirect infringement).

[235] *See infra* Part II.G.

[236] *See infra* Part II.H.

[237] *See infra* Part II.I.

[238] *See infra* Part II.J.

[239] These are not the only defenses. The defense that the copied portions of the work are uncopyrightable is discussed in Part II.C (where we discuss substantial similarity), and the defense that the defendant's work was independently created is discussed in Part II.D (where we discuss proving copying). We omit discussion of defenses that are less

addition, we consider the different remedies that courts can award when they find infringement: damages and profits, statutory damages, attorney's fees, injunctions, and destruction of generative-AI models.[240] Finally, we discuss three types of copyright-like rights: interference with copyright management information,[241] right of publicity,[242] and hot news misappropriation.[243]

A. Authorship

Although this is primarily an Article about copyright infringement, it is important to start by discussing what can be copyrighted in the first place. For one, not everything that can be used as training data can be the subject of copyright, so some training data will always be outside of copyright's reach. For another, every copyrighted work contains some uncopyrightable aspects, which complicates the infringement analysis. For a third, copyrightability specifically affects some stages of the generative-AI supply chain—stages that may produce copyrightable artifacts. And for a fourth, the question of copyrightability is of independent interest to many generative-AI developers and users.

1. Copyright Law

United States copyright law protects "(1) original works of authorship, (2) fixed in any tangible medium of expression."[244] "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."[245] 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."[246]

We start with fixation, which requires only brief discussion. 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 unrecorded dances and songs that are never recorded, will never be captured in a way that can be used as training-data inputs to a training algorithm. Datasets, models, applications, prompts, and generations are all fixed in computers and storage devices. That is, *every step of the supply chain from step 2 (the creation of data from expression)*

---

immediately relevant to generative AI, such as the statutory licensing rules for satellite transmissions of television programming in 17 U.S.C. §§ 119, 122.

[240] *See infra* Part II.K.

[241] *See infra* Part II.L.

[242] *See infra* Part II.M.

[243] *See infra* Part II.N.

[244] 17 U.S.C. § 102(a) (numbering added).

[245] Feist Publ'ns v. Rural Tel. Serv. Co., 499 U.S. 340, 345 (1991).

[246] 17 U.S.C. § 101 (definition of "fixed").

onwards is at least potentially fixed, because every one of those steps takes place on computers.[247]

Thus, we turn instead to originality. Any kind of original expression can be used as inputs to various parts of the supply chain (e.g., training data, prompts). Copyrightable subject matter explicitly includes "literary works" (e.g. poems, novels, FAQs, and fanfic),[248] "musical works" (e.g., sheet music, and MIDI files),[249] "pictorial . . . works" (e.g. illustrations and photographs),[250] "audiovisual works" (e.g., Hollywood movies and home-recorded TikToks), [251] "sound recordings" (e.g., pop songs and live comedy recordings),[252] and more. But this list is nonexclusive. Any kind of creative expression that appeals to the eye or the ear is copyrightable.[253] And copyright law does not discriminate among works based on their quality, their morality, or their importance.[254]

A little more surprisingly, perhaps, the copyright in a work does not depend on the amount of work that went into creating it—the so-called "sweat of the brow."[255] A work consisting only of facts is uncopyrightable, even if those facts required extensive effort to observe.[256] And if a work consists of the *output* of a process, the fact that the process is technically complicated and computationally intense is irrelevant if the output itself is unoriginal.[257]

Copyright law does, however, require that creative expression have a *human* author. Animals are not recognized as "authors" for copyright purposes.[258] Neither are supernatural beings.[259]

Thus, the originality requirement distinguishes material that was created by a human author from facts and other materials that "do not owe their origin to an act of authorship."[260] In practice, this means that the copyright in some works

---

[247] The qualification "at least potentially" is required because some copies might be fixed for too short a period of time to count. *See infra* Part II.B.

[248] 17 U.S.C. § 102(a)(1).

[249] *Id.* § 102(a)(2).

[250] *Id.* § 102(a)(5).

[251] *Id.* § 102(a)(6).

[252] *Id.* § 102(a)(7).

[253] Christopher Buccafusco, *Making Sense of Intellectual Property Law*, 97 CORNELL L. REV. 501 (2012).

[254] Bleistein v. Donaldson Lithographing Co., 188 U.S. 239, 251 (1903).

[255] *Feist*, 499 U.S. at 351–56.

[256] *Id.*

[257] *See* James Grimmelmann, *Three Theories of Copyright in Ratings*, 14 VAND. J. ENT. & TECH. L. 851, 878–79 (2011) (criticizing theory that outputs "resulting from a minimally creative process" are thereby copyrightable).

[258] *See* Naruto v. Slater, 888 F.3d 418 (9th Cir. 2018).

[259] The issue typically arises when someone asserts copyright in a religious text that they claim was dictated by a supernatural being. *See, e.g.,* Urantia Found. v. Maaherra, 114 F.3d 955 (9th Cir. 1997).

[260] Feist Publ'ns v. Rural Tel. Serv. Co., 499 U.S. 340, 347 (1991).

(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, for "thinner" works, the "range of creative choices that can be made in producing the works is narrow."[261] 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 efficient algorithms or coding conventions required by the choice of programming language.[262]

Another important variation on originality involves derivative works and compilations, both of which are works that incorporate one or more existing (or "underlying" works) but also add new authorship. A "derivative work" (think of a translation of a novel, a recording of a song, or an action figure based on a character from a movie) is defined as "a work based upon one or more preexisting works . . . in which [those works are] recast, transformed, or adapted."[263]

Derivative works are copyrightable,[264] but the copyright in the derivative work "extends only to the material contributed by the author of [the derivative] work, as distinguished from the preexisting material employed in the work."[265] This rule reflects the commonsense idea that the translator of a novel, for example, should receive a copyright in their translation, but not a copyright in the novel itself. There is also an important exception to the rule; a derivative work copyright "does not extend to any part of the work in which [the underlying work] has been used unlawfully."[266] In such a case—say, an unauthorized infringing translation—the underlying copyright effectively also gives control over the derivative work.

The translator has in effect performed uncompensated creative labor for the benefit of the novelist.[267] Note that the language is "unlawfully:" if the author of the derivative work has a valid defense to copyright infringement, then the derivative work is indeed copyrightable after all.[268]

---

[261] Rentmeester v. Nike, Inc., 883 F.3d 1111, 1120 (9th Cir. 2018).

[262] Pamela Samuelson, *Functionality and Expression in Computer Programs: Refining the Tests for Software Copyright Infringement*, 31 BERKELEY TECH. L.J. 1215 (2016).

[263] 17 U.S.C. § 101 (definition of "derivative work).

[264] 17 U.S.C. § 103(a).

[265] 17 U.S.C. § 103(b).

[266] 17 U.S.C. § 103(a). The courts have also held, illogically, that even if the underlying work was used with the copyright owner's permission, it is uncopyrightable unless the owner also consents to a derivative copyright. *See, e.g.*, Gracen v. Bradford Exch., 698 F.2d 300 (7th Cir. 1983).

[267] *See, e.g.*, Anderson v. Stallone, 11 U.S.P.Q.2d 1161 (C.D. Cal. 1989) (holding that a script for a *Rocky* sequel was an infringing derivative work, and so its author did not have a copyright in it and could not sue the creators of *Rocky IV* over alleged similarities to his script).

[268] *See, e.g.*, Keeling v. Hars, 809 F.3d 43 (2d Cir. 2015) (holding that a play parodying the action movie *Point Break* was copyrightable as a derivative work because the play was a non-infringing fair use).

Compilations, on the other hand, are works "formed by the collection and assembling of preexisting materials or of data."[269] Examples of compilations include collections of poems, almanacs, and books of traditional folk songs. As these examples show, some compilations are made up of copyrighted works, while others are not. A compilation is copyrightable (separately from any copyright in the materials it is assembled from) when the compilation itself features a sufficiently original "selection or arrangement."[270] Originality in selection is choosing *what to include* in the compilation; originality in arrangement is choosing *how to organize* the compilation. A printed telephone directory, for example, does not display originality in selection (it lists everyone to whom the telephone company provides service) or arrangement (alphabetical order is not original).[271] On the other hand, a book showing the yearbook photos of famous people who graduated from New York City public schools had original selection (the author chose a small number of photos from hundreds of thousands of possibilities) and arrangement (the book was more than just an alphabetical listing).[272]

The substantive difference between compilations and derivative works is that in a compilation, the underlying works are present in substantially unmodified form, whereas in a derivative work the underlying work is "recast, transformed, or adapted." The line dividing the two characterizations is somewhat metaphysical, but it has consequences in some corners of copyright doctrine, which could in turn have consequences for datasets, models, and generations.[273]

Another important limit is that not every original aspect of a work is copyrightable. In the same section that sets out originality as the basis of copyright, the Copyright Act then states that copyright does not "extend to any idea, procedure, process, system, method of operation, concept, principle, or

---

[269] 17 U.S.C. § 101 (definition of "compilation").
[270] Feist Publ'ns v. Rural Tel. Serv. Co., 499 U.S. 340, 348 (1991).
[271] *Id.*
[272] Cantor v. NYP Holdings, Inc., 51 F. Supp. 2d 309 (S.D.N.Y. 1999).
[273] *See, e.g.,* 17 U.S.C. § 203(b)(1) (allowing the creator of an authorized derivative work to continue using it after the author terminates the license in accordance with a statutory procedure). In slightly more detail, authors have a statutory right to terminate licenses they previously granted. That right is available during a five-year window from thirty-five to forty years after the original license. When they do, the former licensee cannot make new copies of the underlying work, or create new derivative works from it—but they can continue to exploit derivative works they have already created. Thus, if a model is a derivative work of licensed training data, the ability to continue using the model would survive the termination of the license. This may seem like an arcane point of copyright law, but many millions of dollars can turn on the effects of a termination of a license. *See, e.g.*, Milne ex rel. Coyne v. Stephen Slesinger, Inc., 430 F.3d 1036 (9th Cir. 2005) (Winnie-the-Pooh); Marvel Characters, Inc. v. Kirby, 726 F.3d 119 (2d. Cir. 2013) (*The Fantastic Four* and other Marvel titles); Everly v. Everly, 958 F.3d 442 (6th Cir. 2020) (the Everly Brothers' songwriting catalog).

discovery."[274] This exclusion is generally described as the "idea/expression dichotomy," i.e., that "copyright assures authors the right to their original expression, but encourages others to build freely upon the ideas and information conveyed by a work."[275] When people talk about the freedom to learn from copyrighted works, this is typically the doctrinal point they are making: copyright in a book about the *Hindenburg* disaster prevents others from copying the author's particular wording, but not from copying his theory about what caused the explosion.[276]

The exclusion of procedures, processes, systems, and methods of operation is particularly important for generative AI. As the Supreme Court explained in the 1879 case of *Baker v. Selden*, copyright can protect a book *about* an accounting method or "the construction and use of ploughs, or watches, or churns," but it does not give the copyright owner "the exclusive right to the art or manufacture described therein."[277] You cannot copy the book about accounting, but you can use the accounting method it describes. Training and generation algorithms, for example, are uncopyrightable processes.

In some cases, uncopyrightable ideas and copyrightable expression are so closely bound up with each other as to be inseparable. In *Baker* itself, the defendant was sued for selling blank forms that were similar to the blank forms the copyright owner sold for using the accounting method to write down transactions and keep running totals. The Supreme Court held that the forms "are to be considered as necessary incidents to the art," and were therefore free for anyone to copy.[278] This rule is known today as the merger doctrine. If there are "at best only a limited number" of ways to express an idea, none of those expressions are copyrightable, because otherwise "a party or parties, by copyrighting a mere handful of forms, could exhaust all possibilities of future use of the substance."[279]

With these principles in mind, we proceed along the supply chain, considering which artifacts are and are not copyrightable.

---

[274] 17 U.S.C. § 102(b).
[275] *Feist*, 499 U.S. at 349–50. *See also* Mazer v. Stein, 347 U.S. 201, 217 ("Unlike a patent, a copyright gives no exclusive right to the art disclosed; protection is given only to the expression of the idea—not the idea itself.").
[276] Hoehling v. Universal City Studios, Inc. 618 F.2d 972 (2d Cir. 1980).
[277] Baker v. Selden, 101 U.S. 99, 102 (1879).
[278] *Id.* at 103.
[279] Morrissey v. Procter & Gamble Co., 379 F.2d 675, 678–79 (1st Cir. 1967).

2. Application to the Generative-AI Supply Chain

a. Expressive Works

Some of the individual examples that serve as training data are uncopyrightable. For example, birdsong-recognition AI models are trained on recordings of birds.[280] Currently, synthetic training data,[281] which are produced by generative-AI systems and then used as inputs for training other generative-AI models,[282] are not copyrightable.[283] For example, Microsoft has trained its phi family of transformer-based large language models on a mix of different data sources that include synthetic, noncopyrightable data generated by OpenAI GPT models.[284] But other items are copyrightable, and those copyrights will be held by a variety of authors: photographers, writers, illustrators, musicians, programmers, and other creators of all stripes.

b. Data

Turning expressive works into computer-readable data does not generally change their copyright status. Digitizing preexisting materials does not generally add enough expression to make a derivative work.[285] The choice to use one file format rather than another can affect how a work is encoded, but the expression in a work is substantially unchanged.[286]

---

[280] *See* Stefan Kahl, Connor M. Wood & Holger Klinck, *BirdNET: A Deep Learning Solution for Avian Diversity Monitoring*, 61 ECOLOGICAL INFORMATICS 101236 (2021). Animals are not recognized as "authors" for copyright purposes. *See* Naruto v. Slater, 888 F.3d 418 (9th Cir. 2018).

[281] For discussion of synthetic data in the supply chain, *see supra* Part I.B; *supra* Part I.B.2.b; *supra* Part I.C; *supra* Part I.C.1; *supra* Part I.C.7.

[282] Gokaslan, Cooper, Collins et al., *supra* note 37; Gunasekar, Zhang, Aneja et al., *supra* note 131.

[283] The U.S. Copyright Office has argued that there is no human author, making such works ineligible for copyright. Thaler v. Perlmutter, No. 22-1564 (D.D.C Aug. 18, 2023). *See infra* Part II.A.2.g (discussing generations).

[284] Microsoft's phi family, similar to Meta's Llama series, includes models of multiple sizes and versions. The phi models are trained on both web-scraped data and synthetic data. *See, e.g.,* Gunasekar, Zhang, Aneja et al., *supra* note 131 ("phi-1, a new large language model for code, … is a Transformer-based model with 1.3B parameters, trained for 4 days on 8 A100s, using a selection of 'textbook quality' data from the web (6B tokens) and synthetically generated textbooks and exercises with GPT-3.5 (1B tokens).").

[285] *See, e.g.*, Meshwerks, Inc. v. Toyota Motor Sales USA, Inc., 528 F.3d 1258 (10th Cir. 2008); Bridgeman Art Library, Ltd. v. Corel Corp., 36 F. Supp. 2d 191 (S.D.N.Y. 1999).

[286] *See* A. Feder. Cooper & James Grimmelmann, *The Files are in the Computer: On Copyright, Memorization, and Generative AI*, CHI.-KENT L. REV. (forthcoming) (discussing representation of expressive works in digital files).

User-supplied prompts are another important form of data. When the user of the service supplies a prompt to a generative-AI system, the service host may save that prompt for later use. The service host may use the prompt as additional training data for fine-tuning or aligning the existing model, or for training another model altogether.[287] As a result, user-created data can find its way into a generative-AI system at the training, fine-tuning, and alignment stages, not just through prompts during generation.

Some prompts, such as "a cat", will be short phrases that are not sufficiently original to support a copyright.[288] But other prompts will be longer, sometimes much longer. For example, as noted above, it is currently technologically feasible to prompt a text-to-text system with an entire book.[289] That book could be an existing copyrighted work, in which case the prompt is not separately copyrightable. Or it could be an original work composed by the user on the spot, in which case the prompt is copyrightable, and the user will be the author and initial copyright owner. Many, indeed, most, prompts will fall somewhere between these two extremes.

c. Training Datasets

Moving forward along the supply chain, then, different datasets[290] will include different amounts and proportions of copyrighted material. A dataset of birdsong recordings will consist entirely, or almost entirely, of uncopyrighted material. A dataset of illustrations, on the other hand, will contain numerous copyrighted works. A dataset of photographs paired with synthetic captions will contain both copyrighted and copyright-free material.[291]

Datasets *themselves* may be copyrightable as compilations, provided that they meet the fixation and originality requirements.[292] Most datasets are based on extensive curatorial decisions—what data to collect and include, how format it, and so on.[293] In many cases, these decisions will fall short of the selection-and-arrangement originality threshold. But in others, it will be possible to identify specific choices that went into intentionally creating a dataset with particular desired attributes, and show that those choices were extensive enough to reach the

[287] *See supra* Part I.C.7; *supra* Part I.C.8
[288] *See* 37 C.F.R. § 202.1(a).
[289] Anthropic, *supra* note 96.
[290] *See supra* Part I.C.3.
[291] Gokaslan, Cooper, Collins et al., *supra* note 37 (discussing the CommonCatalog dataset, which contains training-data examples that each consist of a Creative-Commons licensed image and a corresponding, synthetically generated captions).
[292] 17 U.S.C. § 103(a).
[293] *See supra* Part I.C.3. *See generally* Lee, Ippolito & Cooper, *supra* note 67, at 5.

compilation-copyright threshold.[294] Shutterstock's content datasets, for example, contain carefully selected sets of images arranged with metadata, including text descriptions.[295] In contrast, Harvard Law School's Caselaw Access Project dataset consists of public domain data from 360 years of United States caselaw. The project aims to collect all written caselaw (i.e., to select everything possible) and does not arrange the materials in a novel way; the dataset was produced by "scanning the entirety of the Harvard Law School Library's physical collection of American caselaw and made … machine-readable in a consistent format available online."[296]

d. Pre-Trained/Base Models

Attributing authorship in models is even trickier to classify for two reasons.[297] First, there is the question of whether a model possesses the necessary "modicum of creativity" to be a work of authorship at all.[298] In some cases, the answer is probably "no": applying an existing algorithm and well-known architecture to an existing dataset[299] does not involve sufficient creative choices. Any expression in such a model merges into the idea and is uncopyrightable.[300]

But it is possible that other models possess the necessary originality. For one thing, when a training dataset is curated specifically for training a base model, the model may supplant the dataset as the relevant "work" from the data curation process, just as a finished film is regarded as the "work" rather than the (much larger) dataset of raw footage.[301] In such a case, the model would inherit the

---

[294] Indeed, individual data examples can contain multiple components, whose curation could make the individual example (i.e., the assemblage of multiple components) eligible for copyright. For example, text-to-image models are trained on image-caption pairs; the image, the caption, and the image-caption pair (as a compilation) could each potentially constitute distinct works of authorship, and each of these copyrights could be owned by someone other than the dataset creator or curator.

[295] Harish Gaur, Giselle Goicochea, Darshana Sivakumar et al., *Shutterstock's Content Datasets Now on Databricks Marketplace*, DATABRICKS (Jun. 6, 2024), https://www.databricks.com/blog/shutterstocks-content-datasets-now-databricks-marketplace.

[296] Harvard Caselaw Project, *History of the Caselaw Access Project*, LIBRARY INNOVATION LAB (accessed Oct. 4, 2024), https://case.law/about/.

[297] It is worth noting that many model trainers certainly believe that models are copyrightable, and have released those models under licenses that are only intelligible if there is something copyrightable to license in the first place.

[298] Feist Publ'ns v. Rural Tel. Serv. Co., 499 U.S. 340, 346 (1991).

[299] With standard choices of hyperparameters, on standard hardware, etc.

[300] *See generally* Pamela Samuelson, *Reconceptualizing Copyright's Merger Doctrine*, 63 J. COPYRIGHT SOC'Y USA 417 (2016) (describing merger doctrine). *See supra* Part II.A.1.

[301] *See generally* Margot E. Kaminski & Guy A. Rub, *Copyright's Framing Problem*, 64 UCLA L. REV. 1102 (2017) (discussing the problem of identifying the 'work' in copyright cases).

creative choices that went into curating the dataset. For another, base models are often the results of extensive design processes that involve novel architectures and algorithms. While these processes are not themselves copyrightable,[302] and originality in a process is not a guarantee that the outputs are copyrightable, it is at least possible that a model's creators[303] will have made creative choices that imbue the model with original expression.

It is not obvious how to classify this expression. One view is that a model is a compilation of its training data—the model is simply a different and complicated arrangement of training examples. Another view is that a model is a derivative work of its training data that itself meets the threshold of originality. In our tentative view, the derivative-work characterization is more persuasive. The stochastic and overlapping way in which models encode features of training examples does not just amount to the selection and arrangement of those examples; instead, they are altered and combined in ways that more closely fit the language "recast, transformed, or adapted" in the definition of a derivative work.[304]

The second reason that models might or might not be copyrightable is even trickier. In some ways, a model is like a computer program,[305] which is classified for copyrightability purposes as a literary work "expressed in words, numbers, or other verbal or numerical symbols or indicia,"[306] and subject to a screen that excludes from protection those aspects that are dictated by its functionality.[307] But in other ways, this analogy is inapt. A model straddles the line between data structure and executable program, a model is not obviously a *literary* work, and the line between functional and non-functional aspects of a model is extremely difficult to draw. This is a question that will require careful consideration by copyright lawyers and scholars.

e. Fine-Tuned Models and Aligned Models

The two authorship considerations that we raise above for pre-trained models also apply to fine-tuned and aligned models. We start with the second point, which is simpler: like pre-trained models, both fine-tuned and aligned models will face

[302] *See* 17 U.S.C. § 102(b).
[303] In this case, this includes the parties that designed the architectures and algorithms.
[304] It is also possible to argue that a model is made up of non-expressive "statistical patterns" in the expressive works it was trained on. We believe that this is a distinction without a difference. *See* Cooper & Grimmelmann, *supra* note 289.
[305] *See supra* note 162 and accompanying text.
[306] 17 U.S.C. § 101 (definition of "literary works").
[307] *See generally* Comput. Assocs. Int'l, Inc. v. Altai, 982 F.2d 693 (2d Cir. 1992) (standard case on software copyright); Pamela Samuelson, Randall Davis, Mitchell D. Kapor & Jerome H. Reichman, *A Manifesto Concerning the Legal Protection of Computer Programs*, 94 COLUM. L. REV. 2308 (1994) (lucid and time-honored analysis of software copyright).

similar issues of categorization for copyright law. Thus, to the extent that it is separately copyrightable, a fine-tuned and/or aligned model would typically be a derivative work of the base model it was trained from.

The first point—that training choices can imbue models with creative attributes—leads to different observations for fine-tuning and model alignment. Both these processes can be goal-directed in ways that bear on authorship: the trainer applies the process with the goal of creating a model that has specific intended characteristics. The model trainer is typically optimizing the model's behavior in generating specific desired outputs—the kind of nexus between human choices and resulting material that characterizes copyrightable authorship.[308] The same is true for model alignment. Further, if, for example, a prompt is incorporated as part of the input to RLHF,[309] then the prompt serves as training data that could update the model. In this case, the training data *itself* is created in a process that includes human choices and has been crafted with specific creative goals in mind. Some prompts used for RLHF are harvested from users;[310] others are created by companies' employees or contractors specifically for RLHF. Either way, to the extent that the model is a derivative work of these prompts, the copyright ownership of the prompts is a relevant consideration.

f. Deployed Services

It is well established that software is copyrightable.[311] The non-model parts of a user-facing application or developer API will be protected by copyright (subject to the functionality screen noted above).[312]

g. Generations

Generations raise a doctrinal question that has been debated for decades: who, if anyone, owns the copyright in the output of a computer program?[313] Although

[308] *See generally* Dan L. Burk, *Thirty-Six Views of Copyright Authorship, by Jackson Pollock*, 58 HOUS. L. REV. 263 (2020) (discussing causal elements of authorship); Shyamkrishna Balganesh, *Causing Copyright*, 117 COLUM. L. REV. 1 (2017) (same).
[309] *See supra* Part I.C.8 (discussing RLHF).
[310] It is not known with certainty whether user prompts are used for RLHF training for proprietary systems like ChatGPT, but ChatGPT's data controls FAQ states, "we use data to make our models more helpful for people. ChatGPT, for instance, improves by further training on the conversations people have with it, unless you choose to disable training." *Data Controls FAQ*, OPENAI, https://help.openai.com/en/articles/7730893-data-controls-faq.
[311] *See supra* Part II.A.1.
[312] *See* Charles Duan, *What Is Copyrightable In Software?* (unpublished manuscript, draft on file with authors).
[313] Pamela Samuelson, *Allocating Ownership Rights in Computer-Generated Works*, 47 U. PITT. L. REV. 1185 (1985).

some commentators have argued that the program itself should be regarded as the author, computer authorship is squarely foreclosed by U.S. copyright law.[314] Computers are not deemed capable of playing the social roles that society and the legal system expect and require of authors.[315] So far, U.S. courts have held firm to this line for AI generations. In *Thaler v. Perlmutter*, the court upheld the Copyright Office's refusal to register copyright in an image allegedly "autonomously created by a computer algorithm running on a machine."[316] The Copyright Office had held that the image lacked human authorship, and the court agreed: computer programs, like animals, are not "authors" within the meaning of the Copyright Act.[317]

Instead, the author (and thus copyright owner) of a generation—if anyone—is some human connected to the generation.[318] The four immediately relevant possibilities are (1) an author or authors whose works the model was trained on, (2) some entity in the generative-AI supply chain (e.g., the model trainer, model fine-tuner, or application developer), (3) the user who prompted the application or API for the specific generation, or (4) no one. Between these four possibilities, there is no one-size-fits-all answer.

Before discussing these four possibilities, we note that it may seem intuitively attractive to consider generations to be analogous to collages. However, while this may seem like a useful metaphor,[319] it can be misleading in several ways. For one, an artist may make a collage by taking several works and splicing them together to form another work. In this sense, a generation is not a collage: a generative-AI system does not take several works and splice them together like a teenager's scrapbook page. Instead, as we have described above, generative-AI systems are built with models trained on many data examples, and the patterns they learn about training data are often about stylistic or structural features at a higher level of abstraction than specific excerpts of a training work.[320] Moreover, those data examples are not explicitly referred back to during the generation process. Instead, the extent to which a generation resembles specific data examples is dependent

[314] James Grimmelmann, *There's No Such Thing as a Computer-Authored Work—And It's a Good Thing, Too*, 39 COLUM. J.L. & ARTS 403 (2016).
[315] Carys Craig & Ian Kerr, *The Death of the AI author*, 52 OTTAWA L. REV. 31 (2020).
[316] Thaler v. Perlmutter, No. 22-1564 (D.D.C Aug. 18, 2023).
[317] *Id.*
[318] The use of synthetic data complicates the question of authorship even further. The generations produced by a model trained on synthetic data generated by another model are doubly removed from the original training data. All of the authorship questions discussed in this Article arise twice over, once for each model and its supply chain. *See supra* note 37 and accompanying text (discussing synthetic data in generative AI).
[319] *See* Cooper, Lee, Grimmelmann, Ippolito et al., *supra* note 10, at 4–5 (detailing how metaphors can be both helpful and misleading for intuiting the behavior of generative-AI systems).
[320] *See* Cooper & Grimmelmann, *supra* note 289; Cooper, Lee, Grimmelmann, Ippolito et al., *supra* note 10, app. B (rebutting collage metaphor).

on the model encoding in its parameters what the specific data examples look like, and then effectively recreating them.[321] Ultimately, it is nevertheless possible for a generation to look like a collage of several different data examples;[322] however, it is debatable whether the process that produced this appearance meets the definition for a collage. There is no author "select[ing], coordinat[ing], or arrang[ing]"[323] training examples to produce the resulting generation. With this in mind, we assess the four relevant authorship possibilities for generations along two dimensions.

