Artificial Intelligence and Copyright
88 Fed. Reg. 59942
Docket No. COLC-2023-0006
October 30, 2023

Google LLC appreciates the opportunity to submit these comments in response to the U.S. Copyright Office's notice of inquiry, *Artificial Intelligence and Copyright*, 88 Fed. Reg. 59942 (Aug. 30, 2023). Google has been at the forefront of artificial intelligence (AI) research for many years and has used AI to improve products relied on by billions of people. Our teams across Google, including our AI researchers at Google DeepMind and Google Research, have achieved breakthroughs like AlphaGo, AlphaGo Zero, and AlphaFold; pioneered advances in transformer models; and published more than nine thousand research papers, including many of the leading contributions to the field. As a result, Google brings extensive experience to the questions raised by this notice of inquiry.

We believe that AI is a foundational and transformative technology that will provide compelling benefits to society through its capacity to assist, complement, empower, and inspire people in almost every field of human endeavor. Indeed, it will help us tackle some of society's most pressing challenges and seize new opportunities, from the everyday to the more creative and imaginative. We are excited about our work in AI and the potential of this technology, and we are committed to developing it responsibly in a way that ensures our products become even more useful to people.[1]

The Office's notice underscores some novel copyright law and policy issues raised by AI systems. However, we believe that existing copyright doctrines are sufficiently flexible to handle many of the scenarios that will likely arise with AI, and that courts — informed with the facts of specific cases — are the appropriate first venues for determining how those doctrines should apply. As we'll explain in our comments below, there are well-developed bodies of case law to guide judicial analysis of the most salient questions concerning both copyrightability and infringement in the context of AI systems. The through-line in those precedents is finding the right balance between the legitimate interests of rightsholders and the equally legitimate interests of the public and succeeding generations of creators.

The doctrines of originality and authorship hold the keys to questions of copyrightability for AI-generated outputs, teaching that authorship is a fundamentally human endeavor, even where sophisticated tools enhance human creativity. The doctrine of fair use provides that

---

[1] *See* Kent Walker, *Our commitment to advancing bold and responsible AI, together*, The Keyword (Jul. 21, 2023), https://blog.google/outreach-initiatives/public-policy/our-commitment-to-advancing-bold-and-responsible-ai-together/.

copying for a new and different purpose is permitted without authorization where — as with training AI systems — the secondary use is transformative and does not substitute for the copyrighted work. The volitional conduct doctrine sets the line between direct and secondary infringement, finding the direct infringer by inquiring whose action proximately caused an infringement. And the "staple article of commerce doctrine" protects the development and distribution of new products — like generative AI systems — that are capable of substantial noninfringing uses. As the Supreme Court recently stated, applying existing copyright law to new technologies is "a cooperative effort of Legislatures and courts" and "Congress . . . intended that it so continue."[2] In light of that ongoing effort and the rapidly evolving nature of AI technology, any legislative action now would be premature and could hinder innovation and the many opportunities that come with it.

* * *

I. Our Approach To Developing Artificial Intelligence

Recent years have seen huge breakthroughs in the use and application of artificial intelligence — and AI holds major promise for people around the world. It has the potential to unlock major benefits,[3] from better understanding diseases[4] to mitigating climate change[5] and driving prosperity through greater economic opportunity.

AI also already powers Google's core products, which help billions of people every day. Whether it's asking for movie times, finding the nearest doctor, or finding better routes home, our work in AI is centered on improving people's everyday experiences. Some of our most popular products at Google — like Lens and Translate — have at their core AI technologies like optical character recognition and machine learning. And countless other Google products now have AI built into them, making them more helpful to billions of people.[6]

Many of these improvements are possible thanks to Google Research's introduction of the Transformer model in 2017.[7] The Transformer is considered the foundation of modern language

---

[2] *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1208 (2021).
[3] *See* Yossi Matias, *How we're using AI to help transform healthcare*, The Keyword (Oct. 23, 2023), https://blog.google/technology/health/how-were-using-ai-to-help-transform-healthcare/.
[4] *See* Google DeepMind, *The race to cure a billion people from a deadly parasitic disease,* Google DeepMind Blog (Jul. 28, 2022), https://unfolded.deepmind.com/stories/accelerating-the-search-for-life-saving-treatments-for-leishmaniasis.
[5] *See* Kate Brandt, *New ways we're helping reduce transportation and energy emissions*, The Keyword (Oct. 10, 2023), https://blog.google/outreach-initiatives/sustainability/google-transportation-energy-emissions-reduction/.
[6] *See* Justin Burr, *9 ways we use AI in our product*, The Keyword (Jan. 19, 2023), https://blog.google/technology/ai/9-ways-we-use-ai-in-our-products/.
[7] *See* Jakob Uszkoreit, *Transformer: A Novel Neural Network Architecture for Language Understanding*, Google Research Blog (Aug. 31, 2017), https://blog.research.google/2017/08/transformer-novel-neural-network.html.

models; on top of this architecture we are now able to build AI language models — like BERT, PaLM, MUM, and LaMDA — that can do everything from solving complex math word problems to answering questions in new languages. They can even express their reasoning through chain-of-thought prompting.