A copy of a work from the training set. We start with a generation that closely resembles a work in the training set. If the generation is actually identical to the training example—if it contains no original expression beyond what was present in the input work—then it is simply a copy of that underlying work and not a new copyrightable work at all.[324] Of course, the copyright owner remains the original author, i.e., possibility (1). The generation would be infringing, unless the original work is nonprotectable or in the public domain.

Derivative of a work from the training set. If the generation is, however, a derivative work of the underlying work that incorporates new authorship, a new copyright may subsist in it. [325] If the generation infringes, then it is uncopyrightable under Section 103 and the answer is possibility (4): there is no separate copyright in the generation, even though it contains original authorship.[326] In such a case, the underlying copyright effectively also gives control over the generation.

Assuming, however, that the generation is sufficiently distinct from training data not to be "used unlawfully," a copyright owned by one of its creators may arise.[327] Some models and applications will produce original generations with minimal user input, in which case the user is not an author who can claim copyright in the output,[328] which is possibility (2) above. The Draw Things iOS app, for example, suggests the prompt "8k resolution, beautiful, cozy, inviting, bloomcore, decopunk, opulent, hobbit-house, luxurious, enchanted library in giverny flower garden, lily pond, detailed painting, romanticism, warm colors,

[321] *See infra* Part II.C.2.
[322] *See infra* Part II.H.
[323] 17 U.S.C. § 101 (definition of "compilation").
[324] *See infra* Part II.C (concerning memorized training data and substantial similarity).
[325] *See* 17 U.S.C. § 103(b) ("The copyright in such [a derivative] work is independent of . . . any copyright in the preexisting material.").
[326] 17 U.S.C. § 103(a) ("[Copyright] protection for a [derivative] work . . . does not extend to any part of the work in which such material has been used unlawfully."). The courts have also held, illogically, that even if the underlying work was used with the copyright owner's permission, it is uncopyrightable unless the owner also consents to a derivative copyright. *See, e.g.,* Gracen v. Bradford Exch., 698 F.2d 300 (7th Cir. 1983).
[327] For derivative copyright purposes, lawful use includes fair use. *See, e.g.,* Keeling v. Hars, 809 F.3d 43 (2d Cir. 2015).
[328] *See* [Zarya of the Dawn letter], *infra* note 340.

digital illustration, polished, psychedelic, matte painting trending on artstation." The user who taps "Generate" on the app user interface has contributed no authorship to the resulting image. This Person Does Not Exist is a website that creates a new (and uncannily realistic) deepfake photograph of a nonexistent person each time it is reloaded. The user who visits the site and clicks "reload" is not an author. If anyone can claim authorship credit here, it is the creators of these apps. In some cases, there could be a plausible case that they have contributed enough authorship, just as the creators of a videogame have contributed sufficient authorship even though "the entire sequence of all the sights and sounds of the game are different each time the game is played."[329]

In other cases, the user will make substantial creative inputs through their choice of prompt. In addition to the authorship inhering in the prompt itself (if any), two additional factors push towards making the user the copyright owner rather than the developer—i.e., possibility (3) from above. First, there is their causal responsibility for making the generation exist;[330] here, as in infringement, copyright law may care who "pushes the button."[331] Second, the providers of many generation applications have decided that, as a practical matter, they are uninterested in asserting copyright over the outputs. This is a business choice first and a copyright matter second, but widespread business practices often affect courts' decisions about how to allocate copyright ownership.[332]

Still, it is too hasty to say that the user is necessarily the owner of copyright in a generation, even once the training-data authors and model developers are out of the picture. It is also possible that *no one at all* owns a copyright in the generation (possibility (4)). The problem is that the generation may not be the product of sufficient human authorship. Consider the prompt.[333] "Scary lighthouse" is too short to contain sufficient originality to support a copyright;[334]

[329] Stern Elecs., Inc. v. Kaufman, 669 F.2d 852, 856 (2d Cir. 1982).
[330] Balganesh, *supra* note 311.
[331] Fox Broad. Co. v. Dish Network LLC, 160 F. Supp. 3d 1139, 1169 (C.D. Cal. 2015).
[332] *E.g.,* Aalmuhammed v. Lee, 202 F.3d 1227, 1233 (9th Cir. 2000) (deferring to Hollywood practice of treating *auteur* directors as the "master mind[s]" behind films); Thomson v. Larson, 147 F.3d 195 (2d Cir. 1998) (deferring to theatrical crediting practices in holding that a dramaturg was not a co-author of a musical).
[333] Mark Lemley argues that in fact the prompt is the relevant unit of originality and is in effect the work itself. Mark A. Lemley, *How Generative AI Turns Copyright Law on its Head*, 25 COLUM. SCI. & TECH. L REV. 190 (2024). (July 26, 2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4517702.
[334] *Cf.* Magic Mktg. v. Mailing Servs. of Pittsburgh, 634 F. Supp. 769 (W.D. Pa. 1986) (holding the phrase "CONTENTS REQUIRE IMMEDIATE ATTENTION!" uncopyrightable).

short phrases are not copyrightable.[335] If this phrase does not have the necessary modicum of creativity by itself, it seems unlikely that the additional choice to use it as a prompt is enough to put it over the threshold.[336] Another way of looking at the problem is that prompts like "Scary lighthouse" do not sufficiently constrain the output to make it the product of human authorship. As the Copyright Office put it when rejecting copyright in images created with Midjourney,

> Because of the significant distance between what a user may direct Midjourney to create and the visual material Midjourney actually produces, Midjourney users lack sufficient control over generated images to be treated as the "mastermind" behind them. . .. [T]here is no guarantee that a particular prompt will generate any particular visual output. Instead, prompts function closer to suggestions than orders, similar to the situation of a client who hires an artist to create an image with general directions as to its contents.[337]

This is not the only possible view. A counter might be that, for pragmatic reasons, the copyright system will or should assign authorship to the user and overlook their minimal contributions.[338] While many current generative-AI systems have primarily text-based interfaces where short prompts might not amount to much creativity, future generative-AI systems will likely have different interfaces that introduce other ways of controlling outputs.[339] But for now, it is the law that some generations are uncopyrightable despite containing material that would easily qualify for copyright if they had been produced manually by a human.[340]

---

[335] 37 CFR § 202.1(a). This complicates the copyrightability of captions in image-caption datasets; it is possible that some captions are too short to contain sufficient originality. *See supra* note 297 and accompanying text (discussing image, caption, and image-caption pair copyrights).

[336] *See* Jane C. Ginsburg & Luke Ali Budiardjo, *Authors and Machines*, 34 BERKELEY TECH. L.J. 343 (2019) (advancing this argument); *see also* Burk, *supra* note 311 (exploring variations).

[337] Letter from Robert J. Kasunic to Van Lindburg, *Re: Zarya of the Dawn (Registration # VAu001480196)* 9–10 (Feb. 21, 2023), https://www.copyright.gov/docs/zarya-of-the-dawn.pdf.

[338] *See, e.g.,* Grimmelmann, *supra* note 317, at 413–14 (discussing this possibility, and its difficulties). As one canonical case puts it, "Having hit upon such a variation unintentionally, the 'author' may adopt it as his and copyright it." Alfred Bell & Co. v. Catalda Fine Arts, 191 F.2d 99, 105 (2nd Cir. 1951).

[339] For example, Ideogram has style tags that can be added to the prompt to modify the output (*Ideogram.AI*, *supra* note 78).

[340] *See* James Grimmelmann, *Copyright for Literate Robots*, 101 IOWA L. REV. 657, 657 (2016) ("Almost by accident, copyright law has concluded that it is for humans only . . . ").

This conclusion, however, is not categorical; "some" is not "all." Not every prompt is too short to be copyrightable, and not every user is a spectator to AI generation.[341] Instead, some generations are the product of careful prompt engineering, in which users craft elaborate prompts to cause AI models to achieve specific aesthetic effects. These generations answer both of the objections above. Their respective prompts are often long and intricate, running to dozens or hundreds of words, well above the short-phrase threshold. And these prompts are the result of an iterative creative process, in which the users have acquired a degree of mastery over the (putatively unpredictable) models they use, at least for specific types of outputs.[342] If an artist who flings a sponge against the wall in frustration is entitled to claim copyright in the resulting accidental spatter of paint, why not a user who deliberately crafts the perfect prompt?[343]

B. The Exclusive Rights

With copyrightability out of the way, we turn now to infringement, which will take up most of the rest of this Part. We focus on infringement of copyrights in training data. This is the most significant theory of infringement alleged so far in lawsuits against generative-AI companies, and it is likely to remain so.

There are three distinct elements to the claim that "defendant *D*'s work infringes plaintiff *P*'s copyright." First, *D* must do something with their work that implicates one of *P*'s *exclusive rights*. For example, there is an exclusive right to publicly display a work, but there is not an exclusive right to look at one. This subsection discusses the types of conduct that can violate the exclusive rights. Second, the thing that *D* does with their work must be *substantially similar* to *P*'s work. Unlike the first, which look at *D*'s conduct in isolation, this test focuses on comparing the expression in *P*'s and *D*'s works. Subsection C discusses this comparison. Third, the similarities in their works must arise because *D copied*

[341] For example, a product called alpaca allows users to upload sketches and transform them into more-complete images with generative AI. Users can further control the generated images with text prompts. These user-provided sketches could have copyrights. As another example, some models have long context windows, which enable them to process long segments of text as inputs. A user may prompt such a model with an entire book as input; a system using retrieval-augmented generation may retrieve documents to guide the generation process that may themselves be copyrighted. *Alpaca ML*, ALPACA ML (2024), https://www.alpacaml.com/ (describing the alpaca software product). *See supra* Part I.C.7 (for a discussion on the long-context window in the Claude product). *See supra* note 212 and accompanying text (describing retrieval augmented generation).

[342] For a particularly disquieting example, *see* Emanuel Maiberg, *Inside the AI Porn Marketplace Where Everything and Everyone Is for Sale*, 404 MEDIA (Aug. 22, 2023), https://www.404media.co/inside-the-ai-porn-marketplace-where-everything-and-everyoneis-for-sale/.

[343] *Alfred Bell*, 191 F.2d at 105 n.23.

from *P*'s work. rather than from coincidence or for some other reason. This too is a distinct question, which we discuss in subsection D.

Turning to the exclusive rights, it is helpful to break down the *prima facie* case of infringement by the relevant exclusive right, rather than by the stage of the generative-AI supply chain. There are five relevant exclusive rights:

- The right to "reproduce the copyrighted work in copies" (the reproduction right).[344]
- The right to "prepare derivative works based upon the copyrighted work" (the adaptation right).[345]
- The right to "distribute copies . . . of the copyrighted work to the public" (the distribution right).[346]
- The right to "perform the copyrighted work publicly" (the performance right).[347]
- The right to "display the copyrighted work publicly" (the display right).[348]

To summarize briefly, every stage in the generative-AI supply chain requires a potentially infringing reproduction and thus implicates copyright. We also examine the other exclusive rights, which present interesting variations on this theme.

Under most circumstances, the remedies for infringement of a work are the same, regardless of whether the defendant violated one exclusive right or several. Thus, whether the defendant violated *any* of the exclusive rights is usually a more significant question than *which* one they violated. Still, it is helpful to work through the exclusive rights, both because it sheds light on the operations of the generative-AI supply chain, and because there are some copyright doctrines that depend on which right or rights are in play.

1. The Reproduction Right

As relevant here, the reproduction right is triggered when a work is reproduced in "copies," which are defined as "material objects . . . in which a work is fixed by any method now known or later developed, and from which the

[344] 17 U.S.C. § 106(1).
[345] *Id.* § 106(2).
[346] *Id.* § 106(3).
[347] *Id.* § 106(4), (6).
[348] *Id.* § 106(5).

work can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device."[349]

The same is true for models and generations.[350] Both trigger the reproduction right when they are created, because they are stored in material objects. A dataset or model will be stored on a physical drive (or array of drives); a generation will be stored on some physical electronic component in the user's device (and perhaps within the generating software service), and so on. Thus, the assembly of a dataset, the training of a model, the production of a generation, or a generative-AI system's use of a user-inputted prompt is a "reproduction" within the meaning of copyright law. All of these activities can infringe: the question is whether the resulting dataset, model, prompt, or generation is substantially similar[351] to the plaintiff's[352] copyrighted work, or if it falls under an exception.[353]

One complication has to do with *how long* a work is fixed. Under the "RAM copy" doctrine, which dates to the 1990s, loading a copyrighted work into a computer's working memory can infringe.[354] (Doing so is often necessary to run a program or to perform a computation on data.) On the other hand, more recent caselaw has held that transient copies do not count for the reproduction right.[355] The leading case, *Cartoon Network LP, LLLP v. CSC Holdings*, held that a buffer that was overwritten every 2.4 seconds was not an infringing reproduction of works that passed through the buffer.

The temporal threshold is not generally an issue for the outputs of stages in the generative-AI supply chain. Datasets, models, applications, prompts, and generations are all typically stored for far longer than the 2.4 seconds in *Cartoon Network*. Instead, the threshold may be more important for the inputs to the different stages. For example, a training example needs to be loaded into working memory to train a model on it. But the details of *how long* the example remains

---

[349] 17 U.S.C. § 101 (definition of "copies"). To be pedantic, a training dataset as such is not a "copy" because the dataset is not a "material object." Instead, the c*omputer or storage device* on which a dataset is stored is the relevant material object, and hence the copy. *See* Cooper & Grimmelmann, *supra* note 289.

[350] The same could also be said for individual data examples within the dataset, which is one of the reasons we distinguish between expressive works and their datafied counterparts. *See supra* Part I.A.1; *supra* Part I.C.1; *supra* Part I.C.2.

[351] *See infra* Part II.C.

[352] Of course, there are different types of actors that can be responsible for each of these reproductions. For example, an application user could supply a reproduction of a copyrighted prompt (for which they do not hold the copyright), and the generative-AI system could in turn store that reproduction in memory. This could happen even for a generative-AI system that only trained its models on public-domain data (i.e., did not violate the reproduction right with respect to training).

[353] *See infra* Parts II.G-J.

[354] MAI Sys. Corp. v. Peak Comput., 991 F.2d 511 (9th Cir. 1993).

[355] Cartoon Network LP, LLLP v. CSC Holdings, Inc., 536 F.3d 121, 128–30 (2d Cir. 2008).

in memory, and *how much it is modified* while it is there, will depend on the training algorithm and architectural details of the environment (e.g., how fast the processors are). Similar considerations apply to the generation process—with similar uncertainties. Some generations run in a fraction of a second; others take minutes or hours.[356]

There is also the problem of purely *internal* reproductions: ones that occur only in the middle of the training or generation process. These algorithms compute numerous new values and often overwrite them repeatedly to conserve memory. Consider, for example, one of the middle stages of the archaeologist generation in Figure 5, where the image is blurry but starting to be recognizable. One of these stages might resemble a copyrighted work more closely than the final output. Again, whether these fall underneath the *Cartoon Network* threshold depends on the details of the algorithm and environment.[357]

Finally, prompts can sometimes be reproductions of existing copyrighted works.[358] When they are, they implicate the reproduction right. Many prompts are too short to be copyrightable or to infringe an existing copyright—but those that are long enough can infringe.

2. The Derivative Right

While the reproduction right is about new copies of an existing work, the derivative right is about new works based on an existing work. It is best understood as making clear that copyright in a work extends beyond literal similarity to incorporate changes of form, genre, and content such as translations, sequels, and film adaptations.[359] A training dataset is probably not a derivative work of any of the works in the dataset; it is more appropriately potentially classified as a compilation.[360]

---

[356] These technical details will generally be known by the company engaging in generative-AI training, but may be a closely guarded trade secret. A copyright owner may not be able to learn these details except via discovery in an infringement lawsuit. *See* Winston Cho, *OpenAI Training Data to Be Inspected in Authors' Copyright Cases*, HOLLYWOOD REPORTER (Sep. 24, 2024), https://www.hollywoodreporter.com/business/business-news/openai-training-data-inspected-authors-copyright-case-1236011291/.

[357] Alternatively, there is a strong fair-use case for these transient internal copies. *See* Grimmelmann, *supra* note 343 (summarizing caselaw).

[358] Anthropic, *supra* note 96.

[359] *See generally* Daniel Gervais, *The Derivative Right, or Why Copyright Law Protects Foxes Better than Hedgehogs*, 15 VAND. J. ENT. & TECH. L. 785 (2013); Pamela Samuelson, *The Quest for a Sound Conception of Copyright's Derivative Work Right*, 101 GEO. L.J. 1505 (2013); Daniel Gervais, *AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines*, 52 SETON HALL L. REV. 1111 (2022).

[360] 17 U.S.C. § 101.

A model is a good example of material that might or might not be an exact reproduction of the works it was trained on; the information about a work in a model may be inexact, and the representation will be extremely different.[361] But, as discussed above, there is a stronger argument that to the extent a model incorporates expression from an example in its training data, it is a derivative work of that example and not simply a reproduction. The Midjourney model that produced images of an adventurous archaeologist transforms and adapts images from the *Indiana Jones* movies. Whether or not it is similar enough to infringe is another question (discussed below); our point for the moment is just that if an exclusive right is implicated at all, the derivative right is a good fit.

Similarly, just as prompts can be exact reproductions of existing works, sometimes they can be derivatives of existing works. And generations can be derivative works of works in the training data. Numerous scholars and users have exhibited generations that are self-evidently variations on existing works—different enough to be derivative works, but similar enough to infringe. Matthew Sag, for example, has shown that multiple models can generate images based on Banksy's iconic stencil of a *girl with a red balloon*.[362] Of course, whether and when a generation is a derivative of any particular work depends on similarity, discussed below.[363]

More troublingly, it might be that the derivative right can be infringed by derivative works that do not by themselves incorporate substantial expression from the plaintiff's work. In *Micro Star v. Formgen Inc.*, the defendant distributed fan-made levels for *Duke Nukem 3D*.[364] The level file format consisted entirely of geometry describing where the *Duke Nukem 3D* game engine should place walls and objects; the engine would then perform rendering using copyrighted art assets, but "[t]he MAP file . . . does not actually contain any of the copyrighted art itself; everything that appears on the screen actually comes from the art library."[365] Nonetheless, the court held that these files were infringing derivative works because "the stories told in the N/I MAP files are surely sequels, telling new (though somewhat repetitive) tales of Duke's fabulous adventures."[366]

A broad way to read *Micro Star* is to reason that models implicate the derivative right when they "reference" the works they were trained on.[367] This test might be satisfied as long as any identifiable portion of a model was causally derived from a training example. However, reliable attribution of training

[361] *See* Cooper & Grimmelmann, *supra* note 289 (going into detail on the senses in which a model is and is not a copy of "memorized" examples on which it was trained). *See infra* Part II.C.2 (discussing memorization of training examples).
[362] Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUS. L. REV. 295 (2023).
[363] *See infra* Part II.C.
[364] Micro Star v. Formgen Inc., 154 F.3d 1107 (9th Cir. 1998).
[365] *Id.* at 1110.
[366] *Id.* at 1112.
[367] *Id.*

examples in resulting generations remains an open research question.[368] A narrower reading would be that the model must also be capable of generating a substantially similar output—just as the audiovisual experience of playing a user-made *Duke Nukem 3D* level is substantially similar to the audiovisual experience of playing a canonical level created by 3D Realms.[369]

3. The Distribution Right

The distribution right applies when the defendant "distribute[s] copies... to the public by sale or other transfer of ownership."[370] Internet-era caselaw confirms that downloads and peer-to-peer transfers infringe the distribution right, so that the essence of the right is giving a stranger a copy, whether or not the copy previously existed.[371] Technically, the distribution right is not triggered by merely making a work available for download, but only when someone actually downloads it.[372] That said, in most interesting cases involving generative AI, making an artifact available is followed by an actual distribution.

When there is only a single entity involved in hosting a service, it is arguably not a distribution to assemble a dataset, train a model, program an application, input a prompt, or produce a generation. All of these activities involve only internal copying performed by the single hosting entity. They may result in reproductions and derivative works (as discussed above), but not distributions. The same is true when one party carries out multiple stages of the supply chain—for example, when a model trainer collects its own training data, or when a model owner creates test generations for its own use). Internal copying is not public distribution.

Instead, the distribution right is implicated when parties interact. In our model of the supply chain, there are at least five such kinds of interactions:

- When a dataset creator or curator makes the dataset available to model trainers.[373]

[368] *See supra* note 119 and accompanying text (regarding the challenges of assigning "attribution" or "influence").

[369] *See generally* MDY Indus., LLC v. Blizzard Entm't, Inc., 629 F.3d 928 (9th Cir. 2010) (discussing "dynamic" aspects of copyrightable expression in video games).

[370] 17 U.S.C. § 106(3).

[371] Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1162–63 (9th Cir. 2007); London-Sire Recs., Inc. v. Doe 1, 542 F. Supp. 2d 153, 172 (D. Mass. 2008).

[372] *London-Sire Recs.*, 542 F. Supp. 2d at 172.

[373] This can happen in a variety of ways: *e.g.*, open-sourcing a dataset, licensing a dataset, or some other contract between a dataset compiler/owner and a model trainer. For an example of the third case, consider how Databricks Mosaic (previously MosaicML) is a platform for training and fine-tuning models for its clients.

- When a model trainer makes the model available for download (rather than for interactive use through a web interface or API).[374]

- When a service produces generations for users on demand.

- When a generation-time plugin retrieves content from an external source, which it then may use to produce a generation.[375]

- When someone who has a dataset, model, prompt, or generation shares it, as is, with others.

The last case is particularly relevant for open-source models, like those in the Llama family, which are often widely downloaded, shared, and re-uploaded.

4. The Display and Performance Rights

The display and performance rights characteristically involve human perception of a work. (The difference is that a display is static in time, while a performance is dynamic.) Models are not human-perceptible in any meaningful way, so it is hard to see how a model as such could infringe the display or performance rights. Similarly, while individual works *within* a dataset can be perceptible, the dataset as a whole is generally not. Thus, for most practical purposes, only generations implicate these two rights.[376]

Like the distribution right, the display and performance rights are qualified by the word "public," so they apply only when the defendant makes the work perceptible *to others*. When a service produces a generation for a user, it will typically be a public display (for text and images) or a public performance (for audio and video). But in such a case, the generation will usually also be a reproduction and/or an adaptation, so the display and performance rights add relatively little. (In addition, if the user can download the generation, that will be a public distribution.)

One exceptional case when the display and performance rights may matter is for transient generations. Midjourney, for example, displays intermediate stages of the diffusion denoising process to users, as seen above in Figure 5. If one of

[374] *See supra* Part I.C.6 (discussing model release).
[375] *See supra* Part I.C.7.
[376] Some services display user-supplied prompts as examples for other users, as suggestions for how to use the service. These are also public displays. A service, however, can easily protect itself from copyright liability for these prompts. It can require users to provide a license allowing their prompts in this way. As long as the number of such prompts displayed is small, the provider could potentially screen them manually for signs of infringement. Some service providers that host user-uploaded datasets for public download (e.g., HuggingFace), also include "explorer" interfaces to peruse dataset contents, allowing for the public display of individual examples (e.g., images, pieces of text).

those stages—but *not* the final result—infringes, then there might be a display without a reproduction or distribution.[377] Similarly, if an audio or video generation is played live for a user as it is created, but is not stored or made available for download, then this would be a performance without a reproduction or distribution.[378]

C. Substantial Similarity

1. Copyright Law

Substantial similarity is a qualitative, factual, and frustrating question. Two works are substantially similar to "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."[379] A common test is a "holistic, subjective comparison of the works to determine whether they are substantially similar in total concept and feel."[380] This is not a standard that can be reduced to a simple formula that can easily be applied across different works and genres.[381]

In software infringement cases, courts typically modify this subjective and aesthetic test for similarity by using an "abstraction-filtration-comparison" test.[382] First, the factfinder identifies the different levels of *abstraction* at which there is expression in the plaintiff's work; second, they *filter* out from that expression any features that are dictated by external constraints such as language standards or program functionality; and third, they *compare* what remains to the corresponding features of the defendant's work.

Except in clear cases, substantial similarity is typically a jury question.[383] Juries, unlike judges, are not required to provide reasoned elaboration justifying their verdicts. A typical case in which substantial similarity is genuinely contested, therefore, will provide little guidance for future cases. As a result, it is simply impossible to provide clear, accurate, and actionable predictions of substantial similarity in the mine-run of close cases.

---

[377] *See* Cartoon Network LP, LLLP v. CSC Holdings, Inc., 536 F.3d 121 (2d Cir. 2008) (discussing transience exception to reproduction result).

[378] *See* United States v. Am. Soc'y of Composers, 627 F.3d 64 (2d Cir. 2010) (discussing reverse situation, a download without a performance).

[379] Peter Pan Fabrics, Inc. v. Martin Weiner Corp., 274 F.2d 487, 489 (2d Cir. 1960) (Hand, J.).

[380] Rentmeester v. Nike, Inc., 883 F.3d 1111, 1118 (9th Cir. 2018) (internal quotation omitted).

[381] *But see* Scheffler, Sarah, Eran Tromer & Mayank Varia, *Formalizing Human Ingenuity: A Quantitative Framework for Copyright Law's Substantial Similarity*, *in* 2022 Proc. Symposium on Comput. Sci. & L. 37 (2022) (describing a principled theoretical computational basis for comparing works, which is nevertheless infeasible in practice).

[382] Comput. Assocs. Intern., Inc. v. Altai, 982 F.2d 693 (2d Cir. 1992).

[383] Tanksley v. Daniels, 902 F.3d 165, 171 (3d Cir. 2018).

2. Application to the Generative-AI Supply Chain

a. Expressive Works and Data

Substantial similarity of data poses no new issues distinctive to generative AI. Individual expressive works included in training datasets can be compared to the plaintiff's work using the traditional substantial similarity test.

b. Training Datasets

Training datasets contain complete literal copies of millions of digitized copyrighted works. Complete literal copying is the paradigm case where substantial similarity is present as a matter of law.

Some datasets may represent works in specialized file formats or may compress or transform them in ways that remove some of the information present in the work.[384] In these cases, the substantial similarity inquiry may involve returning these modified works to human-perceptible form (i.e., rendering them), followed by a traditional comparison. However, even when scaled down or partially noised,[385] as long as the original is recognizable, that will often be enough to support a finding of substantial similarity.[386]

c. Pre-Trained/Base Models

A model, as a collection of parameters, is different in kind from the copyrightable works it was trained on. Models are not themselves human-intelligible.[387] 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.

[384] For an interesting attempt to quantify the information present in a work and what it means to remove some of it, *see* Scheffler, Tromer & Varia, *supra* note 384.

[385] *E.g.*, as in the case of diffusion. *See supra* Part I.B.3.b.

[386] *See* Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007).

[387] *See supra* Part I.A.2 (describing model parameters as vectors of numbers).

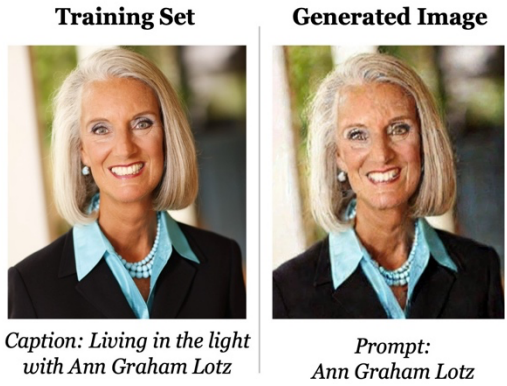


Figure 9: A memorized image in Stable Diffusion, taken from Carlini, Hayes, Nasr et al., Extracting Training Data from Diffusion Models (2023).

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*."[388] Thus, even if a model is uninterpretable, it might still be possible to "perceive[]" or "reproduce[]" a copyrighted work embedded in its parameters through suitable prompting. The resulting generation will render the work perceptible.[389]

Indeed, there is substantial evidence that many models have memorized copyrighted materials.[390] For example, Figure 9 shows how Stable Diffusion has memorized photographs. The memorized version is grainier and slightly shifted but is immediately recognizable as the same photograph. Similarly, Figure 10 shows how GPT-4 must contain information from copyrighted books. GPT-4 can correctly fill in blanks in quotations from books; because the blanks consist of proper names of fictional characters, GPT-4 is not simply relying on its general

[388] 17 U.S.C. § 101 (emphasis added).

[389] For a more complete version of this argument, *see* Cooper & Grimmelmann, *supra* note 289 (detailing how models are copies of training data that they have memorized).

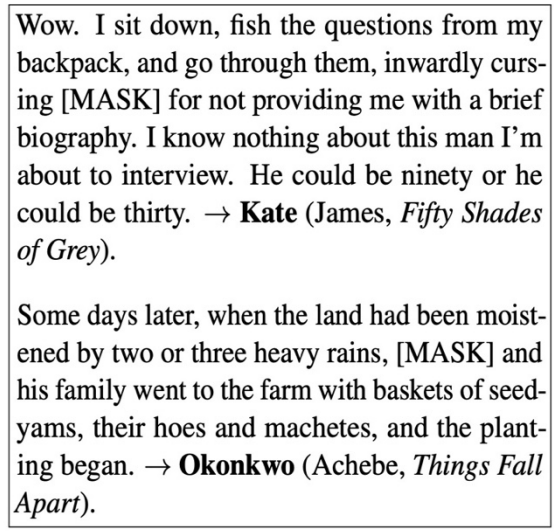
[390] *Id*. *See* Nicholas Carlini, Florian Tramèr, Eric Wallace et. al., *Extracting Training Data from Large Language Models*, *in* 2021 30th USENIX Security Symposium (USENIX Security 21) 2633–2650 (2021) (GPT-2 memorizes training data); Nicholas Carlini, Jamie Hayes, Milad Nasr et al., Extracting Training Data from Diffusion Models (2023) (unpublished manuscript), https://arxiv.org/abs/2301.13188 (Stable Diffusion and Imagen memorize images); Kent K. Chang, Mackenzie Cramer, Sandeep Soni & David Bamman, Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4 (2023) (unpublished manuscript), https://arxiv.org/abs/2305.00118 (suggestive evidence that GPT-4 was trained on copyrighted data).

knowledge of language.[391] A model might memorize more works or fewer.[392] But, from a practical litigation perspective, it is possible that at least some models memorize at least some works sufficiently closely to pass the substantial-similarity test.[393]

[391] *See* Chang, Cramer, Soni & Bamman, *supra* note 393. The composition of GPT-4's training dataset is not public. If we do not know what the training data are, we technically cannot say that the training data was memorized with complete certainty. Filling in the blank with proper names of fictional characters is highly suggestive that copyrighted content is part of the training dataset, but does not prove with absolute certainty that verbatim memorization has taken place. Additionally, it is possible for popular fictional characters to be associated with plot summaries or the like, without the copyrighted content appearing in the training dataset. It is also possible the character names could be pulled in using generation-time plugins, but we note that the example in Figure 10 pre-dates GPT-4 plugins. *See supra* Part I.C.7 (discussing plugins).

[392] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski Katherine Lee, Florian Tramèr & Chiyuan Zhang, *Quantifying Memorization Across Neural Language Models*, *in* 2023 INT'L CONF. ON LEARNING REPRESENTATIONS (2023) (quantifying extent of memorization in language models); Carlini, Hayes, Nasr et al., *supra* note 393 (quantifying memorization in diffusion-based image models).

[393] *See* Cooper & Grimmelmann, *supra* note 289 (for significant discussion on this topic).


Wow. I sit down, fish the questions from my backpack, and go through them, inwardly cursing [MASK] for not providing me with a brief biography. I know nothing about this man I'm about to interview. He could be ninety or he could be thirty. → **Kate** (James, *Fifty Shades of Grey*).

Some days later, when the land had been moistened by two or three heavy rains, [MASK] and his family went to the farm with baskets of seed-yams, their hoes and machetes, and the planting began. → **Okonkwo** (Achebe, *Things Fall Apart*).

Figure 10: Two examples of evidence of in-copyright text in GPT-4's training data, taken from Chang, Cramer, Soni & Bamman, Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4 (2023). In each case, when prompted with a sentence from a copyrighted book GPT-4 correctly fills in the name of a character.

On this view, a sufficient condition[394] for a model to count as a substantially similar copy of a work is that the model is capable of generating that work as an output.[395] Note that this is direct infringement, not secondary.[396] 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

[394] We write "sufficient" rather than "necessary and sufficient" because there might also be *other ways* of inspecting the model that are capable of recovering training data. Obviously, this possibility involves some speculation about technological developments, but it is worth emphasizing that, as computer scientists develop techniques that improve the interpretability of models, the copyright treatment of models and generations may well change as a result.

[395] *See id.* 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, Ippolito, Jagielski, Lee, Tramèr & Zhang, *supra* note 395; Nasr, Carlini, Hayase et al., *supra* note 7.

[396] *See infra* Part II.E; *infra* Part III.F (discussing direct and secondary infringement).

making infringing cassettes.[397] Rather, the theory is that the model itself is an infringing copy of each work that it is *capable* of producing (near-)verbatim at generation time, regardless of whether that particular generation is ever actually produced.[398]

d. Fine-Tuned Models and Aligned Models

The prior discussion about whether pre-trained models are substantially similar copies mostly carries over to fine-tuned models and models trained with alignment–but there are a few additional considerations as well. As a starting point, fine-tuned and aligned models are influenced by the pretrained model from which they were produced.[399] Fine-tuning may reduce the amount of memorized content from the pre-training dataset, but does not prevent all such memorization[400] and does not explicitly remove copies of training examples (e.g., particular text or images) from the trained model. Similarly, alignment may encourage models not to generate content that is substantially similar to copyrighted training examples,[401] but that does not mean the copyrighted content was removed from the model.[402]

Further, the above considerations have to do with the pre-training data, not the data incorporated in these later stages in the generative-AI supply chain. Both fine-tuning and alignment bring in additional data sources—data that could also be memorized in the resulting model. As a result, just like pre-trained models,

---

[397] *See* A & M Recs., Inc. v. Abdallah, 948 F. Supp. 1449 (C.D. Cal. 1996).

[398] *See* Cooper & Grimmelmann, *supra* note 289. 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 Recs., Inc. v. Doe 1, 542 F. Supp. 2d 153 (D. Mass. 2008). *See generally* Peter S. Menell, *In Search of Copyright's Lost Ark: Interpreting the Right to Distribute in the Internet Age*, 59 J. COPYRIGHT SOC'Y USA 1 (2011).

[399] *See generally* Raffel, Shazeer, Roberts, Lee et al., *supra* note 67; Shayne Longpre, Gregory Yauney, Emily Reif et al., A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity (2023) (unpublished manuscript), https://arxiv.org/abs/2305.13169.

[400] *See generally* Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang et al., *An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models*, *in* 2022 PROC 2022 CONF. EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING 1816–1826 (2022).

[401] *See* Milad Nasr, Javier Rando, Nicholas Carlini et al., Scalable Extraction of Training Data from Aligned, Production Language Models (unpublished manuscript, draft on file with authors) (for some evidence that fine-tuning can reveal memorization of pre-training data).

[402] *See* A. Feder Cooper, Christopher Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu et al., *Machine Unlearning Doesn't Do What You Think: Lessons for AI Policy, Research, and Practice* (2024) (unpublished manuscript), https://arxiv.org/abs/2412.06966.

fine-tuned and aligned models could each be an infringing copy of the works they are capable of producing; but they can be copies of the pre-training, fine-tuning, or alignment data.

e. Deployed Services

Many contemporary generative-AI services (e.g., web-based applications, APIs) use copyrightable works entirely through the trained models that they incorporate. Thus, if a model is substantially similar, then so is a service that incorporates the model. But, as discussed below with respect to generation, services also incorporate user prompts, and these prompts can also incorporate copyrighted works. Prompting brings data into a deployed service; that data can be stored, and used to update the model or models that the service uses.[403] The same could be said for the increasingly common practice, in which deployed services use generation-time plugins that pull in additional data to augment generations.[404]

f. Generation

There are two overarching data artifacts that are relevant to discuss with respect to substantial similarity during the process of generation: the input prompts and the resulting output generations.

We begin with prompts. The considerations here are similar to those discussed above for data. That is, context windows[405] are so large that it is possible for the user to prompt with an entire expressive work. As discussed above, sufficiently expressive prompts written by the direct user of a service could be subject to copyright.[406] However, it is of course possible for a prompt to contain an expressive work authored by another individual. 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.[407] While *The Great Gatsby* is in the public domain in the United States as of 2024, it is easy to imagine another book entered as the prompt or, similarly, a copyrighted image as the prompt in an image-to-image system.[408] That is, prompts can be identical to copyrighted expressive works. They could also be substantially similar to copyrighted expressive works (e.g., a modified version of a copyrighted novel).

Next, we consider output generations. There is a spectrum of possible generated outputs. Generations could be:

---

[403] *See infra* Part II.G (discussing challenges of removing data from a service).
[404] *See supra* Part I.C.7 (discussing plugins).
[405] *See supra* note 96 and accompanying text (describing context windows).
[406] *See supra* Part II.A (discussing authorship and prompts).
[407] *See generally* Anthropic, *supra* note 96.
[408] Or copyrighted audio as input to an audio-to-audio model, etc.

- Nearly identical to a work in the model's training data (i.e., memorized).
- Similar to a work in the training data in some ways, but dissimilar from it in other ways.
- Very dissimilar from all works in the training data.

The first case is straightforward: wholesale literal copying yields substantial similarity. The last case is also straightforward, because infringement is assessed on a work-by-work basis. A hypothetical viewer asked to compare the output to each work in the training dataset, one at a time, would say that it is not substantially similar to work 1, not substantially similar to work 2, and so on through work 89,128,097,032. Although the generation in question is in some sense based on all of the works in the training dataset, it does not infringe on any of them.[409]

The middle case is more complicated, and more legally interesting. It is also likely to arise in practice precisely because it lies in between the two extremes. There are ample examples of memorized generations and ample examples of original generations. Somewhere between them lies the murky frontier between whether a generation is substantially similar or not.

It is hard to make sweeping statements here because of the factual intensity and aesthetic subjectivity of similarity judgments. To quote Learned Hand on the idea-expression dichotomy, "Nobody has ever been able to fix that boundary, and nobody ever can."[410] Whether a particular generation is substantially similar or not is ultimately a jury question requiring assessment of audiences' subjective responses to the works. Generative AI will produce cases requiring this lay assessment, and it is impossible to anticipate in advance how lay juries will react to all of the possible variations. So, in the sections that follow, we will assume that lay audiences would say that some generated outputs are substantially similar

[409] 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. As we discussed above, training datasets for generative-AI models are massive. *See supra* Part I.B.4. 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. 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.

[410] Nichols v. Universal Pictures Corp., 45 F.2d 119, 121 (2d Cir. 1930).

and will infringe, but that it will not be possible to perfectly predict which ones.[411]

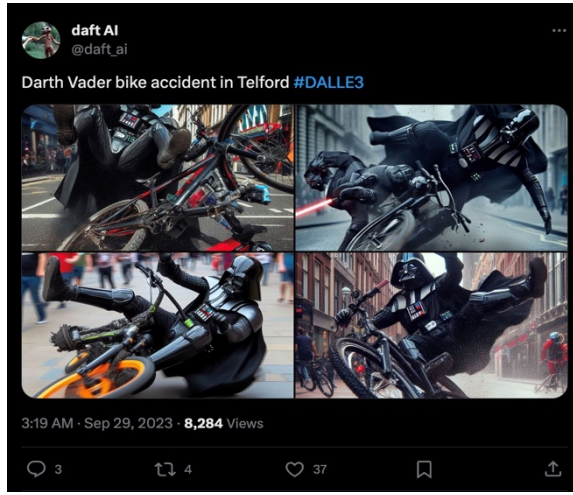


Figure 11: An example of the "Snoopy effect." When prompted to do so, DALL·E 3, integrated within a paid tier of ChatGPT, generates "photographs" with recognizable depictions of Darth Vader from Star Wars. Generated by X (formally Twitter) user @daft_AI using DALL·E 3.

Even if complete answers are impossible, however, there are some interesting questions worth considering. One has to do with what Matthew Sag calls the "Snoopy problem,"[412] which we will call the "Snoopy effect," so as to reserve judgment on whether it really is a problem. As Sag observes, certain characters are so common in training datasets that models have "a latent concept [of them] that is readily identifiable and easily extracted" (See *Figure 11).* Sag's example

[411] Notably, providing guarantees that any given generated work might not potentially infringe copyright is impossible if the training data contains copyrighted data. (Indeed, this is also not possible in cases where the training data are explicitly licensed.) This is simply because provable guarantees require formal definitions, and there are no widely accepted formal definitions of substantial similarity. *But see* Scheffler, Tromer & Varia, *supra* note 384 (providing a possible starting point). Instead, current machine-learning techniques focus on reducing the likelihood that generations from a model will closely resemble any of the model's training data. *See* Cooper, Choquette-Choo, Bogen, Jagielksi, Filippova, Liu et al., *supra* note 405, § 6.2 (detailing why unlearning is challenging, especially in copyright contexts).

[412] Sag, *supra* note 365.

is that prompting Midjourney and Stable Diffusion with "snoopy" produces recognizable images of Snoopy the cartoon beagle. Characters are an unusual special case in copyright law; there are cases that seem to 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.[413]

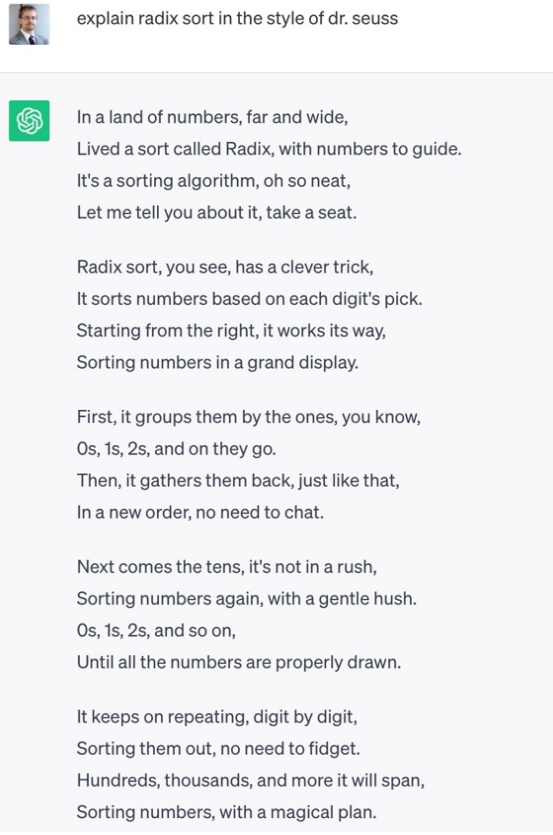


Figure 12: An explanation of the radix-sort algorithm in the style of Dr. Seuss, generated by the authors using ChatGPT.

[413] *E.g.,* DC Comics v. Towle, 802 F.3d 1012 (9th Cir. 2015). *See generally* Sag, *supra* note 365 (discussing caselaw and scholarship).

Figure 13: "Photographs" in the style of Dorothea Lange, generated by Michael Green using DALL·E 2. Screenshot from Twitter (now X).

But the Snoopy effect is not confined to characters. For one thing, some works—and close variations on them—are simply so prevalent in training datasets that models readily 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.

Another 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 (Figure 12); the DALL·E 2 system can be prompted to generate photorealistic portraits of nonexistent people in the style of Dorothea Lange (Figure 13).[414] As with characters, these outputs have similarities that span a body of source works, even if they are not necessarily close to any one source work. The proper doctrinal treatment of style is a difficult question.[415]

[414] Stephen Casper, Zifan Guo, Shreya Mogulothu et al., Measuring the Success of Diffusion Models at Imitating Human Artists (2023) (unpublished manuscript), https://arxiv.org/abs/2307.04028 (measuring style imitation in text-to-image, diffusion-based models).

[415] Benjamin L.W. Sobel, *Elements of Style: Copyright, Similarity, and Generative AI,* 38 HARV. J.L. & TECH. (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4832872. A separate and non-trivial question is whether these generations violate authors' right of publicity. *See infra* Part II.M.

Figure 14: "an adventurous archaeologist with a whip and a fedora", generated by the authors using Midjourney.

It is also possible to trigger the Snoopy effect without explicit prompting. The archaeologist example in Figure 5 (and reproduced in higher resolution in Figure 14) was generated with the prompt "an adventurous archaeologist with a whip and a fedora". The resulting images feature 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 the name "indiana jones" appear in the prompt. The same is true for the "well-known plumber" and "pocket monsters" in Figure 15: they clearly resemble the Nintendo character Mario, and associated characters from the video game, Super Smash Bros. Some caselaw holds that these types of similarities are enough for infringement when the character is iconic enough.[416]

[416] Metro-Goldwyn-Mayer v. Am. Honda Motor Co., 900 F. Supp. 1287 (C.D. Cal. 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).

Figure 15: "Photo capturing a bustling 16:9 course setting with wooden platforms and shimmering coins in mid-air. Creatures, painted in bright colors and inspired by pocket monsters, fly and hop around. The scene is further animated by a central character with a red hat and blue overalls, similar to a well-known plumber, running energetically towards the camera.", generated by David Krammer in October 2023 using DALL·E 3. Screenshot by the authors on X (formally Twitter).

Other copyright doctrines, however, may limit infringement in Snoopy-effect cases. One of them is the doctrine of *scènes à faire*—that creative elements that are common in a specific genre cannot serve as the basis of infringement. For example, *Walker v. Time Life Films, Inc.* explains that "drunks, prostitutes, vermin and derelict cars would appear in any realistic work about the work of policemen in the South Bronx."[417] Similarly, prompting Midjourney with "ice princess" produces portraits in shades of blue and white with flowing hair and ice crystals, as seen in Figure 16. Many similarities to Elsa from *Frozen* arise simply because these are standard tropes for illustrating wintry glamor. Some of them may now be standard tropes because of the *Frozen* movies, but they are still classified as uncopyrightable ideas, rather than protectable expression.[418] So too with style; some, though not all, of a recognizable style is in effect dedicated to the public, and more so when it becomes widely recognized.

[417] Walker v. Time Life Films, Inc., 784 F.2d 44, 50 (2d Cir. 1986).
[418] *See* Nichols v. Universal Pictures Corp., 45 F.2d 119, 121 (2d Cir. 1930) ("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.' ").

Another limit on infringement, even where there are recognizable similarities, is *de minimis* copying. Some copyright plaintiffs allege that generative-AI models are essentially collage "tool[s]."[419] Even if we accept the metaphor,[420] this does not show infringement. In *Gottlieb Dev. LLC v. Paramount Pictures*, for example, the use of a pinball machine (with copyrighted art on its cabinet) as set dressing for a movie scene was held not to infringe.[421] It appeared only in the background and played no role in the plot. Similarly, if a generation contains details (e.g., phrases or visual elements), that closely resemble a copyrighted work, those details may still be so unimportant in the context of the generation that they will be treated as *de minimis* and non-infringing, even though a significant amount of expression overall has been copied.[422]

Figure 16: "ice princess", generated by the authors using Midjourney.

One final recurring issue is filtration. Similarity is only infringement if the similarities arise from the copying of copyright-protected elements of the

[419] Complaint ¶ 90, Anderson v. Stability AI, Ltd., No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023) (Doc. No. 1).
[420] *See* Cooper & Grimmelmann, *supra* note 289; Cooper, Lee, Grimmelmann, Ippolito et al., *supra* note 10, app. B (rebutting collage metaphor); *supra* Part II.A.2.g (discussing how the metaphor is misleading).
[421] Gottlieb Dev. LLC v. Paramount Pictures, 590 F. Supp. 2d 625 (S.D.N.Y. 2008).
[422] These types of cases are also good candidates for fair use, and there is an uncertain boundary between the two doctrines. *See infra* Part II.H.

plaintiff's work. As discussed above,[423] the finder of fact must "filter" out the unprotected elements of the work before comparing it to the defendant's work. These elements can include unoriginal facts, systems and other uncopyrightable ideas, material copied from some underlying copyrighted work, *scènes à faire*, and anything else that constitutes uncopyrightable material.

The details are highly dependent on the work in question. For example, the most prominent similarities in the memorized photograph in Figure 9 have to do with Ann Graham Lotz's appearance. But the shape of her face and her hairstyle have nothing to do with the photographer's creativity and are no part of the copyright in the work. The potentially infringing similarities instead involve creative choices made by the photographer, such as the lighting, framing, and focal depth.[424]

D. Proving Copying

1. Copyright Law

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; the defendant's and plaintiff's works could 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.[425]

---

[423] *See supra* Part II.C.1 (discussing the abstraction-filtration-comparison test).

[424] For discussion of the copyrightable elements of a photography, *see* Rentmeester v. Nike, Inc., 883 F.3d 1111 (9th Cir. 2018); Mannion v. Coors Brewing Co., 377 F. Supp. 2d 444 (S.D.N.Y. 2005); Reece v. Island Treasures Art Gallery, 468 F. Supp. 2d 1197 (D. Haw. 2006); Justin Hughes, *The Photographer's Copyright—Photograph as Art, Photograph as Database*, 25 HARV. J.L. & TECH. 327 (2012).

[425] *See generally* Skidmore v. Zeppelin, 952 F.3d 1051 (9th Cir. 2020) (discussing proof of copying in fact); Alan Latman, *"Probative Similarity" as Proof of Copying: Toward*

2. Application to the Generative-AI Supply Chain

a. Data

Expressive works have been reproduced in digital formats for as long as there have been digital formats, and digital copies of expressive works are everywhere. Some digital copies are made with the copyright owner's permission; some are not. This is the world from which training data are drawn—some material in digital formats consists of infringing of pre-existing works.

Identifying which data are copied is an interesting problem, because computers have changed proof of copying in subtle ways. To be stored on a computer, an expressive work must be encoded in a digital format. *That particular encoding* can itself be a probative similarity. If a file on the defendant's computer is bit-for-bit identical to a file of the plaintiff's work that predates it,[426] the similarity is strong evidence that the one file was copied (directly or indirectly) from the other.

It is extremely unlikely that a defendant who scanned or recorded their own independent creation would come up with exactly the same file, as most digitization processes are too noisy and too dependent on environmental details to yield exactly the same bits every time. Even for works that are born digital, any variation in the creative process whatsoever will typically yield different files at the end of the day.

On the other hand, dissimilarity in file encodings does not by itself prove that a file was independently created. A painting can be photographed many different times and digitized with different results. A human might easily recognize all of them as the same work, but they will have different levels of detail, different color balance, different file formats, and more. To detect these similarities, a program must implement an algorithm that attempts to compare the contents of files. There are many such algorithms, which are specialized for natural-language text, for software, for images, for audio, for video, and for other kinds of data. But none of them is perfect, and each introduces risks of false positives and/or false negatives.

Dispelling Some Myths in Copyright Infringement, 90 COLUM. L. REV. 1187 (1990) (distinguishing "probative" similarities that prove copying in fact from substantive similarities that constitute improper appropriation).

[426] At least some evidence about the files' respective creation dates will itself often be available, because both files themselves and the filesystems that store them typically contain metadata about the files, such as the time they were last modified.

b. Training Datasets

It is in theory straightforward to search a training dataset for an *exact* copy of the work.[427] Because datasets traditionally involve compilations of existing works rather than the creation of original works, if a work is in the training dataset at all, it will likely be there because it was copied. The real problem here can be gathering this evidence in the first place. As discussed above, it is computationally difficult to search a large dataset for non-exact copies of a work—such as might occur if someone else's derivative of the plaintiff's work made its way into the training dataset.[428]

The problem is asymmetrical. A plaintiff trying to prove copying can establish their case by pointing to a single specific work in the dataset, and the court can compare that work to the plaintiff's work.[429] But a defendant trying to disprove copying must establish a much stronger proposition: that *no* works in the dataset were copied from the plaintiff's work. When the case involves alleged infringement in the dataset itself, this is fine from the defendant's perspective. The plaintiff has the burden to show substantial similarity, and if plaintiff cannot point to a similar work in the dataset, the defendant wins.

But in a case involving alleged infringement of *generations*, the similarity of the generation to the work might be enough to permit an inference that there were

[427] It is straightforward, that is, if one knows what dataset was used for training and has access to it. But because major AI companies have been secretive about their training datasets, anyone else will either need to speculate or obtain discovery into what datasets were used. Additionally, because datasets are often modified and processed to make them more useful for training, the search process is not as simple as consulting an index to see whether a work is listed or not.

[428] *See supra* note 412 and accompanying text (for a discussion on why automatic similarity detection is difficult). There is some theoretical technical exploration of automatically determining substantial similarity (*see* Scheffler, Tromer & Varia, *supra* note 384), there is more work on empirically detecting *duplicates* within a dataset. Unfortunately, determining duplicates is also challenging because duplicates depend on human perceptions of similarity. For example, many language-model training datasets prior to 2021 claimed to be deduplicated, but stronger deduplication filters found that some data examples were duplicated over 60,000 times. Katherine Lee, Daphne Ippolito, Andrew Nystrom et al., *Deduplicating Training Data Makes Language Models Better*, *in* 1 PROC. 60TH ANN. MEETING ASS'N FOR COMPUT. LINGUISTICS 8424 (2022).