We believe our approach to AI must be both bold and responsible. To us that means developing AI in a way that maximizes the benefits to society while addressing the challenges. We were one of the first companies to publish a set of AI Principles,[8] and we use an AI risk-assessment framework to identify and mitigate risks. Google DeepMind likewise adheres to our AI Principles and has a dedicated internal governance body — the Responsibility and Safety Council — tasked with upholding them. We also constantly learn from our research, our experiences, our users and the wider community — and incorporate what we learn into our approach to developing and deploying AI.

II. How AI Systems Work

A. Artificial Intelligence and Machine Learning

Before discussing the process of machine learning, it is important to understand some basic terminology and types of AI systems. The term *artificial intelligence* describes a broad and diverse set of technologies. AI systems can be built in many ways. For instance, the computer opponent in a basic chess app might have some deterministic behaviors that are hard-coded by the developer in the form of if-then statements known as decision or production rules. The same is often true even of more advanced AI systems. The Deep Blue chess system that was the first computer to win a match against reigning world champion Gary Kasparov in 1997 was an *expert system* built using a large rule-set meant to imitate experts in a deterministic manner.

Much of the recent progress we've seen in AI is based on *machine learning (ML)*, a subfield of computer science where computers learn to recognize patterns from example data, rather than being programmed with specific rules. Because of the way they are built, ML models are able to complete tasks and solve problems that would have been impossible for expert systems. *Deep learning* is a specific ML technique based on neural networks. Neural networks use nodes or "artificial neurons," inspired by models of brain neurons, as fundamental processing units which receive numeric inputs from, and pass outputs to, other neurons. Deep learning connects multiple layers of these artificial neurons. The interconnections between these neurons, also referred to as nodes, are numerical weights that essentially represent the importance of the contribution of that neuron to the final output. It is also relevant to note that there is no copy of the training data — whether text, images, or other formats — present in the

[8] *See* Google AI, *Our Principles*, Google AI Responsibility, https://ai.google/responsibility/principles/ (last visited Oct. 27, 2023).

model itself. Deep neural networks themselves determine the attributes of the data that they use to recognize patterns, as opposed to a human coder setting those attributes manually.

For example, the AlphaGo model, developed by Google DeepMind, was the first computer program to defeat any professional human Go player and, soon after, the first to defeat a Go world champion. To accomplish that, AlphaGo had to employ an ML model.[9] That is because Go is a profoundly complex game, one googol (10^{100}) times more complex than chess. The power of the decision-making engine underlying Deep Blue would have been no match for a Go world champion.

ML models have long been used for classification or prediction purposes, e.g., a system that can detect cats in photos or predict vehicular traffic patterns. However, recently, the ML models that have captured the most attention — and the type that the Office's study seems most focused on — are *generative AI* models.

B. Generative AI

Generative AI models can use what they have learned to create new content, such as text, images, music, and computer code. A "large language model" (LLM) is a generative AI model that finds patterns in human language, making it suitable for a range of writing tasks, including predicting the next words to complete a sentence or suggesting grammatical edits that preserve what you mean to say. During training, a model evaluates the proximity, order, frequency, and other attributes of portions of words, called tokens, in its training data. In fact, the model itself selects which attributes to use. In this way, training is the discovery of probabilities of relationships between the tokens — ultimately not in any individual text, but in all of the text on which the model is trained. The trained model then comprises a large network of weights that represent these learned relationships. The model can then respond to a prompt and generate new text with a probability of addressing the prompt as determined by its training.

Generative AI models are not databases or information retrieval systems. To be sure, when, for instance, an LLM is prompted for facts, it can generate articulate responses that may give the impression that it is retrieving information. But, fundamentally, the model is generating responses based on a statistical estimation of what a satisfactory response should look like. Put simply, it produces an average group of words, pixels, or sounds related to a prompt. Some have referred to this as, not an answer, but merely "answer-shaped." To understand how generative AI systems are built, it is easiest to take as an example the LLMs — like LaMDA, PaLM, and MusicLM — that underlie many of Google's latest AI advances.

---

[9] *See* Google DeepMind, *AlphaGo*, Google DeepMind Technologies, https://www.deepmind.com/research/highlighted-research/alphago (last visited Oct. 27, 2023).

The technical process of "learning" for an LLM begins with training the model to identify relationships and patterns among words in a large dataset. Through this process, a generative AI model will adjust its parameters to reflect the mathematical relationships in the data. Once the model has adjusted its parameters to accurately reflect these relationships, it can then use them to generate new outputs based on those parameters. The number of parameters needed to capture the complexities and nuances of human language and facts about the world is vast.