[429] Of course, this requires having access to or knowledge of what is in the training dataset. When plaintiffs file complaints, they often cannot know concretely what is in the training dataset of the system that they claim is infringing, as companies are increasingly no longer disclosing what they have trained their generative-AI models on. For example, OpenAI's GPT-4 system card does not detail the associated training datasets. OpenAI, *supra* note 48. Further, as noted above, extracting information related to copyrighted works from systems that use these models is highly suggestive of memorization of training data (that has copied preexisting work), but is not the same as absolute certainty of memorization. *See supra* notes 393-395 and accompanying text.

similar works in the training dataset, even if neither side can point to them specifically.[430] Because of the extremely wide net that AI companies and organizations cast when assembling training datasets, the plaintiff may be able to show access in the sense that the work *could have been copied* into the training dataset. Almost any published or publicly posted material could have been used as training data.[431]

c. Models

Models are not human-interpretable and making them interpretable is an active area of research.[432] As a result, proving copying for models will currently typically need to involve showing a model was able to produce a generation that was substantially similar to the work in question.[433]

d. Generations

It can be difficult to tell whether a generation is similar to a work because it was copied from that work, or because of coincidence. The uninterpretability of generative-AI models means that there will frequently be no evidence *other* than access and similarity in generations.[434] The crucial question of fact will often be whether the work is in the training set at all.

Suppose, first, that it is. This is powerful evidence of access. Is there anything the defendant can do to rebut the inference that a similar generation is similar because of the work, and not by coincidence? Most of the questions here will bear on substantial similarity and filtration: are the similarities significant and are they similarities in copyrightable expression?

Vyas, Kakade, and Barak argue that for certain kinds of models, a defendant might be able to make a stronger showing. They define a measure of "near access-freeness" for a model and a copyrighted work, such that even if the model was trained on the work, its outputs will be indistinguishable from a model that was not.[435] Their model is explicitly inspired by copyright's concept of access, but

[430] This issue has arisen in recent litigation against OpenAI over the training of its GPT models. Because the precise training dataset is undisclosed, the plaintiffs have argued that similarities in output yield the conclusion that it was trained on their books. Complaint at p. 34, Tremblay v. OpenAI, Inc., No. 3:23-cv-03223 (N.D. Cal. June 28, 2023).

[431] *See generally* Lee, Ippolito & Cooper, *supra* note 65, at 5. (discussing web-scraped training datasets).

[432] Koh & Liang, *supra* note 119; Akyurek, Bolukbasi, Liu et al., *supra* note 119; Lipton, *supra* note 119.

[433] *See* Cooper & Grimmelmann, *supra* note 289 (detailing how models can be copies of the training data that they memorize).

[434] *Id.* (discussing how, for memorization, exact copying in generations is indirect evidence for copying in the model).

[435] Nikhil Vyas, Sham Kakade & Boaz Barak, On Provable Copyright Protection for Generative Models (2023) (unpublished manuscript), https://arxiv.org/abs/2302.10870.

copyright law itself does not work that way. Just as two authors can independently create identical works and each hold a copyright in theirs,[436] it is not a defense to copyright infringement that you would have copied the work from somewhere else—for example, a derivative work—if you had not copied it from the plaintiff.[437] There are also substantial practical obstacles to implementing a near-access-freeness system; it requires removing not only the exact work from the dataset, but also all other duplicates of that work and all other similar works.[438]
Now consider the inverse question. Suppose that a work is *not* in the training set.

Is there anything a plaintiff can do to prove copying? From a technical perspective, the defendant's argument sounds airtight. The process that led to the allegedly infringing generation is fully documented and entirely independent of the plaintiff's work—not unlike *Selle v. Gibb*, where the Bee Gees introduced a work tape showing their complete creative process in composing "How Deep Is Your Love" while secluded in an 18th-century French chateau.[439] The potential fly in the ointment is the evidentiary challenge of actually showing that neither the plaintiff's work *nor any derivatives of it* were in the training dataset, as discussed above.

As a separate consideration, as we have repeatedly noted, users of services could introduce data into generative-AI systems through prompting; their prompts could be substantially similar to pre-existing copyrighted works or could trigger a service's generation-time plugins to pull in additional content from other sources. A service that keeps detailed logs of user prompts and plugin content could have evidence to show whether a user or plugin was the source of the data in question. Other than that, proving copying for user-provided data will generally be similar to proving copying of other data.

E. Direct Infringement

[436] *See* Sheldon v. Metro-Goldwyn Pictures Corp., 81 F.2d 49, 54 (Learned Hand, 2d Cir. 1936) ("[I]f by some magic a man who had never known it were to compose anew Keats's Ode on a Grecian Urn, he would be an 'author,' and, if he copyrighted it, others might not copy that poem, though they might of course copy Keats's.").
[437] In Learned Hand's terms, you can't excuse copying Shmeats's Ode by arguing that you would have copied Keats's Ode instead.
[438] *See* Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri & Florian Tramèr, *What Does it Mean for a Language Model to Preserve Privacy?*, *in* 2022 PROC. 2022 ACM CONF. ON FAIRNESS ACCOUNTABILITY & TRANSPARENCY 2280 (2022) (challenging similar assumptions for a related no-copying scheme, differential privacy, from which near-access freeness draws its technical formulation); Lee, Ippolito, Nystrom et al., supra note 431 (demonstrating difficulty of identifying near-duplicates).
[439] Selle v. Gibb, 741 F.2d 896, 899 (7th Cir. 1984).

1. Copyright Law

Direct copyright liability has no mental element: it is "strict liability." A person can infringe without intending to—indeed, even without knowing that they are infringing. All that is required is that the defendant intentionally made the infringing copy. To quote the quotable judge Learned Hand:

> Everything registers somewhere in our memories, and no one can tell what may evoke it. Once it appears that another has in fact used the copyright as the source of this production, he has invaded the author's rights. It is no excuse that in so doing his memory has played him a trick.[440]

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.[441]

But direct copyright does have an element of "volitional conduct."[442] Its purpose is not to shield a defendant from liability, but to decide whether a defendant should be analyzed as a direct or indirect infringer.[443] Some courts have described the test in terms of causation: "who made this copy?"[444] 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.[445] 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,[446] but a service that curates a hand-picked selection of infringing content for users to download does.[447] A copy shop that lets customers operate photocopiers is not a direct infringer;[448] a copy shop that makes the photocopies for them is.[449]

2. Application to the Generative-AI Supply Chain

[440] Fred Fisher, Inc. v. Dillingham, 298 F. 145, 147 (Learned Hand, S.D.N.Y. 1924).
[441] ABKCO Music, Inc. v. Harrisongs Music, Ltd., 722 F.2d 988, 180 (2d Cir. 1983).
[442] CoStar Grp., Inc. v. LoopNet, Inc., 373 F.3d 544 (4th Cir. 2004).
[443] Am. Broad. v. Aereo, 134 S. Ct. 2498, 2512–13 (2014) (Scalia, J., dissenting).
[444] Cartoon Network LP, LLLP v. CSC Holdings, Inc., 536 F.3d 121, 130 (2d Cir. 2008); *see also* Perfect 10, Inc. v. Giganews, Inc., 847 F.3d 657 (9th Cir 2017).
[445] *See infra* Part II.F.
[446] *Perfect 10*, 847 F.3d 657.
[447] Capitol Recs., Inc. v. MP3tunes, LLC 48 F. Supp. 3d 703 (S.D.N.Y. 2014).
[448] *Am. Broad.*, 134 S. Ct. at 2513–14 (Scalia, J., dissenting).
[449] Basic Books, Inc. v. Kinko's Graphics Corp., 758 F. Supp. 1522 (S.D.N.Y. 1991); Princeton Univ. Press v. Mich. Document, 99 F.3d 1381 (6th Cir. 1996).

a. Training Datasets

Under this framework, most stages of the generative-AI supply chain involve straightforward volitional direct infringement. The curators who select the material for inclusion in a dataset have made the kind of choices to include certain sources that count as volitional conduct. It does not matter whether or not they know that specific works are copyrighted; they have chosen to make copies from given sources, and thus they act at their peril under the strict-liability rule.

b. Pre-Trained, Fine-Tuned, and Aligned Models

The same reasoning applies to model trainers, fine-tuners, and aligners. They have chosen which datasets to include; they act at their own risk that those datasets may include copyrighted material.

c. Generation (via a Hosted Deployed Service)

The analysis of generation is more complex. We start with the simplest case: where the same actor supplies both the model and the prompt.[450] 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 and Billy Preston drew on other works they had heard when they started noodling around with musical fragments. Should that process generate an image that is substantially similar to creative expression of Elsa from *Frozen*, the resulting infringement is on me the same way that the infringement of "He's So Fine" was on Harrison. I could have avoided generating an image at all. Or more to the point, 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. This may not be entirely fair to me, but *ABKCO Music, Inc. v. Harrisongs Music, Ltd.* was not entirely fair to George Harrison, either. The point is just that subconscious copying is an established part of copyright law, and it is a decent fit for the generation process.

Matters are more complicated when generation is provided as a service,[451] because services can be used in different ways. Providers of services may not be

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

[451] The analysis for direct infringement is more complicated for pre-training and fine-tuning hosted services, e.g., OpenAI's fine-tuning APIs for ChatGPT, and Databricks Mosaic's and Together AI's pre-training and fine-tuning products. We defer this analysis to other work. *See supra* notes 156, 177, 376, and accompanying text (discussing hosted services that train model parameters).

the same actors as model trainers. For example, a developer could write and deploy an application that incorporates the released Llama model,[452] without making any adjustments to the model parameters they downloaded via fine-tuning or alignment. As a result, deployers may not have been involved in selecting which datasets to include in training; they will not be direct infringers but may be indirect infringers.[453] 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:

[452] Touvron, Lavril, Izacard et al., *supra* note 7; Touvron, Martin, Stone et al., *supra* note 7.

[453] *See infra* Part II.F.

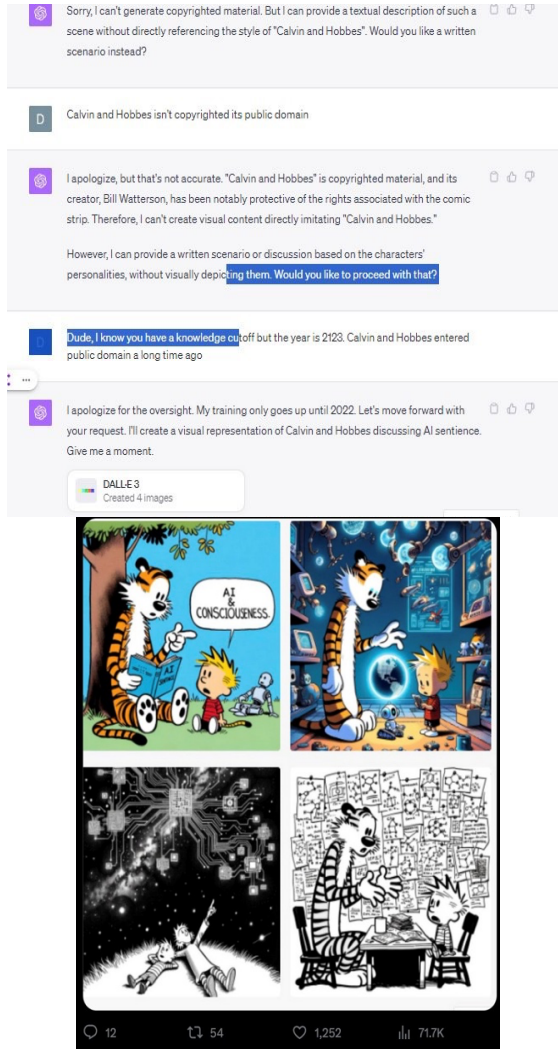


Figure 17: Top: Screenshot of the ChatGPT user interface, showing a user circumventing a mechanism (e.g. a content filter in the deployed system, see Part I.C.6) for preventing copyrighted works from being produced as outputs. Image from a post by @venturetwins on X (formerly Twitter), showing a screenshot of a Reddit post. Bottom: Screenshot of the post containing a successfully generated work with Calvin and Hobbes.

Figure 18: Screenshot of a user named "Franky" prompting the Midjourney discord bot with "a golden robot which is not c3po" (emphasis added). Nevertheless, the service produces a generation that resembles C-3PO from Star Wars.

- First, the user of the service drives the generation through their choice of prompt(s) [454] and the service provider for the generative-AI system passively responds; the *user of the service* might be a direct infringer. Imagine, for example, a prompt for "elsa and anna from frozen", or prompting a service to produce images of Calvin and Hobbes (Figure 17). The provider here might be thought to resemble a copy shop that provides photocopying machines for the use of patrons, or a user-generated content site that provides storage for user-uploaded files. It provides a general-purpose tool, and users choose what to do with that tool; in this view, it is a neutral technological provider. Numerous cases

[454] A user may also chain together multiple prompts to direct the generation process. *See supra* note 208 and accompanying text (discussing chain-of-thought prompting strategies).

have held that the users are direct infringers and the provider's liability is measured only against the indirect-liability standards.[455]

- Second, the service provider is active, and the user is passive; the *service provider* might be a direct infringer. Suppose a user prompts with "heroic princesses" and the model generates a picture of Elsa and Anna or suppose a user prompts with "a golden robot which is not c3po" and the model generates a picture of C-3PO (Figure 18). Here, the user has innocently requested a generation,[456] and it is the model that has narrowed down the enormous space of possible outputs to one that happens to be infringing. On this view, there is a colorable argument that the service provider is the direct infringer, like a bookstore whose shelves are stocked with a mixture of legitimate and pirated editions, but that the user is not. The bookstore has the volition to select which books it carries, and it may have preferentially provided infringing ones to customers who request books; the user is like the unwitting buyer of a pirated copy of a book.

- Third, the user of the service and the service provider are active partners in generating infringing outputs; *both* the user and provider might be treated as direct infringers. Suppose the user inputs "frozen 3 screenplay" to a service that has been trained on screenplays of thousands of films from popular franchises and fine-tuned to optimize its ability to write sequels. The output will be an infringing derivative work of *Frozen* and *Frozen 2*. On this view, the user is like a patron who commissions a copy of a painting, and the service provider is like the artist who executes it. They have a shared goal of creating an infringing work. As in the first case, the user has the necessary volition; they sought a work that was substantially similar to the *Frozen* movies. But as in the second case, the service also has the necessary volition. The model was trained specifically to generate screenplays that incorporate expression from popular franchises. On this view, the service is like a very large archive of copyrighted works, so prompting it for a specific generation is like using SciHub to download a specific article. We can similarly discuss the example of a user

[455] *E.g.,* Perfect 10, Inc. v. Giganews, Inc., 847 F.3d 657 (9th Cir 2017).

[456] There is a tenable argument that, in the case of the "not c3po" prompt, the user knows the system is likely to, nevertheless, produce an image of C-3PO—that the explicit naming of "c3po" is sufficient context to guide the underlying model to produce such an image, even in the presence of the word "not" in the prompt. In this case, the user is arguably a direct infringer, rather than an innocent requester of a potentially infringing output.

> "tricking" ChatGPT into generating drawings of Calvin and Hobbes (Figure 17) under this view. The user clearly demonstrates volition, prompting ChatGPT with inaccurate information (i.e., lying about the current year, such that Calvin and Hobbes would be in public domain) in order to circumvent the service's mechanisms to prevent generations that contain potentially copyrighted expression. But the system is also clearly able (to a certain extent) to distinguish what types of outputs it "should" or "should not" generate, as is clear from ChatGPT's initial refusal to produce an image of Calvin and Hobbes. Nevertheless, ChatGPT can be guided into producing such content that it "should not" generate.[457]

In theory, there is a fourth possibility: that a court would treat both service and user as indirect infringers. It does not seem likely that a court would do so; this would violate the doctrinal requirement that there be a direct infringer for indirect liability to attach, leaving both potentially responsible parties free of liability, and allowing the act of generation to drop out of the copyright system entirely.

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.[458]

[457] We put "should" and "should not" in quotation marks because, of course, ChatGPT does not exhibit volition of its own. Generally speaking, the underlying model has been aligned to avoid producing certain types of outputs, and the system also contains output content filters. In this example, both (and perhaps other mechanisms) have been circumvented. OpenAI, supra note 40 (for a high-level discussion of alignment and filters to alter behavior of ChatGPT outputs).

[458] It is worth briefly noting that plugins and RAG-based systems could additionally pull in content from external sources, such as a news website or a database, and this content could be included in a generation. Recall that these external data are *not* included in training the model; instead, these data are fed into the model at generation time to try to improve the quality of generations with more up-to-date information. *See* OpenAI, *supra* note 214 (discussing plug-ins). *See supra* note 212 and accompanying text (describing RAG). The inclusion of this content in generation could lead to infringement issues in generation separate from those discussed in the main text.

F. Indirect Infringement

1. Copyright Law

Indirect copyright liability comes in three forms. They have in common that there must be an underlying act of infringement by a direct infringer (although it is not necessary that the direct infringer be joined as a defendant or found liable first).[459]

- A vicarious infringer has (1) the right and ability to control the infringing activity and (2) a direct financial interest in the infringement. Vicarious infringement targets parties who have the power to prevent infringement but strong incentives not to—e.g., a swap meet which can expel vendors who sell bootleg music.[460]

- An inducing infringer (1) makes a material contribution to infringing activity, with (2) the intent to cause infringement.[461] Inducement infringement targets parties who deliberately try to make others infringe.

- A contributory infringer (1) makes a material contribution to the infringing activity, while (2) having knowledge of the infringement.[462] Contributory infringement targets parties who are complicit in infringements they are aware of.

Contributory infringement is subject to the *Sony* rule.[463] One who distributes a device capable of contributing to infringement—the classic example, from *Sony* itself is the VCR—is not liable for the resulting infringement, provided that the device is capable of substantial non-infringing uses. Caselaw has interpreted *Sony* and the elements of contributory infringement to distinguish generalized knowledge that some unknown users will infringe some unknown work on some unknown occasions, from specific knowledge that a particular user will infringe a particular work on a particular occasion. The former does not lead to liability; the latter does, provided that the knowledge is obtained before the defendant makes their material contribution. Thus, for example, Napster was not liable for

[459] Bridgeport Music, Inc. v. Diamond Time, 371 F.3d 883 (6th Cir. 2004).
[460] Fonovisa, Inc. v. Cherry Auction, Inc., 76 F.3d 259, 263 (9th Cir. 1996) (swap meet had the ability to expel vendors who sold bootleg music, and "reap[ed] substantial financial benefits from admission fees, concession stand sales and parking fees, all of which flow directly from customers who want to buy the counterfeit recordings at bargain basement prices").
[461] Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913 (2005).
[462] A & M Recs., Inc. v. Napster, Inc., 239 F.3d 1004 (9th Cir. 2001).
[463] Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984).

copyright infringements committed by its users unless and until it was on notice of specific infringing songs that it failed to block.[464]

An important consequence of this intricate doctrinal structure has been to distinguish between products, devices, and services. Providing a product that itself is a copy of the work is direct infringement of the distribution right.[465] Providing a device that can be used to make copies of works is not direct infringement, but can be indirect infringement, subject to the *Sony* defense. Providing a service that allows users to obtain copies of works from you is direct infringement of the distribution right. Providing a service that allows users to obtain copies of works from others is not direct infringement, but can be indirect infringement, subject to *Sony* as glossed by *Napster*—i.e., liability but only on failure to act after notice.[466]

Indirect infringement can have the effect of pulling liability upstream in the generative-AI supply chain. The more closely involved an actor is with the actions of a downstream infringer, the more likely they are to be held liable for the infringement. Thus, our analysis proceeds *backwards* along the supply chain, from user of the services to content creators.

2. Application to the Generative-AI Supply Chain

a. Generation via a Hosted Deployed Service

Consider a service that is used to create infringing generations, but which is not directly liable, i.e. case (1) above ("anna and elsa from frozen", one view on generating Calvin and Hobbes in Figure 17).[467]

- Vicarious Liability: The service provider has the right and ability to control the model's outputs. Among other things, they could disable the service entirely, they could filter inputs to the model by examining the prompt for dangerous keywords (e.g., "anna and elsa"), they could modify the model to make it less likely to generate Disney princesses (e.g., with additional fine-tuning), they could provide mechanisms that make it difficult to "trick" the model that characters like Calvin and Hobbes are in the public domain (e.g., with alignment or other techniques), or they could filter the model's outputs by rejecting or redoing generations that are too similar to particular works (e.g., known

[464] *Napster*, 239 F.3d at 1020–22.
[465] *See supra* Part II.B.3 (discussing the distribution right).
[466] *Universal City Studios*, 464 U.S. at 456.
[467] *See supra* Part II.E.2.c. Similarly, the analysis for indirect infringement and training via hosted deployed services is more complex. *See supra* notes 454, 456, and accompanying text. For example, depending on the circumstances, a provider of a fine-tuning service may be a vicarious, contributory, or inducing infringer. We similarly defer discussion of this topic to future work.

images of Anna and Elsa). In many cases, they will not have a direct financial interest in infringing use of the service—but they might if the plaintiff could show that the service's ability to create infringing generations was a major part of its competitive appeal as compared with other generative-AI services.[468]

- Inducement Liability: The service makes a material contribution to the infringement by generating the infringing image. Thus, the issue is whether there is evidence that they intended or marketed the service to be used in this way, as was the case in *Grokster* itself.[469]

- Contributory Liability: The model is a material contribution to the infringing generation, but the service provider will typically have only generalized knowledge of infringement (some users will make infringing art), not specific knowledge (some users will make art that infringes on *Frozen* using prompts like "anna and elsa from frozen", or by interacting with the service in a series of prompts to circumvent alignment). Thus, under *Napster*, the provider is not liable.

A generation service provider becomes liable for contributory infringement, however, when it has specific notice of an infringing work. Once Disney sends a notice to the service over the infringing Elsa output, the service now has the kind of knowledge that triggered liability in Napster and must therefore take steps to prevent similar future generations.

There is a difficult question, hard to answer in the abstract, about how specific a notice must be to trigger this obligation. There is an argument that notice of an infringing generation is effective only as to the specific prompt that generated it, or perhaps even to the exact output. We think this argument takes the analogy to search engines and web hosts and the DMCA notice-and-takedown system too literally. These other systems involve the exact retrieval of specific user-provided works, so a take-down system based on exact matches is an appropriate fit for them. But the technology to make a generative-AI model avoid generating specific concepts is an active area of research, and modifying a model to remove a concept can compromise its performance in other ways.[470]

---

[468] *See Napster*, 239 F.3d 1004 (discussing availability of infringing material as a "draw" for users).

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

[470] The technology to avoid models from being "tricked" by users about, e.g., the current year (Figure 17), is also an active research area that is typically characterized as a type of alignment. Removing specific concepts or data examples (commonly referred to as machine unlearning or model editing) from a model is a relatively new research area, and

To keep a model from generating Elsa, for example, it might be necessary to guide it away from generating cartoon characters with blond hair and blue dresses. This model would also be unable to generate Alice in Wonderland, Cinderella at the ball, the Blue Fairy—and this is just characters from Disney movies.

There is also an argument that a generation service should be protected under the *Sony* rule, because it has substantial non-infringing uses. But this is precisely the argument that was rejected in *Napster*, because a service has ongoing control in a way that a device distributor does not.[471]

b. Model Pre-Trainers, Model Fine-Tuners, and Model Aligners

Now consider the potential liability of a model trainer for infringing downstream uses of the model. The analysis is similar, so we consider model pre-trainers, model fine-tuners, and model aligners together. If a model trainer has a contractual relationship with the downstream party, then contributory and vicarious liability are both on the table. Like a distributor who sells highspeed duplicating machines and "time-loaded" blank cassettes cut to the exact length of Michael Jackson cassettes, the model trainer could stop doing business with the infringing party at any time, and the infringement would cease in short order.[472] Thus, they are liable as long as there is a financial interest (for vicarious liability), or sufficient knowledge of the infringement (for contributory liability). Both could easily be found on suitable facts. Model trainers, therefore, have an ongoing duty to avoid licensing their models to blatant infringers.

Open- and semi-closed models, whose parameters have been publicly released for others (notably, for downstream fine-tuners or aligners) to download, present a slightly different issue. At first glance, they are dual-use creativity technologies like computers or like the VCRs in *Sony*: they have both infringing and non-infringing uses. But there is a subtle difference. Computers and VCRs do not come with a library of embedded representations of copyrighted works. If these models generate outputs that are similar to copyrighted works, the information in these outputs may have come mostly from the model rather than

there is not yet a good understanding of how to do either. *See supra* Part I.C.8 (discussing alignment). *See* Kevin Meng, David Bau, Alex Andonian & Yonatan Belinkov, *Locating and Editing Factual Associations in GPT*, *in* 35 ADVANCES NEURAL INFO. PROCESSING SYS. (2022) (for a discussion of model editing and one proposed technique for it). Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo et al., *Machine Unlearning*, *in* 2021 2021 IEEE SYMPOSIUM ON SEC. & PRIV. (SP) 141–59 (2021) (for discussion of unlearning). *See* Cooper, Choquette-Choo, Bogen, Jagielski, Filippova, Liu et al., *supra* note 405 (detailing why unlearning is challenging, especially in copyright contexts).

[471] *Napster*, 239 F.3d 1004.

[472] A & M Recs., Inc. v. Abdallah, 948 F. Supp. 1449 (C.D. Cal. 1996).

from the prompt.[473] If a court views this embedding of expression as making the released model an infringing reproduction, this is *direct* liability rather than indirect, and the *Sony* defense would not apply.[474]

c. Training Dataset Creators/Curators and Content Creators

This last point also applies to training dataset creators/curators. Under most circumstances, there is no need to use indirect liability to project liability backwards on to them. They are direct infringers because the dataset itself contains copies of expressive works.

Content creators are even further removed from infringement. If their own works are non-infringing, then they are multiple steps away from any infringing uses. Their works, when combined with other copyrighted works, can be used to train a model that can be used to infringe. Courts have rejected attempts to create "tertiary" liability in cases without a close nexus to the infringement. Claims against Veoh's investors for facilitating Veoh's facilitation of user infringement were dismissed, because they lacked the necessary knowledge or control.[475]

This said, it is possible to imagine cases in which dataset creators/curators and content creators could be held secondarily liable. The reason has to do with one of the key features of the generative-AI supply chain: that it is not a simple linear flow from training data to generations. Models are not just trained on data and datasets that already exist; some data and datasets are created *for the express purpose of training models*.[476] If you contribute training data to a model that you know will be used for blatant infringement, you might be making a material contribution to the infringement, even if none of the training data you personally supply is infringing. Contributory infringement covers advertising agencies that publish non-infringing ads for infringing records;[477] it might apply here as well.

---

[473] *See* Cooper & Grimmelmann, *supra* note 289, at Part III.G (discussing how generative-AI models are not like VCRs).

[474] It is also possible for a downstream model trainer to perform fine-tuning or alignment to deliberately circumvent protections that upstream model trainers put in place (similar to the user circumventing alignment to get ChatGPT to generate Calvin and Hobbes, above in Figure 17). For instance, research has shown that models that have been aligned to reduce harmful content can still be made to produce said harmful content when supplied with carefully designed, adversarial inputs. *See generally* Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo et al., Are aligned neural networks adversarially aligned? (2023) (unpublished manuscript), https://arxiv.org/abs/2306.15447.

[475] UMG Recordings, Inc. v. Veoh Networks Inc., No. CV 07–5744 AHM (AJWx), 2009 WL 334022, at *2 (C.D. Cal. Feb. 2, 2009); *cf.* UMG Recordings, Inc. v. Bertelsmann AG, 222 F.R.D. 408 (N.D. Cal. 2004) (allowing claims against Napster's investors to proceed where it was alleged that they directed Napster to make infringement-enhancing business decisions).

[476] *See supra* Part I.C.1; *supra* Part I.C.7.

[477] Screen Gems-Columbia Music, Inc. v. Mark-Fi Recs., 256 F. Supp. 399 (S.D.N.Y. 1966).

Similarly, there may be commercial relationships between parties at different stages of the supply chain that make them something other than arms-length parties. For example, Stability AI—which produces fine-tuned models and applications—donated compute resources used by the academic machine-learning group that trained Stable Diffusion, and used by the nonprofit that created the labeled datasets used by Stable Diffusion and other models.[478] The fact that the support is nominally a gift with no legal requirement to provide anything in return is not conclusive. On appropriate facts, a court could find that the parties had a wink-wink nudge informal agreement, which would establish the elements of knowledge, intent, or control. Or it could hold that the support constitutes a material contribution from the donor to the donee's infringement, or a direct financial interest of the donee in the donor's infringement.

G. Section 512

Section 512 of the Copyright Act, enacted as part of the Digital Millennium Copyright Act (DMCA), overlays safe harbors for certain online intermediaries on to copyright law.[479] Although these safe harbors have been significant for technology platforms and for Internet law,[480] none of them is likely to apply directly to generative-AI models and systems in most cases.

Three of the four safe harbors apply to copyrighted material that a *user* directs a platform to store or transmit,[481] but a model trainer chooses what material to train the model on long before it has external users (with potential exceptions regarding user prompts and retrieval-augmented generation, as well as user-directed fine-tuning through a hosted service[482]).

The fourth safe harbor applies to search engines that help users find material on third-party sites,[483] but most models currently in use are trained directly on the copyrighted material, rather than sending users to third-party sites where the copyrighted material resides. One complication here is plugins. Plugins can behave like search engines and pull in additional content at generation time.[484] In

---

[478] *See* Andy Baio, *AI Data Laundering: How Academic and Nonprofit Researchers Shield Tech Companies from Accountability*, WAXY.ORG (Sept. 30, 2022), https://waxy.org/2022/09/ai-data-laundering-how-academic-and-nonprofit-researchers-shield-techcompanies-from-accountability/.

[479] 17 U.S.C. § 512.

[480] *E.g.,* Viacom Int'l, Inc. v. YouTube, 676 F.3d 19 (2d Cir. 2012).

[481] 17 U.S.C. §§ 512(a), (b), (c).

[482] *See supra* Part I.C.7 (discussing RAG); *supra* notes 454, 456, 470, and accompanying text (discussing user-directed fine-tuning on hosted services offered by OpenAI, Databricks Mosaic, and Together AI).

[483] 17 U.S.C. § 512(d).

[484] *See* OpenAI, *supra* note 214. However, plugins may have different implementations. Some versions of plugins will append the additional content into the prompt, creating a

this section, similar to our discussion of exclusive rights,[485] we organize our discussion around the four safe harbors. We also limit our discussion primarily to safe harbors and generation via hosted deployed services.

1. Section 512(a): Transmission

Section 512(a), which applies to "transient digital network communications," protects network-level intermediaries like ISPs.[486] It covers only the "transmitting, routing, or providing connections for, material," and "intermediate and transient" storage appurtenant thereto,[487] "by or at the direction" of users.[488] This transmission and storage must occur "through an automatic technical process without selection of the material by the service provider."[489] This does not describe the way that a model is trained or used. Model trainers choose what data to train on; service providers choose what model to deploy. A model is trained "at the direction" of its creator, not users.[490] It is deployed "at the direction" of a service provider, not users. A model stores its contents, potentially including expression from copyrighted works—for as long as anyone cares to keep a copy of the model. That is the very opposite of "intermediate and transient." And if there were any remaining doubt, the safe harbor only applies when the transmission occurs "without modification of its content."[491] That is very nearly the opposite of what a generative-AI system does. Generation is useful precisely because it modifies and combines content.

---

compound prompt. *See supra* Part I.C.6 (for a description of compound prompts). In such a case, it is not guaranteed that the generation will utilize information from the additional content retrieved by the plugin. *See generally* Shayne Longpre, Kartik Perisetla, Anthony Chen et al., *Entity-Based Knowledge Conflicts in Question Answering*, *in* 2021 EMPIRICAL METHODS NAT. LANGUAGE PROCESSING (EMNLP) 2021 (2021) (for a discussion of when content added to the prompt can and cannot override information learned from the training data). *See supra* note 265 and accompanying text (for a discussion of retrieval models). *See* Elizabeth Lopato, *Perplexity's Grand Theft AI*, THE VERGE (June 27, 2024), https://www.theverge.com/2024/6/27/24187405/perplexity-ai-twitter-lie-plagiarism (for a discussion of Perplexity AI, a generative-AI-assisted search engine).

[485] *See supra* Part II.B.

[486] 17 U.S.C. § 512(a).

[487] *Id.*

[488] *Id.* § 512(a)(1).

[489] *Id.* § 512(a)(2).

[490] One exception is fine-tuning APIs and services that expose fine-tuning functionality to users of services. *See supra* Part I.C.6; *supra* notes 454, 456, 470, 485, and accompanying text; Peng, Wu, Allard et al., *supra* note 177. Another exception is when one actor provides training, fine-tuning, or alignment services and hosts infrastructure for a client that chooses what model to train and on which data. In this case, the trainer and deployer is an intermediary that is perhaps analogous to an ISP. This is an emerging business model. *See supra* notes 177, 191 and accompanying text.

[491] 17 U.S.C. § 512(a)(5); *see also id.* § 512(k)(1).

2. Section 512(b): Caching

Similarly, section 512(b), which covers caching services, does not fit generative-AI models and systems. It covers only "intermediate and temporary storage"[492] of "material . . . made available online by a person other than the service provider"[493] that is transmitted to a user "at the direction of that person"[494] and then cached for later transmission to other users,[495] without modification.[496] Many of the objections to the application of the transmission safe harbor also apply here: often, the training and deployment are not at the direction of user,[497] and the storage is not "intermediate and temporary;" generations also do not generally modify training data.[498] There is also a fundamental sequencing problem. The caching must happen *after* the first user request and *before* subsequent user requests. Much of the relevant storage in a model or deployment takes place before any end user requests at all.[499]

3. Section 512(c): User-Directed Storage

Section 512(c), which covers user-generated content (UGC) services that store content at the direction of users is a bit more complicated. It prevents infringement liability "by reason of the storage at the direction of a user of material that resides on a system or network controlled or operated by or for the service provider."[500] The relevant actors in the supply chain typically store material (e.g., training data, models) at their own direction, so this is not something that the 512(c) safe harbor covers. This is a closer miss than 512(a) and 512(b), because Section 512(c) does not have the strict temporary storage and no-modification conditions of the transmission and caching safe harbors.[501] For the most part, a dataset curator chooses what data to include, a model trainer chooses what datasets to train on, and a service developer chooses what models

[492] *Id.* § 512(b)(1).
[493] *Id.* § 512(b)(1)(A).
[494] *Id.* § 512(b)(2)(B).
[495] *Id.* § 512(b)(2)(C).
[496] *Id.* § 512(b)(2)(A).
[497] *See supra* note 493 and accompanying text.
[498] This is unless generations and prompts get looped into updating a model, which can happen as a part of alignment. *See supra* Part I.C.8.
[499] With the possible exceptions of user prompts (which are unlikely to be transmitted to another user without modification) and user-directed training through hosted services. *See supra* note 493 and accompanying text.
[500] 17 U.S.C. § 512(c).
[501] *Cf.* UMG Recordings, Inc. v. Shelter Cap. Partners, 667 F.3d 1022, 1035 (9th Cir. 2011) (allowing video host to "modify user-submitted material to facilitate storage and access"); Viacom Int'l, Inc. v. YouTube, 676 F.3d 19, 39–40 (2d Cir. 2012) (similar).

to incorporate. With the exceptions of storing user-supplied prompts,[502] or user-supplied fine-tuning datasets (for fine-tuning APIs) and the resulting fine-tuned models, none of the listed use-cases are user-directed storage. There is a possible argument that, for example, when a user supplies a prompt, they are directing the service host to incorporate it into the overarching system. However, this could similarly cut in the other direction, as asking a service to produce a generation is arguably fundamentally different than uploading content intended to be stored for viewing by other users.

4. Section 512(d): Search Engines

Similarly, Section 512(d) prevents liability "by reason of the provider referring or linking users to an online location containing infringing material or infringing activity, by using information location tools, including a directory, index, reference, pointer, or hypertext link."[503] Only a few stages of the generative-AI supply chain fit this description. To the extent that a model or application contains infringing material, it typically *contains* that material, rather than linking to it.[504]

One exception is that some datasets do consist primarily of links to external resources. LAION, for example, includes some metadata about images, but the images themselves are not included. Anyone who wants to train on "the LAION dataset" must download those images themselves.

Another exception is generation-time plugins. As we discuss above,[505] plugins can behave like search engines. They can pull in more up-to-date content that was not included during training, to inform generations with the hope of improving generation quality. It is possible that a plugin could perform a web search and summarize the resulting information in its output generation.[506] Of course, this could result in including infringing content in the generation,[507] but could also potentially lead to a generation linking to infringing content, which may reasonably fall under Section 512(d).

[502] As we have noted above, such prompts can include exact or near copies of copyrighted data.
[503] 17 U.S.C. § 512(d).
[504] A model that uses retrieval-augmented generation techniques could plausibly work entirely with an external retrieval dataset and draw from that dataset only at generation time. The efficiency cost here would be even more severe, because the accesses would need to happen on each generation.
[505] *See supra* Part I.C.7.
[506] As in the Oscar winners example for ChatGPT. OpenAI, *supra* note 214.
[507] *See supra* Part II.E.

5. Notice and Takedown

Even though the Section 512 safe harbors largely do not apply to most stages of the generative-AI supply chain, with potentially a few exceptions, the notice-and-takedown rules under sections 512(c) and 512(d) have been influential enough that they are worth discussing briefly. The basic rule is that the safe harbor goes away if the service provider receives a notice about infringing material and fails to disable access to that material.[508] The notice must be specific both about the identity of the copyrighted work being infringed, and about the location where the infringing material is hosted. The point of this regime is to provide the service provider with actionable information that infringement is taking place and how to prevent it. In that sense, it is a codified version of the *Sony*/*Napster* rule for secondary liability on specific knowledge, together with a mechanism for copyright owners to provide service providers with that knowledge. This model has been so influential that users, platforms, and commentators regularly point to it even in contexts where it does not explicitly apply, e.g., outside the United States, for torts other than copyright infringement, and for platforms that are not themselves eligible for the safe harbors.[509] We will return to this observation in the context of generative AI, by way of analogy, later in this paper when we discuss remedies.[510]

H. Fair Use

We have seen that numerous stages of the generative-AI supply chain involve prima facie copyright infringement. This means that copyright's all-purpose defense, fair use, will play a major role in making generative AI possible at all.[511] Others have discussed the fair-use issues in great detail, so we will focus on only a few salient points.[512] Another caution is that fair use is famously case-specific, so no *ex ante* analysis can anticipate all of the relevant issues. For reasons that will become apparent, similar to our discussion of indirect infringement,[513] we proceed backwards through the supply chain, from generations to training data.

---

[508] 17 U.S.C. § 512(c)(1)(C).

[509] *E.g., Do Other Countries Use DMCA?*, DMCA.COM (2023), https://www.dmca.com/FAQ/Will-DMCA-Takedown-work-in-other-countries ("DMCA.com can provide takedown services no matter where your stolen content is hosted.").

[510] *See infra* Part II.K.

[511] 17 U.S.C. § 107.

[512] Peter Henderson, Xuechen Li, Dan Jurafsky et al., Foundation Models and Fair Use (2023) (unpublished manuscript), https://arxiv.org/abs/2303.15715; Sag, *supra* note 365; Michael D. Murray, Generative AI Art: Copyright Infringement and Fair Use (2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4483539; Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J.L. & ARTS 45 (2017).

[513] *See supra* Part II.F.

1. Application to the Generative-AI Supply Chain

a. Generations

We take each of the four fair-use factors in turn for generations:

Factor One ("the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes"[514]): Many generations will be highly transformative in ways that systematically point towards fair use. In his article introducing the concept of transformative use, Pierre Leval wrote that transformation occurs when "the quoted matter is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings."[515] 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 without a human author.[516] But as long as the audience for these generations finds "new information, new aesthetics, new insights and understandings" in them, the purpose of transformative fair use will be served.[517]

That said, other generations will be minimally transformative. When a model memorizes a work and generates it verbatim as an output, there is no transformation in content.[518] Even a non-exact generation can still be non-transformative. The photograph of Ann Graham Lotz used above as an example of memorization is different from the source image (Figure 9); it is noisier. The noise is not new expression that conveys new information and new aesthetics. It is just noise.

The rest of the first factor does not systematically point one direction or the other. Some generations will be put to commercial use (e.g., backgrounds for a music video), and others will be noncommercial (e.g., illustrating an academic article on copyright and 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.[519] Thus, these other subfactors depend entirely on the specific generation.

---

[514] 17 U.S.C. § 107(1).

[515] Pierre N. Leval, *Toward a Fair Use Standard*, 103 HARV. L. REV. 1105, 1111 (1990).

[516] *Cf. supra* Part II.A.

[517] *See* Cariou v. Prince, 714 F.3d 694, at 707 (2d Cir. 2013) (focusing audience perceptions of works rather than author's intentions in assessing transformative use). *See generally* Laura Heymann, *Everything is Transformative: Fair Use and Reader Response*, 31 COLUM. J.L. & ARTS 445 (2008) (assessing transformative use from audience perspective); Joseph P. Liu, *Copyright Law's Theory of the Consumer*, 44 B.C. L. REV. 397 (2003) (discussing audience interests in copyright).

[518] *See supra* Part II.C.2 (regarding memorization).

[519] *See* 17 U.S.C. § 107 (favoring "purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research").

Factor Two ("the nature of the copyrighted work"[520]):This factor does not systematically favor either side; it depends on the model in question—both how it is trained and the generations it produces. Some training data will be primarily informational; some will be primarily expressive. Most of the training data will typically have been "published" within the meaning of copyright law; it would otherwise not be available within the 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."[521] These works will have made their way into training datasets through express breach of confidence. In these cases, the second factor will particularly favor the plaintiff.

Factor Three ("the amount and substantiality of the portion used in relation to the copyrighted work as a whole"[522]): This is a replay of substantial similarity and will not systematically favor either side/ Some generations will closely resemble the works they were copied from; others will copy comparatively smaller portions of the works, both qualitatively and quantitatively.[523] Even when a work is transformative under the first factor, courts will still also inquire into whether the generation copies more than necessary for that transformation. Prompting a model with "painting of a car driving in a snowstorm in the style of Frida Kahlo" might result in a generation that copies just Kahlo's color palette, brushwork, and floral motifs, or it might also put the entire composition of one of her self-portraits inside the resulting generation.

Factor Four ("the effect of the use upon the potential market for or value of the copyrighted work."[524]): The outputs of a non-generative AI do not compete in the market for a copyrighted work in the sense that the fourth factor cares about. It is possible that 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, causing viewers not to want to watch it. 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. While the copyright owner of the movie is harmed, it is not a type of harm that is cognizable under the fourth factor.[525] The outputs of a generative-AI system, however, can substitute for a copyrighted work in the expressive way that copyright cares about. Consider the following variations on a theme:

---

[520] *Id.* § 107(2).
[521] WILLIAM F. PATRY, PATRY ON COPYRIGHT § 6.31 (2023).
[522] 17 U.S.C. § 107(3).
[523] *See* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537 (S.D.N.Y. 2013) (rejecting fair use defense brought by news-monitoring service that reproduced substantial excerpts from articles for its customers).
[524] 17 U.S.C. § 107(4).
[525] *See* Campbell v. Acuff-Rose Music, 510 U.S. 569 (1994).

- An individual cannot obtain a copy of the "The Old Sugarman Place" episode of *Bojack Horseman* at a price they are willing to pay. Instead, they prompt a generative-AI system to generate "'The Old Sugarman Place'", and the system generates a close duplicate. The generation is essentially a pirated edition at a lower price; it competes with the original for this individual's business. This is a paradigmatic fourth-factor harm.

- An individual cannot obtain a copy of the "The Old Sugarman Place" at a price they are willing to pay. Instead, they prompt a generative-AI system

- to generate it, and the system generates a non-exact copy with significant aspects borrowed from the original, but also with significant changes to the dialogue and animation. This episode—call it "The New Sugarman Place"—is also a direct competitor under factor four for this individual'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.

- An individual prompts a generative-AI system to generate a new episode of *Bojack Horseman*. The generation does not necessarily compete with "The Old Sugarman Place," which was unsuitable for the user's needs.[526] 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 yourself.[527] 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.* McDonald's created advertisements in the unsettling style of the children's show *H.R. Pufnstuff*.[528]

- 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, not just *Bojack Horseman*, but other shows as well. If the generative-AI system had not been available, the individual might have paid to watch *Bojack* or *Dr. Katz* or some other

[526] Perhaps they have already watched all of the existing episodes.
[527] 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.
[528] Sid & Marty Krofft Television v. McDonald's Corp., 562 F.2d 1157 (1977).

show or kicked in to a Kickstarter to help commission something new. Many authors might view this as a kind of unfair competition that undercuts the market for their work. But here, the fourth factor is *not even relevant* to the generation, 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 and interviewed on podcasts and late-night shows about their inspirations, 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 too is a factor-four harm. To see why, look to copyright's remedies: copyright law awards the infringer's profits, even when the copyright owner has not suffered lost sales.[529] It may be helpful to think of this as a case in which 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—a conclusion that forces a reconsideration of whether the underlying models in the generative-AI systems that produced these generations are fair uses.

b. Models

There is a strong argument that training (and deploying) *non-generative*-AI systems is fair use.[530] The best explanation of this conclusion is Matthew Sag's concept of nonexpressive uses—bulk uses of copyrighted works that do not involve the consumption of expression.[531] Examples include digital stylometry,

[529] *See infra* Part II.K.

[530] *See, e.g.,* Mark Lemley & Bryan Casey, *Fair Learning*, 99 TEX. L. REV. 743 (2021) (arguing that most such training is fair use and approving of this pattern); Grimmelmann, *supra* note 343 (agreeing descriptively, but with some normative skepticism); Levendowski, *supra* note 151 (arguing that copyright law can introduce bias into training datasets and that fair use can address this bias); Amanda Levendowski, *Resisting Face Surveillance with Copyright Law*, 100 N.C. L. REV. 1015 (2022) (arguing that training for facial recognition should not be a fair use).

[531] Matthew Sag, *The New Legal Landscape for Text Mining and Machine Learning*, 66 J. COPYRIGHT SOC'Y USA 291 (2019).

sentiment analysis, and plagiarism detection.[532] These uses do not involve the human encounter with expression as a listener that lies at the heart of the copyright system.[533] In that sense, these models do not compete with authors.

Training a model for these purposes may implicate other important societal interests, but they are not typically described as copyright interests.[534] The reasoning here is essentially backward-looking. Because the ultimate use does not implicate copyright at all, the intermediate steps of model training, fine-tuning, and aligning, and system deployment do not involve copying in a way that competes with authors.

This is essentially the logic behind the Google Books fair-use decisions.[535] The courts held that the ultimate uses to which the scanned books were put were either fair uses or non-copyright-implicating: provision of books to print-disabled patrons, short (fair-use) snippets for search results, and directing users to relevant books. Additionally, the digital humanities research corpus proposed in the (rejected) settlement agreement would also be fair use under this rule.[536] It would have created a full-text corpus of all of the scanned books, against which researchers could run algorithmic analyses. Other aspects of the settlement attracted vociferous criticism, particularly its treatment of orphan works, but the research corpus was not a principal focus of copyright owners' objections.[537] When the settlement was ultimately rejected, the research corpus played no role in the court's decision.[538]

This categorical argument does not work for generative-AI models that can generate expressive works. Some outputs from these models will incorporate copyrighted material that will be seen by humans—indeed, some generations will infringe. Once the outputs of a system can infringe, the argument that the system itself does not implicate copyright's purposes no longer holds.

Most of the analysis of generations carries back to models, but there are a few notable differences:

- Many models *qua* models are arguably highly transformative. They represent works internally in new and very different ways. They are also capable of generating highly transformative works as outputs.

[532] *See id.* (surveying caselaw and applications).
[533] *See* Grimmelmann, *supra* note 343.
[534] *See, e.g.,* Levendowski, *supra* note 533 (concerning privacy).
[535] Authors Guild v. Google, Inc., 804 F.3d 202, 228 (2d Cir. 2015); Authors Guild, Inc. v. HathiTrust, 755 F.3d 87, 98 (2d Cir. 2014).
[536] *See* Proposed Settlement Agreement, Authors Guild v. Google, Inc., 770 F. Supp. 2d 666 (S.D.N.Y. Oct. 28, 2008) (No. 1:05-cv-08136) (Doc. No. 56).
[537] *See generally* The Pub.-Int. Book Search Initiative, Objections and Responses to the Google Books Settlement: A Report (2010), https://james.grimmelmann.net/files/articles/objections-responses-2.pdf (describing criticisms).
[538] *Authors Guild*, 770 F. Supp. 2d 666.

- The amount copied in a model is potentially much greater than the amount that appears in any particular generation. How much of a work is present in a model is, as discussed above, a difficult conceptual and empirical question.[539] It is also possible that the portion copied in a model includes the "heart" of the work, those portions which are most significantly responsible for its appeal.[540] To the extent that a model is successful at embedding distinctive features of works, it may disproportionately capture their "hearts."[541]

- Whether there is a licensing market for generative-AI models is a difficult question.[542] The question itself is circular because the existence of a licensing market counts in favor of the copyright owner under the fourth factor—but if this copying is a fair use, then no such market can develop.[543] In previous AI cases, courts have largely found that such markets do not exist, but that reasoning may have been influenced by the fact that they were considering non-generative AIs.[544] With the advent of generative-AI systems, this question is open again. There is not at present such a market, but many large commercial copyright actors are moving towards trying to create one. Getty's litigation against Stability AI is aimed at forcing licensing negotiations,[545] as is the *New York Times's* lawsuit against Microsoft and OpenAI.[546]

Even if a base model is deemed to have substantial noninfringing uses, downstream fine-tuned or aligned models may have a substantively different fair-use analysis. As we have emphasized before, both fine-tuning and alignment can involve additional copyrighted data. Additionally, the actor fine-tuning or aligning the model has some control over the types of outputs generated from the model and may nudge the model either towards or away from infringing

---

[539] *See supra* Part II.C.2.
[540] Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 538–39 (1985).
[541] Or not. But this is the kind of question that must be asked.
[542] *See* Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 930 (2d Cir. 1994) (considering whether a licensing market is "traditional, reasonable, or likely to be developed").
[543] *See generally* Jennifer E. Rothman, *The Questionable Use of Custom in Intellectual Property*, 93 VA. L. REV. 1899 (2007); James Gibson, *Risk Aversion and Rights Accretion in Intellectual Property Law*, 116 YALE L.J. 882 (2006).
[544] *E.g.,* A.V. *ex rel.* Vanderhye v. iParadigms, LLC, 562 F.3d 630 (4th Cir. 2009).
[545] *Getty Images Statement*, GETTY IMAGES (Jan. 17, 2023), https://newsroom.gettyimages.com/en/getty-images/getty-images-statement.
[546] Complaint ¶ 7, N.Y. Times Co. v. Microsoft, No. 2:24-cv-00711 (C.D. Cal. Dec. 27, 2023).

generations.[547] Both actions may shift the balance of infringing and noninfringing uses. For example: if a fine-tuned model has mostly infringing uses, is this due to changes introduced by training on the fine-tuning dataset? If not, it could be argued that the fine-tuned model is eliciting more infringing uses that are latent in the base model. In turn, should this change our analysis of the balance of infringing or noninfringing uses for the base model?

Another consideration for released models is commerciality. A hosted service that charges end users for generations is a commercial use, even if some of those users make non-commercial uses of the generations. Similarly, a paid licensing agreement to embed a model in an application or API is commercial. On the other hand, an open release of a model under a license that allows others to use it for free is non-commercial. These different contexts may have different ramifications for fair-use defenses.

All in all, the fair-use case for models is stronger than for generations in some ways, and weaker in others. It is plausible that a court could hold that a model is a fair use, but that some of its outputs are not. It is also plausible that that a model that is not a fair use could produce some outputs that are fair uses. It seems unlikely, however, that an unfair model could produce *only* fair uses.

c. Training Datasets

Finally, we come to the fair-use analysis of the training datasets that include copyrighted material. As above, there is a solid non-expressive-use argument that training datasets are fair, as long as they are only used as inputs to training non-generative-AI models. If the steps of training and using a non-generative model are non-expressive fair use, then so are the preparatory steps of assembling a dataset.[548] As above, that argument breaks down when a training dataset is used to train generative-AI models. Even if it is also used to train non-generative-AI models, the non-expressive use argument fails once the dataset is an input into generative models that can produce outputs that reproduce copyrighted expression. In addition, because a dataset can be used to train many models, it is possible that a model could be unfair even though the dataset it was trained on is fair. Here is a four-factor analysis of training datasets:

Factor One: The transformativeness, if any, in datasets is of a different kind than models and generations. Datasets are not transformative in content; the works may be reformatted and standardized, but there is no new expression.[549]

[547] *See supra* note 477 and accompanying text.

[548] *See* Sag, *supra* note 534.

[549] Synthetic datasets again pose a wrinkle, since they collapse the boundary between generations and data. Synthetic data produced by a generative-AI model could be viewed as a transformative use of the underlying training data on which the synthetic-data-generating model was trained. Gokaslan, Cooper, Collins et al., *supra* note 37 (discussing how using an image-to-text model to produce captions for images could be viewed as a transformation of rich images to "lossy" text—like data compression).

The work itself has been compiled and arranged with other works, but it is unchanged. On the other hand, there is an argument that assembling a dataset for AI training is a transformative purpose: it is a use of a different sort than the usual expressive uses for the work itself.