LLMs are developed in multiple stages, including *pre-training* and *fine-tuning*. Pre-training is a way of training an ML model on a wide variety of data. This gives the model a head start when it is later trained on a smaller dataset of labeled data for a specific task. When an LLM is pre-trained, training material is analyzed to examine and extract statistical relationships among the individual words and sentences, e.g., their frequency, importance, and semantic relationship to each other. The AI "model" is simply the encapsulation of those statistical facts in numbers. And given enough content — on the scale of hundreds of billions of words — the model may be able to embody human language as a whole in the form of its parameters and nodes. Importantly, given the volume of data that models need to train on, any particular work standing alone is not essential or even necessary for that training. Instead, it is the total *collection* of works that is needed to train an AI model. Following pre-training, the model can be refined through a process called fine-tuning. Fine-tuning an LLM is the process of adapting a pre-trained LLM to improve its performance on a specific task. The model learns from additional example data to help hone its capabilities. For instance, one can fine-tune a general purpose LLM to teach it how to summarize technical reports in general by using a smaller set of examples of technical reports and accurate summaries.

III. How AI Will Unlock Scientific Discoveries, Help Organizations Tackle Societal Challenges, and Improve Our Everyday Lives

AI's potential societal benefits to the U.S. and the world cannot be overstated. The technology's uses are extensive. From powering research that enables new scientific breakthroughs to product integrations designed to make everyday life easier, we're exploring responsible and innovative AI technologies that make a true difference for humanity. We are excited about the promise AI holds for solving some of the most persistent challenges facing our world.

AI has the potential to significantly improve healthcare, including maternal care, cancer treatments, and tuberculosis screening. For example, understanding how a protein folds is important for medical research, but it is also time-intensive and painstaking. Google DeepMind's AlphaFold predicted 200 million protein structures that previously would have taken several years each to discover, effectively saving hundreds of millions of years of

researchers' time.[10] Structural biologists who use AlphaFold have seen their productivity grow 20% faster than those who do not. Google Research also recently announced a new LLM that could be a helpful tool for clinicians: Med-PaLM.[11]

AI can also help with mitigating and adapting to climate change: by tracking wildfire boundaries in real time; helping to reduce carbon emissions by decreasing stop-and-go traffic; and providing critical flood forecasts.[12] Partnerships in the field of climate science will help organizations develop innovative solutions. For example, Google Research recently teamed up with American Airlines and Breakthrough Energy to bring together large amounts of data — like satellite imagery and weather and flight path data — to develop forecast maps to test if pilots can choose routes that avoid creating contrails (i.e., the thin, white lines sometimes seen behind airplanes that account for roughly 35% of aviation's global warming impact). This partnership showed that contrail avoidance has the potential to be a cost-effective, scalable solution to reduce the climate impact of flying.[13]

In addition, AI is powering progress in making the world's information accessible to people everywhere. Google's Data Commons project synthesizes publicly available data from government agencies and other authoritative sources into an open source, API-accessible knowledge graph available to everyone. It links references to unique entities (such as cities, counties, organizations, etc.) that exist across different datasets to nodes on the graph, such that users can access data about a particular entity aggregated from different sources without the significant data wrangling procedures required to clean or join records. Data Commons is also now using LLMs to create a natural language interface that allows users to ask questions, making it even more useful.[14]

And through our 1,000 Languages Initiative, we are working to build an AI model that will support the world's 1,000 most-spoken languages, bringing greater inclusion to billions of people in historically marginalized or underserved communities all around the world. While more than 7,000 languages are spoken globally, only a few are well represented online today. That means traditional approaches to training LLMs on text from the World Wide Web fail to capture the diversity of global communication. We've already made significant progress towards this goal with a Universal Speech Model trained on more than 400 languages.[15]

---

[10] *See* Demis Hassabis, *Putting the power of AlphaFold into the world's hands*, Google DeepMind Blog (July 22, 2022), https://deepmind.google/discover/blog/putting-the-power-of-alphafold-into-the-worlds-hands/.
[11] *See* Google Research, *Med-PaLM*, https://sites.research.google/med-palm/, (last visited Oct. 27, 2023).
[12] *See* Yossi Matias, *How we're using AI to combat floods, wildfires and extreme heat*, The Keyword (Oct. 10, 2023), https://blog.google/outreach-initiatives/sustainability/google-ai-climate-change-solutions/.
[13] *See* Carl Elkin & Dinesh Sanekommu, *How AI is helping airlines mitigate the climate impact of contrails*, The Keyword (Aug. 8, 2023), https://blog.google/technology/ai/ai-airlines-contrails-climate-change/.
[14] *See* Data Commons, *About Data Commons*, https://www.datacommons.org/about (last visited Oct. 27, 2023).
[15] *See* Jeff Dean, *3 ways AI is scaling helpful technologies worldwide*, The Keyword (Nov. 2, 2022), https://blog.google/technology/ai/ways-ai-is-scaling-helpful/.