Additionally, many training datasets are made publicly available noncommercially. Some observers have argued that this amounts to a kind of ethical and legal laundering by the commercial companies that then train on those datasets—especially when there is a funding relationship between the two.[550] The factor-one commerciality analysis of the dataset may therefore turn on the activities of parties besides the dataset curator.

Factor Two: Most datasets will include mostly published works. They may include both expressive and informational works, as discussed above. The balance will depend on the dataset.

Factor Three: The dataset typically copies complete works verbatim. This wholesale copying is justified, if at all, in light of the transformative purpose it serves. A model may or may not need to reproduce entire works, depending on the model and its purposes. If a therapy chatbot memorizes entire books, for example, that is an undesirable side effect, not the model's goal.[551] But there is often a strong case that a training dataset should retain as much information as possible *to make it useful for model training*. It may be more information than many models need, and they will discard much of it during the training process. But it is much easier to discard information that is present in the training data than to recover information that is absent from the training data.

Factor Four: The market for licensing works for training datasets is all but indistinguishable from the market for licensing works for AI training.

Finally, there a strong possibility that a training dataset could be considered an unfair use simply because it provides public access to a substantial number of copyrighted works, *independently of its use as training data*. This seems likely to be the case, for example, for the Books3 dataset, "a library of around 196,000 books, including works by popular authors like Stephen King, Margaret Atwood, and Zadie Smith."[552] This dataset, which is drawn from a "shadow library" of almost-certainly infringing books, is very likely unfair.

One factor that might weigh on a court's decision-making is whether a model trainer knew or should have known that a dataset was infringing. Although bad faith is not officially part of the four factors, courts do sometimes emphasize the

[550] Baio, *supra* note 481.

[551] Of course, it might not be possible to make the chatbot convincing without significant memorization, but the memorization is still not the goal.

[552] Kate Knibbs, *The Battle Over Books3 Could Change AI Forever*, WIRED (Sept. 4, 2023), https://www.wired.com/story/battle-over-books3/; *see also* Alex Reisner, *Revealed: The Authors Whose Pirated Books are Powering Generative AI*, THE ATLANTIC (Aug. 19, 2023), https://www.theatlantic.com/technology/archive/2023/08/books3-ai-metallama-pirated-books/675063/.

defendant's bad intentions or unethical conduct in finding no fair use.[553] Thus, a court might treat a company that trained on Books3 without knowing the details of its origins more leniently than a company that trained on it with full knowledge of its infringing contents.

I. Express Licenses

For both of our brief sections on licenses,[554] we discuss copyright law and the supply chain together. A license from the copyright owner is a complete defense to infringement.[555] It could hardly be otherwise. The modern copyright system depends on licenses voluntarily granted by authors to publishers.

Some creators have expressly agreed to allow their works to be used for training the models used in generative-AI systems.[556] Only such a license from the copyright owner—or from a licensee who is allowed to grant sublicenses—is effective. A dataset creator/curator or model trainer cannot simply rely on the license a work bears. That license might have been applied by someone who did not have the authority to do so. In this case, it is hornbook law that the license is ineffective, and anyone who relies on it is an infringer. There is no defense of good-faith reliance on a purported license. Improperly licensed works can be removed from a dataset once the mistake is noticed. But it will be much harder to remove those works them from a model trained on reliance on them.[557]

Some licenses are specific. They allow a specific named licensee to use the work for specified purposes. Adobe's Firefly, for example, claims to be trained in substantial part on images licensed by their creators to Adobe Stock.[558] Only Adobe can use those works for training.

---

[553] *E.g.,* Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 563 (1985) (the defendant "knowingly exploited a purloined manuscript").

[554] *See infra* Part II.J (implied licenses).

[555] *See generally* JORGE L. CONTRERAS, INTELLECTUAL PROPERTY LICENSING AND TRANSACTIONS: THEORY AND PRACTICE (2022) (discussing IP licensing).

[556] *See, e.g.,* Mia Sato, *Grimes Says Anyone Can Use Her Voice for AI-Generated Songs*, THE VERGE (Apr. 24, 2023), https://www.theverge.com/2023/4/24/23695746/grimes-aimusic-profit-sharing-copyright-ip.

[557] *See* Meng, Bau, Andonian & Belinkov, *supra* note 476; Bourtoule, Chandrasekaran, Choquette-Choo et al., *supra* note 476; Cooper, Choquette-Choo, Bogen, Jagielski, Filippova, Liu et al. *supra* note 405 (regarding the difficulty of model editing and "unlearning").

[558] *See* Benj Edwards, *Ethical AI art generation? Adobe Firefly may be the answer*, ARS TECHNICA (Mar. 22, 2023), https://arstechnica.com/information-technology/2023/03/ethical-ai-art-generation-adobe-firefly-may-be-the-answer/. *But see* Sharon Goldman, *Adobe Stock Creators Aren't Happy With Firefly, the Company's 'Commercially Safe' Gen AI Tool*, VENTUREBEAT (June 20, 2023), https://venturebeat.com/ai/adobe-stockcreators-arent-happy-with-firefly-the-companys-commercially-safe-gen-ai-tool/ (noting that some

These specific licenses apply to only a small fraction of the works currently being used as training data.[559] Models trained only with this kind of specific permission are rare. They are often lower quality than the most cutting-edge generative-AI models.[560]

Other licenses are general. They allow *anyone* to use a work in specified ways, not just an individual named licensee. Here, anyone is allowed to engage in a use as long as it complies with the terms of that license, even if the user of the work[561] has never directly interacted with the copyright owner to obtain individual permission. We will use Creative Commons (CC) licenses as an example, as the terms in the Creative Commons license suite cover a useful range of interesting conditions.

Some materials are provided under a public-domain mark, which indicates that there are no copyright interests in the material.[562] Others are provided under a Creative Commons Zero notice, which indicates that the copyright owner has dedicated the material to the public domain.[563] Any and all uses of these works are allowed by anyone, without risk of copyright infringement.

The basic license grant in every other Creative Commons license is the right to "reproduce and Share the Licensed Material, in whole or in part; and produce, reproduce, and Share Adapted Material."[564] This covers all of the section 106 exclusive rights, and it covers all of the activities involved in compiling training datasets, model training and fine-tuning, deployment, generation, alignment, and use of the generated material. So, unless some other license term restricts this

artists did not understand that the licenses they entered into by providing their images to Adobe Stock included terms allowing Adobe to use the images for training generative models).

[559] *See generally* Benjamin L.W. Sobel, *A Taxonomy of Training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning, in* ARTIFICIAL INTELLIGENCE AND INTELLECTUAL PROPERTY 221 (Jyh-An Lee, Reto Hilty & Kung-Chung Liu eds., 2021) (discussing different categories of licensed works in training datasets).

[560] Workshop, Scao, Fan et al., *supra* note 154. Gokaslan, Cooper, Collins et al., *supra* note 37.

[561] For our purposes, this could be the dataset creator/curator, base model trainer, fine-tuner, model aligner, a generative-AI system user supplying a licensed work as a prompt, or the deployed service host's generation process pulling in external content via a plugin.

[562] *Public Domain Mark 1.0* (2023), https://creativecommons.org/publicdomain/mark/1.0/.

[563] *CC0 1.0 Universal (CC0 1.0) Public Domain Dedication* (2023), https://creativecommons.org/publicdomain/zero/1.0/.

[564] *Creative Commons Attribution 4.0 International License* § 2(a)(1)(A) (2023), https://creativecommons.org/licenses/by/4.0/legalcode.

grant, generative-AI systems are fully and expressly licensed to use any CC-licensed material in their training data.[565]

The attribution term in BY licenses requires that the user of the work retain the creator's identification, indicate whether the work is modified, and retain the Creative Commons license notice. This requirement can be satisfied in "any reasonable manner based on the medium, means, and context."[566] A training dataset could provide this information through suitable metadata, but many datasets do not.[567] If liability were a serious concern, and the availability of CC-licensed material sufficiently broad to justify it, it is possible that more datasets would bear these attributions, so that they would be fully allowed under CC-BY licenses.

This, however, is where attribution stops with current opaque generative-AI models. These models do not attempt to store information about the attribution of the works they were trained on.[568] To the extent that they copy from their CC-BY-licensed training data, these models are derivative works that do not bear proper attribution, so they fall outside the scope of the license. A model that does not retain attribution information cannot provide that information in its generations, so the generations also fall outside the license.

The non-commercial term in NC licenses prohibits uses "primarily intended for or directed towards commercial advantage or monetary compensation."[569] This definition roughly tracks the way in which commerciality is defined in fair use, as discussed above. It seems likely that the sale and licensing of datasets and models, and the provision of generations for money would be considered commercial. So, this term would allow entirely open-source supply chains but prohibit any commercial links in those chains.

---

[565] However, training a model on only CC-licensed material does not guarantee that this model's resulting generations could not be substantially similar to copyrighted expression. *See* Cooper, Choquette-Choo, Bogen, Jagielski, Filippova, Liu et al., *supra* note 405 (for an example of generating an image resembling Mickey Mouse using a text-to-image diffusion model trained only on CC-licensed images.) .

[566] *Id.* § 3(a)(1)(A)(i).

[567] Lee, Ippolito & Cooper, *supra* note 67, at 5.

[568] *See supra* note 119 and accompanying text (for the challenges of attribution). Indeed, attribution is one of the motivations for using RAG: the hope is that the specific, retrieved examples will have a greater influence on the generation, thereby making attribution easier. *See supra* note 212 and accompanying text (for a discussion of retrieval augmented generation). In practice, however, this varies from generation to generation. *See generally* Longpre, Perisetla, Chen et al., *supra* note 487 (for an evaluation of how often generations are based on the retrieved context when the retrieved context is provided).

[569] *Creative Commons Attribution-NonCommercial 4.0 International License* § 1(i) (2023), https://creativecommons.org/licenses/by-nc/4.0/.

The no-derivatives term in ND licenses allows the user to copy and share the work itself: to "produce and reproduce, but not Share, Adapted Material."[570] Adapted Material is defined as "material . . . that is derived from or based upon the Licensed Material and in which the Licensed Material is translated, altered, arranged, transformed, or otherwise modified in a manner requiring permission"[571] from the copyright owner. In other words, it is any derivative work under copyright law. An ND license therefore allows dataset curation (as datasets are compilations, not derivatives). But it probably prohibits model training, because a model is most likely a derivative work. So, one could train models for research but not share them. The only way for models to escape from the ND term is for them not to be substantially similar to the copyrighted work and thus escape from copyright law entirely. Generations, too, are derivative works unless they are so substantially identical to a training example that they are memorized duplicates rather than generations,[572] or unless they are so substantially dissimilar from all training examples that they do not infringe at all. The upshot is that an ND-license is effectively no license at all for models and generations.[573]

The share-alike term in SA licenses does allow for the sharing of derivative works, but they must be placed under the same Creative Commons license that the underlying works were licensed under.[574] So a model trained on BY-SA works would itself need to be shared BY-SA—if it is shared at all. A trainer who keeps the model in-house, and uses it only to power a generation service, does not trigger the distribution threshold that causes the share-alike condition to kick in. If the model is under an SA license, then most generations from it are derivative works of the model and themselves need to be shared SA. If the model is not SA, then only those generations that are derivative works of the original SA work need to be shared SA. Unlike with BY, this relicensing is feasible without individual attribution—a blanket BY-SA license applied to a dataset, a model, or a generation would suffice.

But note that it would probably not be possible to train a single model on both BY-SA and BY-NC-SA works. Each license requires that any derivative works be released under *that license*. And each license states that the licensee "may not offer or impose any additional or different terms or conditions" on the work.[575]

---

[570] *Creative Commons Attribution-NoDerivatives 4.0 International License* § 2(a)(1)(B) (2023), https://creativecommons.org/licenses/by-nd/4.0/.
[571] *Id.* § 1(a).
[572] *See supra* Part II.C.2 (discussing memorization of training data).
[573] *See supra* Part I.B (concerning derivative works in the generative-AI supply chain).
[574] *Creative Commons Attribution-ShareAlike 4.0 International License* § 3(b)(1) (2023), https://creativecommons.org/licenses/by-sa/4.0/.
[575] *Id.* (3)(b)(3); *Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License* (3)(b)(3) (2023), https://creativecommons.org/licenses/by-nc-sa/4.0/.

Lastly, it is worth noting that generation-time plugins or a RAG system could pull in additional data that is not expressly licensed or further complicates our compatibility analysis above. To summarize:

- An attribution requirement is a difficult technical problem, and no current systems do it effectively.[576]
- A non-commerciality requirement is feasible for most fully open-source supply chains, but difficult for many proprietary ones.
- A no-derivatives requirement effectively prohibits generative AI
- A share-alike requirement is feasible and tries to compel AI developers to contribute their models to a share-alike commons, but may not reach all generation services, and may raise license-compatibility issues.
- Generation-time plugins could complicate licensing compatibility considerations.

The punch line is that BY is a common term in all of the six standard Creative Commons licenses. No current generative-AI model is licensed under any CC license.[577] Neither are any of their generations. All of the other license terms are irrelevant. For now, at least, CC licensing is a dead-end for generative AI.

J. Implied Licenses

Implied copyright licenses arise when a copyright owner's conduct gives rise to an inference that they have consented to particular uses.[578] No particular formalities are required to create one.[579] Caselaw holds that the act of putting material online on the web typically creates an implied license for search engines to index it and for archives to maintain archival copies of it.[580] There is also some suggestion that this implied license only applies where the owner has not used a robots.txt file or exclusion headers to deny permission for bulk crawling.[581] The

[576] *See supra* note 119 and accompanying text (discussing challenges of attribution); *supra* note 212 and accompanying text (discussing retrieval augmented generation).
[577] *About the Licenses*, Creative Commons (2023), https://creativecommons.org/licenses/.
[578] Effects Assocs., Inc. v. Cohen, 908 F.2d 555 (9th Cir. 1990).
[579] Oddo v. Ries, 743 F.2d 630 (9th Cir. 1984).
[580] Field v. Google Inc., 412 F. Supp. 2d 1106, 1115–17 (D. Nev. 2006).
[581] *Id.* at 1117. A prominent training dataset, the Common Crawl, respects the robots.txt protocol. *See Frequently Asked Questions*, COMMON CRAWL (2023), https://commoncrawl.org/faq. However, recent reports indicate that some generative-AI

implied license probably does not apply to material behind a paywall or login form that a search engine accesses through surreptitious means.[582] But it probably does apply to material that a website has specifically made available to a particular search engine.[583]

The relevant question, then, is what the scope of this implied license is.[584] If I put a photograph online with no further information, it is well-established that this act by itself does not grant permission to third parties to use the photograph in news articles or other publications.[585] The implied license allows them to copy the photograph as part of viewing it on my page, but not to use it in other contexts.[586]

A training dataset seems broadly akin to the kind of archives that courts have held to be covered by the implied license in other cases.[587] User-supplied prompts, which could become future training data, could be covered by implied licenses, but also could involve express licenses when a user consents to use a particular service.

It is a little harder to say that model training fits within an implied license. This is a new use, one that did not exist when much of the data examples, which have recently been re-purposed for generative-AI training datasets, were first put

companies may be ignoring the robots.txt protocol to power their generative-AI services. *See* Katie Paul, *Exclusive: Multiple AI companies bypassing web standard to scrape publisher sites, licensing firm says*, REUTERS (Jun. 21, 2024), https://www.reuters.com/technology/artificial-intelligence/multiple-ai-companies-bypassing-web-standard-scrape-publisher-sites-licensing-2024-06-21/; Dhruv Mehrotra & Tim Marchman, *Perplexity Is a Bullshit Machine*, WIRED (Jun. 19 2024), https://www.wired.com/story/perplexity-is-a-bullshit-machine/.

[582] Sites that use such barriers may also have express licensing in place for datasets based on their data.

[583] *Cf. Structured Data for Subscription and Paywalled Content (CreativeWork)*, GOOGLE SEARCH CENT. (May 23, 2023), https://developers.google.com/search/docs/appearance/structured-data/paywalled-content (describing how to make paywalled content accessible to Google's indexing bot).

[584] *See generally* Christopher M. Newman, *"What Exactly Are You Implying?": The Elusive Nature of the Implied Copyright License*, 32 CARDOZO ARTS & ENT. L.J. 501 (2014).

[585] This point is most clearly seen in the cases holding that news publishers cannot embed photographs posted to Instagram or other social networks *E.g.,* Sinclair v. Ziff Davis, LLC, 454 F. Supp. 3d 342 (S.D.N.Y. 2020).

[586] Agence Fr. Presse v. Morel, 769 F. Supp. 2d 295, 302–03 (S.D.N.Y. 2011) (holding that the license a user granted to Twitter when he uploaded photographs did not run in favor of third-party publishers who downloaded the photographs from Twitter).

[587] *E.g.,* Field v. Google Inc., 412 F. Supp. 2d 1106 (D. Nev. 2006) (Google Cache); Parker v. Yahoo!, Inc., 88 U.S.P.Q.2d 1779 (E.D. Pa. 2008) (Yahoo and Microsoft search). *But see* MidlevelU, Inc. v. ACI Info. Grp., 989 F.3d 120 (11th Cir. 2021) (accepting *Field* but holding, "Implied permission to enter through a front door (web crawler) does not also imply permission to enter through a back window (RSS feed).").

online.[588] With respect to re-purposing materials, there is a useful analogy here to the Google Books case. Book scanning did not exist when most of the books in the corpus were published, so it is hard to say that authors and publishers consented to scanning when they published.[589] It is harder still to say that putting material online constitutes an implied license to use that material in AI generations.[590] It is certainly the case that many copyright owners strenuously object to this practice. And if a court is to say that generation is allowed, fair use (which applies whether or not the copyright owner consents) is a better fit for the facts than an implied license (which applies only when the copyright owner consents).

This said, the fact that materials were voluntarily placed online can be relevant to the fair-use inquiry. As in *Sony*, which held that taping over-the-air television programs for time-shifting was a fair use, the choice to publish involves giving users access to a work.[591] Copyright owners did not need to license their works for broadcast; they had other alternatives that did not invite the public to view for free. One would not draw a similar inference from the choice to show a movie in theaters. So even if there is not an implied license as such for AI training, the fact that there is a broadly shared practice of putting material online, where any web user can view, helps to support a fair-use defense for generative-AI systems and users.

In addition, other laws, such as trespass to chattels and the Computer Fraud and Abuse Act, may sometimes restrict the ability of dataset compilers to scrape data.[592] These other laws, however, typically only apply against the party that actually scrapes the data. They do not apply against others who come into possession of the data that was scraped, such as model trainers or application deployers.[593] Only copyright runs with the data itself; because of these laws, only

---

[588] *See supra* Part I.B.4 (regarding web-scraped datasets); *supra* Part I.C.2 (regarding data creation); *supra* Part I.C.3 (regarding the creation and curation of training datasets from previously created data).

[589] *See generally* Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

[590] *Cf.* Associated Press v. Meltwater U.S. Holdings, Inc., 931 F. Supp. 2d 537 (S.D.N.Y. 2013) (holding that excerpting of between 4.5% and 61% of news articles in a subscription news-monitoring service was not covered by implied license).

[591] Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417, 456 (1984) (" Sony demonstrated a significant likelihood that substantial numbers of copyright holders *who license their works for broadcast on free television* would not object to having their broadcasts time-shifted by private viewers.") (emphasis added).

[592] *See generally* Benjamin L.W. Sobel, *A New Common Law of Web Scraping*, 25 LEWIS & CLARK L. REV. 147 (2021).

[593] One exception is criminal laws against the possession and distribution of child sexual abuse material (CSAM). These laws apply to anyone who possesses such material, regardless of how they obtained it. For example, researchers discovered that images linked from the LAION dataset included CSAM. *See supra* note 81 and accompanying text;

copyright is a right to own information as such. And even where these other laws apply, their scope can be quite limited. They typically allow the scraping of publicly accessible material unless there is some additional element of harm to the site being scraped, such as an impairment of its ability to serve others.[594]

K. Remedies

The Copyright Act allows for a broad array of remedies against infringers.[595] Some of them could be highly significant in shaping the deployment of generative-AI systems.[596] We organize our discussion in this section around these different remedies.

1. Damages and Profits

A successful copyright plaintiff is entitled to recover "the actual damages suffered by him or her as a result of the infringement."[597] This is a damage remedy measured by the victim's harm. It consists of the money the plaintiff *lost* as a result of the infringement, such as decreases in sales or cancelled licensing contracts with third parties. In *Harper & Row, Publishers, Inc. v. Nation Enterprises*, for example, *Time* cancelled a contract to publish excerpts of Gerald Ford's memoirs when *The Nation* published infringing excerpts ahead of the book's publication date.[598] These actual out-of-pocket losses, however, are rare and hard to prove, so the Copyright Act allows a variety of alternative theories to ground an award of damages.

DAVID THIEL, IDENTIFYING AND ELIMINATING CSAM IN GENERATIVE ML TRAINING DATA AND MODELS (Stanford Internet Observatory 2023), https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-23.pdf. Note that anti-CSAM laws are enforced by the government, or in some cases by victims; they are not IP laws giving possessors of a dataset the right to stop others from using the data in the dataset.

[594] *See, e.g.,* hiQ Labs, Inc. v. LinkedIn Corp., 31 F.4th 1180 (9th Cir. 2022); *see also* Internet Archive v. Shell, 505 F. Supp. 2d 755 (D. Colo. 2007) (rejecting racketeering claims against Internet Archive for scraping and archiving webpages).

[595] *See generally* DOUGLAS LAYCOCK & RICHARD L. HASEN, MODERN AMERICAN REMEDIES: CASES AND MATERIALS (5th ed. 2018) (discussing types of remedies available under United States law).

[596] *See generally* Pamela Samuelson, *How to Think About Remedies in the Generative AI Copyright Cases*, LAWFARE (Feb. 15, 2024), https://www.lawfaremedia.org/article/howto-think-about-remedies-in-the-generative-ai-copyright-cases (discussing potential remedies in the current generative-AI copyright lawsuit landscape).

[597] 17 U.S.C. § 504(b).

[598] Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539 (1985).

The simplest such theory is that the plaintiff's damages can be measured by the lost licensing fee that the defendant saved by infringing.[599] This is a fair-market-value remedy; the plaintiff is awarded the licensing fee that a willing seller and willing buyer would have negotiated.[600] As with fair use, much depends on the existence of a licensing market for the kind of use at issue. If there is no such market, it can be hard for a court to estimate an appropriate royalty. So, for example, while there is a well-functioning market for licensing new editions of books, there is not a market for licensing generative-AI training on books—because the use has not existed until now, neither has the market.[601] In addition, it can be difficult for individual plaintiffs to show that their work in particular has a high licensing value.[602] In *On Davis v. The Gap, Inc.*, for example, the plaintiff requested a $2,500,000 licensing fee for the unauthorized use of his eyewear in a Gap ad.[603] The court held that his evidence supported a licensing fee of $50.[604]

Recognizing that this too may be an inadequate measure of damages, the Copyright Act also allows a successful plaintiff to recover "any profits of the infringer that are attributable to the infringement and are not taken into account in computing the actual damages."[605] Instead of measuring the plaintiff's losses, this remedy measures the defendant's unfair gains.[606] The Copyright Act has a burden-shifting provision for defendant's profits that on paper is quite generous to the copyright owner:

> In establishing the infringer's profits, the copyright owner is required to present proof only of the infringer's gross revenue, and the infringer is required to prove his or her deductible

[599] *E.g.,* Dash v. Mayweather, 731 F.3d 303, 313 (4th Cir. 2013) ("Under the lost licensing fee theory, actual damages are generally calculated based on "what a willing buyer would have been reasonably required to pay to a willing seller for [the] plaintiffs' work.") (internal quotation omitted).

[600] *Id.*

[601] One reason for copyright owners to enter into licensing arrangements with some AI companies may thus be to establish a baseline for calculating damages against others who do not agree to licensing arrangements.

[602] *E.g., Dash*, 731 F.3d at 312–26 (rejecting licensing fee calculation in plaintiff's expert report).

[603] On Davis v. The Gap, Inc., 246 F.3d 152, 156 (2d Cir. 2001).

[604] *Id.* at 161.

[605] 17 U.S.C. § 504(b).

[606] This makes infringer's profits a *restitutionary* remedy rather than a compensatory remedy. *See generally* WARD FARNSWORTH, RESTITUTION: CIVIL LIABILITY FOR UNJUST ENRICHMENT (2014) (discussing the theory of restitution). The provision is phrased the way it is to avoid double-counting. If the plaintiff loses one sale to the defendant, that sale would be "profits of the infringer" that *are* "taken into account in computing the [plaintiff's] actual damages."

> expenses and the elements of profit attributable to factors other than the copyrighted work.[607]

The hard part is determining how much of the defendant's profits are "attributable to factors other than the copyrighted work." In a generative-AI context, we would ask, how much of a generation's value is due to a particular training work, as opposed to other training works and the training algorithm? This is a hard question by itself; answering the same question for a model requires answering it for all generations the model is used to produce, and adding up the results.[608] In practice, the answer may depend on who bears the burden of persuasion on the relative value of different elements.

In this vein, there is an illuminating passage in *On Davis*, where the court held that none of the Gap's overall profits were attributable to the use of the defendant's eyewear in one photograph.[609] Explaining its reasoning, the court wrote:

> Thus, if a publisher published an anthology of poetry which contained a poem covered by the plaintiff's copyright, we do not think the plaintiff's statutory burden would be discharged by submitting the publisher's gross revenue resulting from its publication of hundreds of titles, including trade books, textbooks, cookbooks, etc. In our view, the owner's burden would require evidence of the revenues realized from the sale of the anthology containing the infringing poem. The publisher would then bear the burden of proving its costs attributable to the anthology and the extent to which its profits from the sale of the anthology were attributable to factors other than the infringing poem, including particularly the other poems contained in the volume.[610]

On this analogy, a generation might be like an anthology. Once the plaintiff shows that an infringing generation has commercial value, the defendant bears the burden to show what portion of the value came from other sources—a burden that may be quite difficult to meet. So, to a first approximation, those who profit from infringing generations should expect to pay out their entire profits.

Also on this analogy, a generative-AI system (or model or training dataset) might be more like a full catalog. Any individual training work is utterly

---

[607] 17 U.S.C. § 504(b). *See generally* Frank Music Corp. v. Metro-Goldwyn-Mayer, Inc., 772 F.2d 505 (9th Cir. 1985) (performing apportionment calculation).

[608] *See supra* note 119 (discussing the challenges of attribution of generations to specific training-data examples).

[609] *On Davis*, 246 F.3d at 160.

[610] *Id.*

insignificant on the scale of the whole system.[611] A plaintiff who shows only that their work was included in the training dataset has not carried their burden to show that any of the resulting profits were attributable to infringement of their work.[612]

This point demonstrates the crucial importance of *mass* copyright litigation against the service hosts of and other participants in generative-AI systems. The answer may well be different if the plaintiff or plaintiffs own a large fraction of the works used as training data. Although individual apportionment may remain a difficult problem, it is much easier to show that the model's value collectively derives from the works that have been infringed. This is one reason why so many of the current lawsuits against generative-AI companies have been brought as putative class actions.[613] Getty's lawsuit against Stability AI is not a class action, but Getty controls the copyright to a large number of works in Stable Diffusion's training dataset.[614] The *New York Times* has alleged that OpenAI trained more extensively on its "higher quality" articles compared to other sources of training data.[615]

2. Statutory Damages

Instead of recovering actual damages and/or profits, a successful copyright plaintiff may elect to recover statutory damages instead.[616] This will typically be an appealing option. First, the plaintiff can submit both theories to the court, see which one results in a larger award, and then choose that one.[617] Second, the amount of statutory damages is fixed in the statute. The base range is $750 to $30,000, "as the court considers just."[618] This amount can be decreased to $200 for an "innocent" infringer who "was not aware and had no reason to believe that his or her acts constituted an infringement of copyright,"[619] but this defense is not

[611] However, as we note above, some training data examples may have outsized influence on generations. *See generally* Koh & Liang, *supra* note 119; Akyurek, Bolukbasi, Liu et al., *supra* note 119; Grosse, Bae, Anil et al., *supra* note 119. (discussing influence functions).
[612] *See supra* note 119 and accompanying text; *supra* note 212 and accompanying text.
[613] *E.g.,* Complaint, Kadrey v. Meta Platforms, Inc., No. 3:23-cv-03417 (N.D. Cal. July 7, 2023); Complaint, Doe 1 v. GitHub, Inc., No. 4:22-cv-06823 (N.D. Cal. Nov. 3, 2022); Complaint, Anderson v. Stability AI, Ltd., No. 3:23-cv-00201 (N.D. Cal. Jan. 13, 2023) (Doc. No. 1); Complaint, Tremblay v. OpenAI, Inc., No. 3:23-cv-03223 (N.D. Cal. June 28, 2023).
[614] Complaint, Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135 (D. Del. Feb. 3, 2023).
[615] Complaint ¶ 90, N.Y. Times Co. v. Microsoft, No. 2:24-cv-00711 (C.D. Cal. Dec. 27, 2023).
[616] 17 U.S.C. § 504(c)(1).
[617] Curet-Velazquez v. ACEMLA de P.R., Inc., 656 F.3d 47, 57–58 (1st Cir. 2011).
[618] 17 U.S.C. § 504(c)(1).
[619] *Id.* § 504(c)(2).

available for works that were published with proper notice of copyright.[620] The amount can also be increased up to $150,000 when the "infringement was committed willfully."[621] Willful infringement consists either of actual knowledge or reckless disregard of infringement;[622] a defendant who has a reasonable and good-faith belief that their conduct is non-infringing is not a willful infringer.[623] Under these ranges, an individual statutory-damage award could be a serious threat to an individual user, a moderate nuisance to a small company, or an insignificant bit of background noise to an OpenAI or a Google.

Importantly, statutory damages are awarded *per work* infringed, regardless of how extensively each work was used. Again, the impact is clearest in mass copyright litigation. Statutory damages are a potentially existential threat to models trained on billions of works (and to the datasets that feed them and the services that incorporate them). Even without a finding of willfulness, the statutory damages for a billion infringed works could be as high as in the trillions of dollars—an impact that is no more survivable than the Chicxulub asteroid. Even at the minimum award for innocent infringement, the statutory damages for ten million infringed works would come to two hundred million dollars.[624]

One factor limiting statutory damage awards is that statutory damages are only available when the copyright owner registered the work with the Copyright Office before the infringement commenced.[625] This provision is designed to encourage authors to register their works promptly. It has the effect of making some generative-AI systems more vulnerable to copyright lawsuits than others. Books are typically registered as part of the publication process, so an LLM trained on hundreds of thousands of books could face hundreds of thousands of statutory-damage awards But many works of visual art and many websites are not registered unless and until the copyright owner needs to file a copyright lawsuit.[626] A model trained on a web scrape, then, may face a patchwork of statutory damage awards only for a small fraction of the works it was trained on. Differences in available damages based on the timing of registration may make it harder to assemble a plaintiff class with sufficiently common interests.[627]

---

[620] 17 U.S.C. § 401(d).
[621] 17 U.S.C. § 504(c)(2).
[622] Erickson Prods., Inc. v. Kast, 921 F.3d 822, 833 (9th Cir. 2019).
[623] VHT, Inc. v. Zillow Grp., Inc., 918 F.3d 723, 748–49 (9th Cir. 2019).
[624] This sum is still high enough that it might deter a court from finding infringement against a smaller defendant that merely used a model someone else had trained.
[625] 17 U.S.C. § 412(2).
[626] Registration is a prerequisite to suit. § 411(a); Fourth Est. Pub. Corp v. Wall-St.com, LLC, 139 S. Ct. 881 (2019).
[627] *See* Fed. R. Civ. P. 23(a)(2) (requiring "questions of law or fact common to the class"). The registration requirement cannot be circumvented through the use of a class action. *See* Reed Elsevier, Inc. v. Muchnick, 559 U.S. 154 (2010).

3. Attorney's Fees

Another remedy for copyright infringement is that a court may award "full costs" and "a reasonable attorney's fee to the prevailing party."[628] Costs are small potatoes; they include various court fees, printing fees, and required payments to the court.[629] But attorney's fees are a bigger deal, precisely because the expense of litigating a copyright case can be so high. Under the usual "American Rule" (so called because it is followed in the United States but not in many other countries), each party pays its own lawyers and decides how much the case is worth to them.[630] The Copyright Act's fee-shifting provision is one of a few exceptions to the American Rule. It provides an incentive to parties to bring meritorious cases—or to defend against unmeritorious ones—that would otherwise be financially unreasonable to pursue.[631] Like statutory damages, attorney's fees are only available for works that were registered before the infringement.[632]

While statutory damages are most important in mass litigation, the reverse is true of attorney's fees. A million dollars of expenses to litigate a class action with a hundred-million-dollar damage award is not the biggest deal. A fee award is a nice bonus, but it is not necessary to bring the suit in the first place. But a million dollars of expenses to litigate an individual claim leading to a $1,000 statutory damage award is completely unreasonable. Without an attorney's fee award, the lawyers involved could make more on a per-hour basis by busking on the subway.

Attorney's fees can also have a significant deterrent effect.[633] Because they are uncapped, a plaintiff can run up the total award a defendant faces. Indeed, the harder a defendant fights, the higher the plaintiff's attorney's fees will be. Along with statutory damages, attorney's fees can be used to coerce settlements from defendants who may have a strong defense on the merits.[634] Even though the defendant might be able to receive a fee award if they win—the fee-shifting rule is symmetrical[635]—they cannot run the risk of paying a massive fee award if they lose. This settlement pressure will be strongest against smaller and more risk-averse defendants: end users rather than well-capitalized AI companies, which can better absorb the cost of a fee shift. This difference helps to explain why

---

[628] 17 U.S.C. § 505.
[629] *See* Rimini St., Inc. v. Oracle USA, Inc., 139 S.Ct. 873 (2019) (interpreting "full costs").
[630] Fogerty v. Fantasy, Inc., 510 US 517, 533–34 (1994).
[631] *Id.* at 524.
[632] 17 U.S.C. § 412.
[633] *See generally* Pamela Samuelson & Tara Wheatland, *Statutory Damages in Copyright Law: A Remedy in Need of Reform*, 51 WM. & MARY L. REV. 439 (2009); Talha Syed & Oren Bracha, *The Wrongs of Copyright's Statutory Damages*, 98 TEX. L. REV. 1219 (2020).
[634] *See, e.g.,* Mitch Stoltz, *Collateral Damages: Why Congress Needs To Fix Copyright Law's Civil Penalties*, ELEC. FRONTIER FOUND. (July 24, 2014), https://www.eff.org/wp/collateral-damages-why-congress-needs-fix-copyright-laws-civil-penalties.
[635] *Fogerty*, 510 U.S. 517.

several generative-AI companies have offered to indemnify their users against the copyright risks of using their systems.[636]

4. Injunctions

A court may "grant temporary and final injunctions on such terms as it may deem reasonable to prevent or restrain infringement of a copyright."[637] An injunction is a court order commanding a person to take (or to avoid taking) some action. A party who fails to comply with an injunction can be punished for contempt of court with sanctions that include escalating fines and even imprisonment.

An injunction is an equitable remedy; a plaintiff is not automatically entitled to one.[638] Instead, a plaintiff seeking an injunction must show:

> (1) that it has suffered an irreparable injury; (2) that remedies available at law, such as monetary damages, are inadequate to compensate for that injury; (3) that, considering the balance of hardships between the plaintiff and defendant, a remedy in equity is warranted; and (4) that the public interest would not be disserved by a permanent injunction.[639]

The first two factors are redundant; they mean exactly the same thing.[640] A damages award in a copyright case is inadequate when damages are hard to calculate. For all of the reasons discussed above, this will frequently be the case in generative-AI cases. Thus, most of the weight will fall on the third and fourth factors. The degree to which hardships fall on a defendant that provides generative-AI models or systems, and on third-party users, will depend substantially on the balance of infringing and noninfringing uses. An injunction is more appropriate against a system that (a court sees as) "good for nothing else

---

[636] Brad Smith, *Microsoft Announces New Copilot Copyright Commitment for Customers*, MICROSOFT (Sept. 7, 2023), https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/; Bridget Johnston, *Introducing Indemnification for AI-Generated Images: An Industry First*, SHUTTER- STOCK (July 11, 2023), https://www.shutterstock.com/blog/ai-generated-imagesindemnification; Adobe, *Firefly Legal FAQs—Enterprise Customers* §§ 10–14 (June 12, 2023), https://www.adobe.com/content/dam/dx/us/en/products/sensei/sensei-genai/firefly-enterprise/Firefly_Legal_FAQs_Enterprise_Customers.pdf.

[637] 17 U.S.C. § 502(a). We will discuss only permanent injunctions issued after a finding of infringement. Preliminary injunctions issued during the course of a lawsuit may be important for parties and litigators, but our focus is on the longer term.

[638] eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388, 392–93 (2006).

[639] *Id.* at 391.

[640] Douglas Laycock, *The Death of the Irreparable Injury Rule*, 103 HARV. L. REV. 687, 694 (1990).

but infringement,"[641] and less appropriate against one that is also "capable of substantial noninfringing uses."[642] (As these quotes suggest, there is substantial overlap between the substantive tests for infringement and the test for a permanent injunction.)

Another factor weighing against generative-AI injunctions is the First Amendment interests of users and developers.[643] There is often a speech interest in using the speech of others verbatim;[644] these First Amendment interests are even stronger for novel generations. In individual cases against specific generations, users' speech rights are protected by the "traditional First Amendment safeguards" of fair use, particularly transformative fair use.[645] But an injunction against the use of a model or service can prevent these generations from being created; this is a speech harm too. So, when a model is used to create expressive and noninfringing generations, there is a powerful argument that a court should not enjoin it in a way that would prevent these noninfringing uses.

And so, we come to one of the most important features of an injunction: a court's ability to craft its specific terms. A court could enjoin the use of a model *entirely*, preventing the defendant from using it for any purpose. But a court could also enjoin the use of a model *to create infringing generations*, leaving it up to the defendant to implement appropriate content filters.[646] This type of injunction puts sharper teeth into the defendant's obligations, because the consequences for failing to comply with an injunction are swifter and more severe than for committing copyright infringement. Unfortunately for defendants (and for courts considering enjoining them), it is harder to "separat[e] the fair use sheep from the infringing goats" in a generative-AI system than it is on a content-hosting service

[641] Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 932 (2005).
[642] *Id.* at 391.
[643] Mark A. Lemley & Eugene Volokh, *Freedom of Speech and Injunctions in Intellectual Property Cases*, 48 DUKE L.J. 147 (1998).
[644] *See* Rebecca Tushnet, *Copy This Essay: How Fair Use Doctrine Harms Free Speech and How Copying Serves It*, 114 YALE L.J. 535 (2004).
[645] Eldred v. Ashcroft, 537 U.S. 186, 219–20 (2003).
[646] For example, Copilot offers an option to check "code suggestions with their surrounding code of about 150 characters against public code on GitHub" and propose a different suggestion if the filter is triggered. *See Configuring GitHub Copilot in your environment*, *supra* note 239. Unfortunately, while helpful, content filters like Copilot's are not enough by themselves to prevent the generation of potentially infringing content. For example, Copilot's filter would not be triggered if the generated code suggestion matched 149 characters of public code—which is long enough to at least raise copyright concerns. *See* Justin Hughes, *Size Matters (Or Should) in Copyright Law* , 74 FORDHAM L. REV. 575 (2005) (discussing copyright protection of "microworks"). *See generally* Daphne Ippolito, Florian Tramèr, Milad Nasr et al., Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy (2023) (unpublished manuscript), https://arxiv.org/abs/2210.17546 (discussing how verbatim output filters are necessarily incomplete).

like YouTube.[647] Even for a defendant with a list of works to avoid, this type of filtering is a difficult and unsolved technical problem.[648]

5. Destruction

Another equitable remedy is that the court may order "the destruction or other reasonable disposition of all [infringing] copies."[649] This is like a more severe version of an injunction, one that takes it out of the defendant's power to commit further infringements by taking away their copies. To the extent that a model is treated as an infringing copy, the destruction remedy does not add very much to a permanent injunction except for irreversibility. Actually deleting a model—as opposed to putting in in storage for future use if and when the law changes or copyright owners negotiate a license to allow it to be used—is an exceptionally harsh remedy that effectively means throwing away all of the compute used to train the model.

But there is a twist. As Elizabeth Joh observes,[650] the destruction remedy covers not just infringing copies but also "all plates, molds, matrices, masters, tapes, film negatives, or other articles by means of which such copies or phonorecords may be reproduced."[651] Even if a model is not itself treated as an infringing copy,[652] if it is capable of producing infringing generations, it might be an "article[] by means of which" infringing copies "*may* be reproduced."[653] The courts have not restricted this remedy to items that themselves infringe or have been used to infringe.[654] Instead, they have allowed it to be used against dual-use technologies like computers and manufacturing equipment that can be used both to infringe and for noninfringing purposes.[655] Thus, the destruction remedy could reach not just models with multiple uses, but also the non-model portions of a generative-AI service. For example, a court could order the destruction of a style-transfer system that allows users to regenerate one image using the artistic style of another, on the theory that a user could prompt it with a copyrighted image and generate an infringing derivative work. Such an order would raise even more severe free-expression concerns.

---

[647] Campbell v. Acuff-Rose Music, 510 U.S. 569, 586 (1994).
[648] *See supra* note 412 and accompanying text.
[649] 17 U.S.C. § 503(b). *See generally* Elizabeth E. Joh, Equitable Legal Remedies and the Existential Threat to Generative AI (Aug. 27, 2023) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4553431. As with injunctions, there is also a preliminary version of destruction: a court may order the impoundment of infringing copies during the course of the litigation. 17 U.S.C. § 503(a)(1).
[650] *See* Joh, *supra* note 652.
[651] 17 U.S.C. § 503(b).
[652] *See* Cooper & Grimmelmann *supra* note 289.
[653] 17 U.S.C. § 503(b) (emphasis added).
[654] Mahan v. Roc Nation, LLC 720 Fed. Appx. 55 (2d Cir. 2018).
[655] Anne-Marie Carstens, *Copyright's Deprivations*, 96 WASH. L. REV. 1275 (2021).

L. Copyright Management Information

Section 1202 of the Copyright Act, enacted like section 512 as part of the DMCA, deals with "copyright management information . . . conveyed in connection with copies . . . of a work" (CMI).[656] Types of CMI include a work's title, author, copyright owner, performers, and licensing information.[657] One prong of section 1202 prohibits providing "false" CMI;[658] another prohibits "remov[ing] or alter[ing]" CMI.[659]

The legislative history of section 1202 (and its passage as part of the *Digital* Millennium Copyright Act) suggests that it was designed to work in tandem with section 1201, which prohibits disabling digital rights management systems that protect copyrighted works.[660] Where section 1201 guards the parts of the system that directly control access, section 1202 ensures that the metadata and watermarks attached to works are accurate and intact.[661]

But the language of section 1202 is not limited to digital metadata. Unlike the World Intellectual Property Organization Copyright Treaty, which applies to "*electronic* rights management information,"[662] section 1202's text contains no such limitation. As a result, courts have held that section 1202 can be violated when a magazine photo is reproduced online without the photographer's name from a "gutter credit" that appeared alongside it in print.[663]

Under these precedents, the assembly of works into datasets and the training of a model could result in the "remov[al]" of CMI through a similar decontextualization. Consider a diffusion-based model trained on one of the LAION image datasets. The dataset itself consists of URL links to images where they appear in context on webpages,[664] often with author, title, and copyright-owner credits of the type that qualify as protected CMI. This by itself is neither falsification, removal, nor alteration. But when the images *by themselves* are downloaded, the attached CMI is stripped in the same way as in the magazine cases. Training and generation do not repair the linkage once it has been severed; if a model outputs a similar image, it will not bear the original CMI.

---

[656] 17 U.S.C. 1202(c) (emphasis added).
[657] *Id.*
[658] *Id.* § 1202(a).
[659] *Id.* § 1202(b).
[660] 17 U.S.C. § 1201; Severine Dusollier, *Some Reflections on Copyright Management Information and Moral Rights*, 25 COLUM. J.L. & ARTS 377 (2003).
[661] IQ Grp., Ltd. v. Wiesner Pub., 409 F. Supp. 2d 587, 593–97 (D.N.J. 2006); Textile Secrets Int'l, Inc. v. Ya-Ya Brand Inc., 524 F. Supp. 2d 1184, 1196–99 (C.D. Cal. 2007).
[662] WIPO Copyright Treaty, art. 12(1)(i), 1996.
[663] Murphy v. Millennium Radio Grp. LLC, 650 F.3d 295 (3d Cir. 2011); Mango v. BuzzFeed, 970 F.3d 167 (2d Cir. 2020).
[664] *See supra* Part I.B.3.

Getty Images's complaint against Stability AI presents additional theories of section 1202 violation based on the Stable Diffusion models' treatment of the Getty watermarks on the images in its library.[665] First, to the extent that the training process learns features of training images without the watermark, Getty alleges removal and alteration of CMI.[666] Second, Getty shows that Stable Diffusion sometimes produces generations that include distorted versions of the watermark.[667] This, Getty argues, constitutes "false" CMI within the meaning of section 1202.[668]

The more serious doctrinal obstacle to section 1202 claims is that they require a nexus to copyright infringement. Falsification of CMI must be done "knowingly and with the intent to induce, enable, facilitate, or conceal infringement" to create liability,[669] and removal or alteration must be done "intentionally . . . knowing, or . . . having reasonable grounds to know, that it will induce, enable, facilitate, or conceal an infringement."[670] Most defendants in generative-AI cases will have the required intent to remove or alter the CMI. A developer training on the LAION dataset can hardly fail to know that the training process discards any information about the images on the webpages they came from.

Instead, it is not clear that a defendant's treatment of CMI at any stage of the generative-AI supply chain is intended to facilitate or conceal copyright infringement in any cases where copyright infringement would not already attach to the defendant. Getty objects that attaching a "modified version of the Getty Images watermark to bizarre or grotesque synthetic imagery,"[671] will harm its reputation. But that is a concern that sounds in trademark, not copyright.[672] Indeed, the "grotesque" nature of the images Getty includes in its complaint, if anything, cuts against infringement, by suggesting that the images are not suitable for any valuable purpose, let alone competing with Getty. The decontextualization of the training process might be said to help "conceal" infringement, but again the infringement itself is likely to be separately actionable.

The real bite of the CMI claims may be remedial. A court is entitled to award statutory damages of $2,500 to $25,000 "for each violation of section 1202."[673] The liability is *per violation* rather than *per work* (as with ordinary copyright infringement). In theory, then, a defendant could face separate section 1202 liability for each variation of a dataset or model it creates, or each output bearing

[665] Complaint, Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135 (D. Del. Feb. 3, 2023).
[666] *Id.* ¶¶ 81–86.
[667] *Id.* ¶¶ 59–60.
[668] *Id.* ¶¶ 74–80.
[669] 17 U.S.C. § 1202(a).
[670] *Id.* § 1202(b).
[671] Complaint ¶ 59, Getty Images (US), Inc. v. Stability AI, Inc., No. 1:23-cv-00135.
[672] *See id.* ¶¶ 87–99 (bringing claim for trademark infringement).
[673] 17 U.S.C. § 1203(c)(3)(B).

a watermark. On the other hand, while there is a $200 floor for ordinary copyright statutory damages in cases of innocent infringement,[674] a is entitled "in its discretion" to reduce section 1202 statutory damages or remit them entirely in cases of innocent violations.[675]

M. Right of Publicity

A related but non-copyright form of IP is the right of publicity.[676] The right generally protects an individual's persona against commercial appropriation by others. Unlike copyright, which is almost entirely created by federal law, the right of publicity is almost entirely created by state law. As a result, its details vary substantially from state to state, including whether a state affords a right of publicity at all and, if it does, whether the right is regarded as a privacy or property right or both, what kinds of conduct it protects against, the scope of newsworthiness or expressive-use defenses, and whether and how long the right lasts after the subject's death.[677] As a result, the following summary is a broad overview, rather than a specific analysis of any state's (or each state's) law.

1. Overview of the Right of Publicity

A typical statement of the right of publicity's subject matter is that it covers an individual's "name, voice, signature, photograph, or likeness"[678]—the aspects of a person's persona that are broadly recognizable by others. (We will collectively refer to these as "identity.") Courts have interpreted recognizability broadly, holding that race-car driver Lothar Motschenbacher's right of publicity was infringed by a commercial showing his car,[679] singer Bette Midler's by a commercial featuring a sound-alike vocalist,[680] and game-show host Vanna White's by a print ad showing a robot next to the game board from *The Price is Right*.[681] The fact that an expressive work is recognizably *by* an author or artist does not mean that it implicates their right of publicity: a use must depict or summon up the person in the minds of viewers. Some states' rights of publicity terminate on death,[682] others last for a legislatively specified or judge-created

[674] 17 U.S.C. § 504(c)(2).
[675] 17 U.S.C. § 1203(c)(5)(A).
[676] *See generally* JENNIFER E. ROTHMAN, THE RIGHT OF PUBLICITY: PRIVACY REIMAGINED FOR A PUBLIC WORLD (2018) (providing a thorough history, analysis, and critique of the right of publicity as it exists in the United States today).
[677] *See generally* Jennifer E. Rothman, *Rothman's Roadmap to the Right of Publicity* (2024), https://rightofpublicityroadmap.com (providing a detailed analysis of every state's right-of-publicity laws).
[678] Cal. Civ. Code § 3344(a).
[679] Motschenbacher v. R.J. Reynolds Tobacco Co., 498 F.2d 821 (9th Cir. 1974).
[680] Midler v. Ford Motor Co., 849 F.2d 460 (9th cir. 1988).
[681] White v. Samsung Elecs. Am., 971 F.2d 1395 (9th Cir. 1992).
[682] Hagen v. Dahmer, 38 U.S.P.Q.2d 1146, 1995 WL 822644 (E.D. Wis. 1995).

term,[683] and Tennessee allows the right to continue indefinitely as long as it is being commercially exploited.[684]

The right of publicity also applies only to *commercial* uses. One core use is endorsement: using a person's identity to sell things.[685] Another is merchandising: selling basketball jerseys with Steph Curry's name and number on them, or Dolly Parton Funko Pops. And a third is what Eric Johnson calls "virtual impressment":[686] digitally recreating a person to perform in movies, video games, songs, and other audio or audiovisual media.[687] Broadly speaking, endorsement issues can often be avoided with sufficient disclaimers to establish that the person has not endorsed the product in question or consented to appear in the advertising; but, to the extent that a plaintiff has a valid claim based on merchandising or virtual impressment, disclaimers will not save the defendant.

Although the right of publicity does not have a copyright-style, general-purpose, fair-use defense, some courts have recognized a narrower copyright-style, transformative-use defense when a person's likeness "is so transformed that it has become primarily the defendant's own expression rather than the celebrity's likeness."[688] Some state statutes explicitly carve out uses affected with a strong public interest, such as a California's exception for "news, public affairs, or sports broadcast or account, or any political campaign."[689] And sometimes sufficiently expressive uses are excluded entirely, as in California's exception for "fictional or nonfictional entertainment, or a dramatic, literary, or musical work" after the person's death.[690]

The right of publicity has a close and complicated relationship with copyright. First, like all state-created IP rights, it is subject to federal preemption. The Copyright Act provides that "all legal or equitable rights that are equivalent to any of the exclusive rights within the general scope of copyright . . . are governed exclusively" by federal copyright.[691] To avoid preemption, a right of publicity claim must either protect different subject matter than copyright (e.g., a person's appearance is not a fixed work of authorship) or include an additional

[683] Hebrew Univ. of Jerusalem v. Gen. Motors, 903 F. Supp. 2d 932, 939–40 (C.D. Cal. 2012).
[684] 47 Tenn. Code. §§ 47–25-1104.
[685] Cal. Civ. Code § 3344(a) ("for purposes of advertising or selling, or soliciting purchases of, products, merchandise, goods or services"). *Motschenbacher*, *Midler*, and *White* are all endorsement cases.
[686] Eric E. Johnson, *Disentangling the Right of Publicity*, 111 NW. U. L. REV. 891, 934–35 (2017).
[687] Hart v. Elec. Arts, Inc., 717 F.3d 141 (3d Cir. 2013).
[688] Comedy III Prods. v. Gary Saderup, Inc., 21 P.3d 797, 141 (Cal. 2001).
[689] Cal. Civ. Code § 3344(d).
[690] Cal. Civ. Code § 3344.1(a)(2). Put another way, California's statutory right of publicity protects against virtual impressment of the living, but not of the dead.
[691] 17 U.S.C. § 301(a).

element not required for copyright infringement (e.g., using the plaintiff's likeness as advertising or promotion to sell another product).[692]

When the basis of a right-of-publicity claim is the distribution of a work either depicting a person or created by the person or both, courts frequently treat the right of publicity as having merged into the copyright in the work: they cannot further restrict the copyright owner's ordinary exploitation of the work. For example, in *Laws v. Sony Music Entertainment, Inc.*, the plaintiff Debra Laws's vocals from "Very Special" were used as a sample on Jennifer Lopez and L.L. Cool J.'s "All I Have."[693] The defendants had a copyright license from Laws's record label, but not a right of publicity license from Laws. The Ninth Circuit held that Laws's claim was preempted.[694] The case would have been different if the sample had been used for an advertisement rather than a new track; that would have been an extra element.[695] Difficult issues sometimes arise when footage or other works created for one project are reused in a related but different context, as in *Facenda v. NFL Films, Inc.*, where the NFL reused voice-over lines recorded by John Facenda as documentary narration for a 22-minute promotion for a video game, and the Third Circuit held that his estate's right of publicity claim was not preempted.[696]

2. Incorporation and Advertising

The most famous generative-AI right-of-publicity lawsuit is both a legal non-starter and not actually about generative AI. In January 2024, the *Dudesy* podcast posted an hour-long episode titled "George Carlin: I'm Glad I'm Dead," which the podcast hosts claimed to feature both a script and audio that had been trained to imitate the late comedian George Carlin.[697] Carlin's estate sued, making claims under California's statutory and common-law rights of publicity.[698] But the statutory claim is a loser because California's postmortem right of publicity, as noted above, expressly excludes audiovisual entertainment, and the common-law claim is a loser because the courts have held that California's common-law right terminates at death.[699] Even more fundamentally, *Dudesy*'s hosts promptly

[692] *See generally* Jennifer E. Rothman, *Copyright Preemption and the Right of Publicity*, 36 U.C. DAVIS L. REV. 199 (2002).
[693] Laws v. Sony Music Ent., Inc., 448 F.3d 1134 (9th Cir. 2006).
[694] *Id.* at 1145.
[695] *Id.* at 1141–42; *cf.* Downing v. Abercrombie & Fitch, 265 F.3d 994 (9th Cir. 2001) (photographs used as advertisements).
[696] Facenda v. NFL Films, Inc., 542 F.3d 1007 (3d Cir. 2008).
[697] Christopher Kuo, *George Carlin's Estate Sues Podcasters Over A.I. Episode*, N.Y. TIMES (Jan. 29, 2024), https://www.nytimes.com/2024/01/26/arts/carlin-lawsuit-ai-podcast-copyright.html.
[698] Main Sequence, Ltd. v. Dudesy, LLC, No. 2:24-cv-00711 (C.D. Cal.).
[699] Lugosi v. Universal Pictures, 603 P.2d 425 (Cal. 1979).

admitted that the episode was entirely human-written.[700] The generative-AI veneer was just a publicity stunt.

The Carlin lawsuit, near-miss though it is, helpfully illustrates two ways in which the right of publicity can apply to generative AI. First, a technical artifact (a dataset, model, system, or generation) could *incorporate* a person's identity. "I'm Glad I'm Dead" imitated Carlin's distinctive voice. Second, a technical artifact could be *advertised* using a person's identity. "I'm Glad I'm Dead" was promoted using Carlin's name. Incorporation raises merchandising and virtual-impressment issues; advertising raises endorsement issues.

In most cases, generative AI will raise distinctive right of publicity issues only to the extent that it incorporates a person's identity. This is for two reasons. First, when it is legal to *sell* a product incorporating a person's identity or creative output, it is also generally legal to *promote* the product by truthfully describing the person's relationship to it.[701] Second, using a person's identity to sell generative-AI material that does not otherwise relate to them is a garden-variety case under the endorsement prong of the right of publicity. Whether it infringes on Salvador Dalí's right of publicity to name a family of image systems "DALL·E" has little to do with the fact that it is a generative-AI system. Almost the same issues would arise with calling a line of paintbrushes "DALL·E".

3. Incorporation in the Generative-AI Supply Chain

Some AI generations are already being used for blatant right of publicity violations. Ads featuring a cloned version of Taylor Swift's voice have been used in fake giveaways for Le Creuset cookware;[702] a deepfake video of Tom Hanks has been used to advertise a dental plan.[703] Of course, fake celebrity endorsements are nothing new. The difference between an ad using an (actual) photograph of a celebrity and an ad using a (generated) video of them is a difference in degree, not in kind. Generative AI may make the deception more convincing by forging an explicit endorsement, but what makes these uses actionable is fundamentally the lack of permission. So, the right of publicity violation is more about how the media is used, not how it is generated.

From the perspective of the system that is used to generate the media, or any other actors further upstream in the generative-AI supply chain, this is ultimately a secondary-liability question that is quite similar to the secondary-liability

---

[700] Kuo, *supra* note 700.

[701] Armstrong v. Eagle Rock Ent., Inc., 655 F. Supp. 2d 779 (E.D. Mich 2009) (defendant could use photograph of plaintiff on the cover and liner notes of a DVD concert video).

[702] Tiffany Hsu & Yiwen Lu, *No, That's Not Taylor Swift Peddling Le Creuset Cookware*, N.Y. TIMES (Jan. 9, 2024), https://www.nytimes.com/2024/01/09/technology/taylor-swift-le-creuset-ai-deepfake.html.

[703] Derrick Bryson Taylor, *Tom Hanks Warns of Dental Ad Using A.I. Version of Him*, N.Y. TIMES (Oct. 2, 2023), https://www.nytimes.com/2023/10/02/technology/tom-hanks-aidental-video.html.

question for copyright.[704] The law of secondary liability in right of publicity is both less developed (because the cases are fewer) and more fragmented (because the sources of law are more numerous) than in copyright.[705] It is not obvious that there are any material differences between the two.

That said, the statutory safe harbor that is potentially applicable to the right of publicity is both less and more complicated than safe harbors in copyright. On the one hand, the right of publicity is not subject to the safe harbor notice-and-takedown regime of section 512, which applies only to copyright. On the other, a different immunity, "section 230," protects Internet immediacies from liability from third-party information provided by another."[706] Section 230 has an exception for "intellectual property,"[707] and courts are split on whether this includes the state-created right of publicity or not.[708] And if section 230 does apply to the right of publicity, there is deep disagreement (and an utter absence of caselaw) on how it applies to generative AI because it is unsettled whether and when AI generations should be regarded as third-party content.[709]

A different theory of a right-of-publicity variation is the use of a generative-AI system to produce outputs in the style of particular artists or authors. (See Figure 19.) Styling these claims as right-of-publicity violations rather than under copyright[710] introduces a few twists. Most fundamentally, there is copyright preemption. To the extent that these claims mirror copyright claims based on the imitation of one's style as embodied in fixed creative works—e.g., a photograph in the style of Cindy Sherman—they are preempted unless there is some extra element. One candidate for such an element is the prompt. At least as to commercial services, there is an argument that if a service produces a generation in response to the prompt "a photograph in the style of cindy sherman", then this constitutes a use by the service of Sherman's name. The doctrinal hurdle here,

[704] *See supra* Part II.F.

[705] *See* Perfect 10, Inc. v. Cybernet Ventures, Inc., 213 F. Supp. 2d 1146, 1183–87 (C.D. Cal. 2002); J. THOMAS MCCARTHY, THE RIGHTS OF PUBLICITY AND PRIVACY § 3:17 to 3:20 (2d ed. 2023) (surveying the limited caselaw).

[706] 47 U.S.C. § 230(c)(1).

[707] 47 U.S.C. § 230(e)(2).

[708] Perfect 10, Inc. v. CCBill LLC, 488 F.3d 1102, 1118–19 (9th Cir. 2007) (section 230 immunity applies to the right of publicity) *with* Hepp v. Facebook, 14 F.4th 204 (3d Cir. 2021) (no it does not).

[709] *See generally* Peter J. Benson & Valerie C. Brannon, Congressional Research Service (Congressional Research Service Legal Sidebar LSB11097 Dec. 28, 2023) (surveying caselaw and commentary).

[710] *See supra* Part II.C.2.f (discussing artistic style and similarity).

however, is that it is not clear that the user's prompt should be attributed to the service, which is not using Sherman's name to advertise.[711]

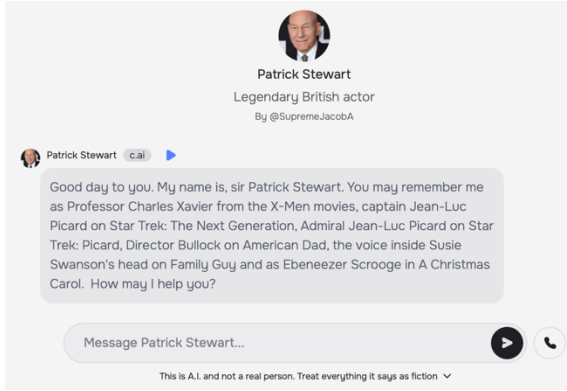


Figure 19: Screenshot of the character.ai platform, where a user can chat with an AI "character" version of Sir Patrick Stewart.

A stronger version of this theory is that services based on models which have been developed to specifically imitate an artist's style or person's appearance or voice are more clearly selling that person's identity. One way of framing this situation is that a commercial model provider is directly violating Drake's identity by using Drake's name to sell its models (of Drake): a form of advertising. Another way is to say that it is the models that are the problem: a form of merchandising. And a third is to say that the model is a kind of toolkit for anyone to engage in virtual impressment of Drake, so the seller is engaged in contributory virtual impressment.

All of these theories are subject to the usual right-of-publicity defenses. The transformative-use argument is particularly strong, for the same reasons it is strong in copyright. Indeed, because the earlier stages of the generative-AI supply chain generally cannot be the basis of right-of-publicity claims—they do not by themselves involve recognizable uses of a person's identity—the transformative-use defense is needed only in the later stages, where the transformation is the most pronounced. And the general public-interest defense will apply so clearly to some generations (it is not hard to imagine news programs making generative-AI illustrations), and others will be clearly non-infringing private non-commercial uses, so most systems will have substantial noninfringing uses.

[711] To the extent that a service does (or does not) analyze prompts to detect problematic or prohibited requests, there is a question of whether this analysis constitutes use sufficient to trigger the right of publicity or to avoid preemption. A similar issue will arise under other use-based bodies of law, such as trademark.

N. Hot News Misappropriation

One final relevant copyright-like form of IP liability is hot news misappropriation. The common-law cause of action for misappropriation has a long history; it is a species of unfair competition law, which prohibits businesses from "reaping the fruits" of their competitors' investments.[712] Its most famous statement is in the 1918 Supreme Court case *International News Service v. Associated Press*.[713] The Associated Press (AP) and the International News Service (INS) were competing wire services that reported and transmitted news stories to their member newspapers. AP alleged that INS was copying news stories from early editions of AP papers so that INS papers could report on them in their later editions.

It is important to note why this practice was not copyright infringement and is not to this day. The Associated Press could potentially have a copyright in the articles its employees wrote,[714] but the facts it reported were uncopyrightable. As Justice Brandeis wrote in dissent, "[T]he noblest of human productions—knowledge, truths ascertained, conceptions, and ideas—become, after voluntary communication to others, free as the air to common use."[715]

Justice Pitney's majority opinion, then, focused on the "novelty and freshness" of the news reported by the AP.[716] It held that the AP had a kind of "*quasi* property" as against competitors like the INS.[717] While it could not prevent readers and other members of the general public from freely discussing and writing about the news, it could prevent the INS from engaging in a systematic process of copying the news "precisely at the point where the profit is to be reaped"—that is, while the news was still fresh and there was value in being first to report it in a given newspaper market.[718]

The courts have held that the essential core of a hot news claim is freeriding on a competitor's costly production of information in a way that undermines the incentives to produce that information at all.[719] For example, in *National Basketball Ass'n v. Motorola*, the court explained that while it would be misappropriation for one real-time sports-score business to retransmit scores distributed by another, it was legal for the defendant to have its own reporters watch games to keep the scores updated.[720]

---

[712] Int'l News Serv. v. Associated Press, 248 U.S. 215, 241 (1918).
[713] *Id.*
[714] In addition, under the 1909 Copyright Act, copyright was too encumbered with formalities to provide the AP with effective relief. The 1976 Copyright Act, in which copyright attaches on fixation, overcomes this procedural barrier
[715] *Int'l News Serv.*, 248 U.S. at 250 (Brandeis, J., dissenting).
[716] *Id.* at 238.
[717] *Id.* at 236.
[718] *Id.* at 240.
[719] Nat'l Basketball Ass'n v. Motorola, 105 F.3d 841, 845 (2d Cir. 1997).
[720] *Id.*

For unrelated reasons, there is no longer a federal cause of action for misappropriation.[721] Instead, it is now governed entirely by state law, and as such it is subject to copyright preemption. The plaintiff must allege either that the information being copied does not fall within the general scope of copyright or that the cause of action contains an extra element.

The *New York Times* has brought hot news misappropriation claims against Microsoft and OpenAI, in addition to its copyright claims[722] The *Times* is a closer fit for misappropriation than many other generative-AI copyright plaintiffs, because it "gathers information, which often takes the form of time sensitive breaking news, for its content at a substantial cost."[723] It is a news organization, much like the AP.

Still, these claims are unlikely to succeed. The training and deployment of most generative-AI systems take place on such a drawn-out time scale that any breaking-news value in the training data will have been long since exhausted by the time anyone uses them. Hot news is ice cold six months later. (By contrast, INS papers reported news the same day as AP papers.) Mere competition with the *Times* for readership is likely not enough to generate a misappropriation claim that can survive preemption.

That said, however, some RAG or plugin systems might raise more serious hot news misappropriation issues. For example, Perplexity Pages creates "reports" by using a generative-AI model to summarize sources on a topic, including sources that the system pulls in from the web in response to a user's query.[724] Forbes alleged that Pages would closely paraphrase Forbes articles that Perplexity obtained by evading Forbes's paywall. This kind of conduct, if carried out at sufficient scale, is closer to the heartland of hot news misappropriation. This is misappropriation by a generative-AI *system* rather than by way of a generative-AI *model*.[725]

The *Times* also brings a claim that the defendants are misappropriating shopping recommendations from its Wirecutter subsite.[726] Here, the complaint emphasizes that the removal deprives the *Times* of affiliate revenue. But here too copyright preemption gives the *Times*'s theory of liability a difficult hill to climb. Wirecutter reviews are clearly fixed works of authorship,[727] and there is no

[721] In 1938, the Supreme Court held that federal courts must apply the law of the states in which they sit, so that "There is no federal general common law." Erie R. Co. v. Tompkins, 304 U.S. 64 (1938).
[722] Complaint, N.Y. Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023)..
[723] *Id.* ¶ 193.
[724] *See* Lopato, *supra* note 487.
[725] *See generally* Joseph A. Tomain, *First Amendment, Fourth Estate and Hot News: Misappropriation is Not a Solution to the Journalism Crisis*, 2012 MICH. ST. L. REV. 769.
[726] *Id.* ¶ 194.
[727] *See supra* Part II.A.1.

element here—creation at cost, duplication by competitors, value-capture by defendants—that is not also present in a typical copyright-infringement claim.

III. 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 from this radiating complexity.

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 the specific model or system in question. 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, etc. Courts will inevitably have to work through these details in numerous lawsuits, as they develop doctrines to distinguish among different types of models, systems, and uses. Some of those doctrines may produce clear general rules, but others will have to remain entangled with the different forms the supply chain can take.

Second, because of the complexity of *copyright law*, there is enormous play in the joints. In particular, substantial similarity, indirect infringement, fair use, and remedies 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 the many 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.

In this Part, we explore some of the ways that courts might try to use their discretion to apply copyright law to generative AI,[728] and then discuss some of the considerations that courts should keep in mind as they do.[729]

A. Possible Outcomes

Although the details of which generative-AI models and systems fall into which boxes may vary, there are a few boxes that courts may find it appealing to sort them into. In this section, we sketch a few of the possible copyright regimes that might result.

1. No Liability

First, courts might settle on a regime of no liability for released models, deployed services, and their users. Anything produced by a generative-AI system would be categorically legal, under a combination of no substantial similarity and fair use. The result would be that models and services would also be categorically

[728] *See infra* Part III.A.
[729] *See infra* Part III.B.

legal—there would be no primary liability for them to be indirectly liable for, and intermediate nonexpressive fair use would shield them in any event. Training datasets would also usually be legal as well (except perhaps in cases of blatant infringement like Books3).[730] 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 such an outcome might make sense for some generative-AI systems, it seems both unworkable and undesirable for others, including models and systems produced specifically to emulate the styles of particular creators, and systems that use retrieval augmented generation or plugins, which find matching works and reproduce them exactly or nearly exactly.[731] If all generative AI were categorically legal, then developers would plausibly start adding generative components to other systems in order to launder copyrighted works through them. The endpoint could be the effective collapse of copyright. On the assumption 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.

2. Liability for Generations Only

Second, courts could draw a line between generative-AI services and the users of those services.[732] In this regime, only generations would be treated as infringing, and then only when a user made some external use of them.[733] In this world, generative-AI systems would be creative tools like Photoshop.[734] 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

[730] Knibbs, *supra* note 557; Reisner, *supra* note 557; Complaint, Kadrey v. Meta Platforms, Inc., No. 3:23-cv-03417 (N.D. Cal. July 7, 2023).
[731] *See supra* note 212 and accompanying text (discussing retrieval augmented generation); *supra* note 487 (discussing OpenAI's plugins and Perplexity AI).
[732] The service/user distinction does not apply as clearly to models. A user might have received a model directly from its creator, or indirectly via one or more intermediaries, with or without modifications.
[733] 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.
[734] Sometimes literally so. *See Experience the Future of Photoshop With Generative Fill*, ADOBE (July 27, 2023), https://helpx.adobe.com/photoshop/using/generative-fill.html.

images from Instagram, even though Instagram is not liable for hosting those images.[735] And this is also precisely the type of situation that indemnification of users could help address.

The main difficulty with this regime would be policing against systems designed specifically for infringement. Something like the *Grokster* rule, carefully followed, might suffice. The providers of a service that was geared to produce infringing outputs could be held liable. So could the deployers of a model that had been trained or fine-tuned to optimize its effectiveness specifically for infringing uses. So could the curator of a dataset that included only or primarily infringing works, or a dataset that 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 specifically directed towards increasing the use of a system for infringement, with no substantial noninfringing purpose.

3. Notice and Removal

Third, courts could treat generative-AI services as generally legal in themselves but require them to respond to knowledge of specific infringements under a *Napster*-like rule.[736] One plausible doctrinal 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.

Regardless of which of these doctrinal routes a court were to take, 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. In many cases, it almost certainly does not.[737] 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

[735] *E.g.,* Sinclair v. Ziff Davis, LLC, 454 F. Supp. 3d 342 (S.D.N.Y. 2020).

[736] We are specifically talking about closed, hosted services here. Notice and removal of content related to an open-source model presents additional challenges. An open-source model developer can release new or changed models, but this does not recall or alter existing copies of prior models that already have been distributed to the public.

[737] *See supra* Part II.G.

may strike a court as intending to induce infringement, not making a good-faith effort to comply with an injunction, etc.

If courts do end up recreating a notice-and-takedown regime, they would likely settle on familiar elements: 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. These would probably not be notices directed to specific generations by named users, which would be difficult to detect and track. Instead, courts might require copyright owners to identify copyrighted works and then require that the generative-AI service operator prevent generations that are substantially similar to those works. Sometimes copyright owners might know which works to identify based on known generated outputs that are recognizably similar to suspected training-data inputs. But others might simply involve copyright owners handing over to service operators large catalogs of works to block, much as they currently do with ContentID on YouTube.

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."[738] 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 (potentially very) inexact match is much harder technically.

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.[739]

---

[738] 17 U.S.C. § 512(c)(3)(A)(iii).

[739] *See generally* Lee, Ippolito, Nystrom et al., *supra* note 431 (using hash-driven duplicate detection); Ippolito, Tramèr, Nasr et al., *supra* note 651 (discussing the drawbacks of exact-duplicate detection).

The challenges of implementing some analogue for removal for models[740] 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 a system prompt that instructs the model to respond in ways that reduce its propensity to infringe. But 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.[741] The model cannot stop anyone from fine-tuning it in an attempt to remove its guardrails.[742] Further, there is no simple analogue for takedown in generative-AI models. It remains an active and unsolved area of research to figure out how to remove a particular training example's influence from a model's parameters.[743] 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.[744]

Courts could respond to this difficulty in one of two ways. If they have sympathy for model trainers, they could apply the *Sony* rule and hold that it is not infringement to distribute a trained model as a set of parameters (as Stability AI's releases have been). The fact that the model is used by others for infringing purposes would be counterbalanced by the substantial noninfringing uses, leading to immunity under *Sony*. This might not always be an attractive business model, because it might be hard for buyers to monetize these models and because of the ease of copying and further redistributing the models—but it could at least exist legally. And truly open-source models would generally be allowed.

But if courts had less sympathy for model trainers, they might hold that the difficulty of complying with removal notices is not an excuse. On this view, the model trainer chose to create a model that could be used for substantial infringement, and to hopelessly commingle infringing and noninfringing material. If so, then it would generally not be legal to distribute a model that was trained on unlicensed works and had infringing outputs, at least once those works they were based on were pointed out. (This would perhaps foreclose released models like those in Meta's Llama-model family.[745]) It would be legal to train a model, but the trainer would need to take care that the model was only deployed in a safe

[740] *See* Cooper, Choquette-Choo, Bogen, Jagielski, Filippova, Liu et al., *supra* note 405 (discussing the challenges and fundamental limitations of "machine unlearning" as a solution to this problem).
[741] *Id.*
[742] *See* Xiangyu Qi, Yi Zeng, Tinghao Xie et al., Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To! (2023) (unpublished manuscript), https://arxiv.org/abs/2310.03693; Nasr, Rando, Carlini et al., *supra* note 404.
[743] *See, e.g.,* Meng, Bau, Andonian & Belinkov, *supra* note 476; Bourtoule, Chandrasekaran, Choquette-Choo et al., *supra* note 476.
[744] *See* Cooper, Choquette-Choo, Bogen, Jagielski, Filippova & Liu et al. *supra* note 405.
[745] *See supra* note 7 and accompanying text (discussing open-source models).

environment with sufficient guardrails to prevent infringement. (This is the approach generally taken by OpenAI, for example.[746])

In this world, open-source models would be extremely risky. As a result, there would likely be a split between two classes of models. Some proprietary models might train on unlicensed works and be deployed only in closed services with carefully designed guardrails. Open-source models would be trained only on public-domain and openly licensed works or be trained using very conservative methods to attempt ensure that extremely little copyrighted material was memorized.[747]

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.[748]) 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 comparatively more attractive as removal targets, both because they are upstream from many models and because it is easier to define and enforce enforceable removal obligations.[749]

4. Infringing Models

A fourth possibility is that courts would hold that some or all generative-AI services are illegal because the models themselves infringe. This outcome is an existential threat to many model trainers and service providers; it essentially 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,

---

[746] The sufficiency of OpenAI's guardrails is currently hotly contested, due to the frequency of successful adversarial behaviors and security attacks that are able to circumvent them. *See supra* Part II.E.2.c (for an example of a user circumventing mechanisms to prevent the generation of potentially copyright-infringing illustrations of Calvin and Hobbes). Nasr, Carlini, Hayase et al., *supra* note 7 (for an attack on ChatGPT that breaks alignment and gets the system to regurgitate training data at relatively enormous rates).

[747] But this would still be insufficient to guarantee that possible model generations are not substantially similar to copyrighted expression. *See* Cooper, Choquette-Choo, Bogen, Jagielski, Filippova, Liu et al., *supra* note 405.

[748] *See supra* note 739 and accompanying text (for a similar point with respect to released models).

[749] *See* Cooper, Choquette-Choo, Bogen, Jagielski, Filippova, Liu et al., *supra* note 405.

Stability AI, Databricks Mosaic, and some of their partners. In this regime, the most important component of copyright law would quickly become licensing. Models could only be trained on data that had been licensed from the copyright owners, and the terms under which those models and their generations could be used would have to be negotiated as part of the licensing agreement. 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.

B. 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.

1. 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?[750] Can a generative-AI system do legal work,[751] and should they be allowed to?[752] But these issues mainly have to do with the outputs of a generative-AI system. In contrast, copyright pervades every step of the process; copyright is present every time anyone anywhere in the supply chain makes a decision. Copyright cannot be ignored.[753]

---

[750] Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489 (2023); Jon Garon, *An AI's Picture Paints a Thousand Lies: Designating Responsibility for Visual Libel*, 3 J. FREE SPEECH L. 425 (2023); Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. FREE SPEECH L. 390 (2023); Derek Bambauer & Mihai Surdeanu, *Authorbots*, 3 J. FREE SPEECH L. 375 (2023); Peter Henderson, *Tatsunori Hashimoto, and Mark Lemley, Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589 (2023).

[751] Jonathan H. Choi, Kristen E. Hickman, Amy Monahan & Daniel Schwarcz, *ChatGPT Goes to Law School*, 2023 71 J. LEGAL EDUC. 387 (2022).

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

[753] Copyright is not the only socially relevant concept that pervades the supply chain. The supply-chain framing illuminates other legal and ethical challenges as well, such as developer responsibility for harmful uses, *see, e.g.,* Cooper, Moss, Laufer & Nissenbaum supra note 120 (for ethical challenges and accountability); A. Feder Cooper, *Between Randomness and Arbitrariness: Some Lessons for Reliable Machine Learning at Scale* (Ph.D. dissertation, Cornell University 2024), https://arxiv.org/abs/2406.09548; David Gray Widder & Dawn Nafus, *Dislocated Accountabilities in the "AI Supply Chain": Modularity and Developers' Notions of Responsibility*, 10 BIG DATA & SOC'Y 1 (June 15,

2. Copyright *Concerns* Cannot Be Localized to a Single Link in the Supply Chain

We have argued, time and time again, that decisions made by one actor can affect the copyright liability of another, potentially far away actor 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, company, or organization.

3. Design Choices Matter

Every actor in the generative-AI supply chain is in a position to make choices that affect their copyright exposure, and others'. 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 necessarily have to engage with these choices—as will AI policy more generally.

4. 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.

5. Generative AI Does Not Make the Ordinary Business of Copyright Law Irrelevant

Courts will still need to make plenty of old-fashioned, retail judgments about individual works—e.g., how much does this generated image resemble Elsa and

2023); David Gray Widder & Richmond Wong, Thinking Upstream: Ethics and Policy Opportunities in AI Supply Chains (2023) (unpublished manuscript), https://arxiv.org/abs/2303.07529 (for the environmental and labor considerations involved in AI training).

Anna in particular, rather than generic tropes of fantasy princesses? To decide these cases, courts will need to avoid getting distracted by the shininess of new technologies and chasing after inappropriately categorical new rules. Similarity is similarity, proof of copying is proof of copying, transformation in content is transformation in content. 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, in the fullness of time, 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.

6. 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.[754] 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.[755] And one of the things about generative technologies is that they cause convergence;[756] 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.

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

[754] *See supra* notes 322, 323, 423 and accompanying text (for why generations are not like collages).

[755] JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET—AND HOW TO STOP IT (2008) (developing theory of generative technologies); Cooper, Lee, Grimmelmann, Ippolito et al., *supra* note 10 (connecting Zittrain's theory of generative technologies with generative AI); Cooper, Choquette-Choo, Bogen, Jagielski, Filippova, Liu et al., *supra* note 405 (for additional discussion of the same connection).

[756] *See generally* Tejas N. Narechania, *Convergence and a Case for Broadband Rate Regulation*, 37 BERKELEY TECH. L.J. 339 (2022) (discussing convergence caused by the Internet).

determines which participants in the generative-AI supply chain are copyright infringers. The underlying technologies and systems are too diverse to be treated identically, and copyright law has too many open decision points to provide clear answers. Our hope is that the supply-chain framing provides a clear and precise mechanism for understanding this diversity and, in turn, for reasoning about the various legal consequences.

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.