AI can also make a difference in our everyday lives. For example, AI is already powering many products that millions (and in some cases billions) of people use, such as Google Maps, Google Translate, Google Lens, and more. And now we are leveraging AI to help people ignite and assist their creativity with Bard, increase their productivity with Workspace tools, and revolutionize the way they access knowledge in Search. These types of tools have the potential to make everyday experiences easier, more productive, and more creative.

IV. How We Are Working With the Creative Community To Unlock New Opportunities

We also believe in AI's potential to amplify and augment human creativity, unlocking new opportunities for artists, creators, journalists, musicians, and consumers to engage creatively with new tools and expand the pie for everyone. In fact, we are already seeing creators exploring new areas, including the creation of new types of music, books, photography, clothing, pottery, games, and other art inspired in collaboration with AI models. We're excited about how AI can supercharge human creativity — not replacing it, but enhancing, enabling, and liberating it.

With this in mind, we are committed to building tools that increase access to information and create new and expanded economic and creative opportunities for artists, small businesses, and creators of all kinds. To do this, we are working closely with the creative community to put these tools in the hands of creators and to tackle new challenges as they emerge.

For example, we are working closely with our music partners to develop an AI framework to help us work toward our common goals. This includes YouTube's Music AI Incubator. The incubator will help inform YouTube's approach as we work with some of music's most innovative artists, songwriters, and producers in the industry, across a diverse range of culture, genres, and experience. We also announced a set of principles that will govern YouTube's work on AI.[16]

In addition, we announced Lab Sessions, a series of experimental collaborations with visionaries — from artists to academics, scientists to students, creators to entrepreneurs — to help them use AI to compose new music, support the creative writing process, better learn sign language, and more.

Through the Google News Initiative, we are supporting training programs for journalists — so they can use AI in their work — and research into how AI can support the news ecosystem.[17] In addition, we have built research tools like Pinpoint, which helps journalists and academics

---

[16] *See* Neal Mohan, *Our principles for partnering with the music industry on AI technology*, Youtube Official Blog (Aug. 21, 2023), https://blog.youtube/inside-youtube/partnering-with-the-music-industry-on-ai/.
[17] *See* The London School of Economics and Political Science, *JournalismAI Home Page*, Dept. of Media and Communications, https://www.lse.ac.uk/media-and-communications/polis/JournalismAI (last visited Oct. 27, 2023).

explore and analyze large collections of documents.[18] Recently, a study by JournalismAI showed that almost three quarters (73%) of news organizations surveyed believe generative AI applications, such as Bard or ChatGPT, present new opportunities for journalism. Some respondents noted that AI can free up journalists' capacity for more creative work by taking on time-intensive tasks such as interview transcription and fact-checking.[19]

We are also prioritizing approaches that will allow us to send valuable traffic to web publishers, including news publishers. We've heard from web publishers that they want greater choice and control over how their content is used for emerging generative AI use cases. That is why we announced Google-Extended, a new control that web publishers can use to manage whether their sites help improve Bard and our Vertex AI generative APIs, including future generations of models that power those products.[20] And it's why we're committed to engaging with the web and AI communities to explore additional machine-readable approaches to choice and control for web publishers.[21]

V. Copyright Implications of AI Training and Use

A. Training

The U.S. copyright system provides a limited monopoly right to authors, permitting them to restrict certain uses of their work for which they have a right to be paid.[22] The "reward to the owner," however, is a "secondary consideration."[23] "The sole interest of the United States and the primary object in conferring the monopoly lie in the general benefits derived by the public from the labors of authors."[24] Our copyright law thus purposefully strikes a "balance between the interests of authors ... in the control and exploitation of their writings ... on the one hand, and society's competing interest in the free flow of ideas, information, and commerce on the other hand."[25] Ultimately, "copyright is intended to increase and not impede the harvest of knowledge."[26]

Copyright protection must also be balanced against First Amendment considerations. As the Supreme Court has noted, "copyright's limited monopolies are compatible with free speech

---

[18] *See* Brendan McCarthy, *How technology powered a Pulitzer Prize-winning investigation*, The Keyword (Nov. 16, 2021), https://blog.google/products/news/boston-globe-pinpoint-pulitzer-prize/.
[19] *See* Charlie Beckett**,** *How AI is generating change in newsrooms worldwide*, The Keyword (Sept. 20, 2023), https://blog.google/technology/ai/how-ai-is-generating-change-in-newsrooms-worldwide/.
[20] *See* Danielle Romain**,** *An update on web publisher controls*, The Keyword (Sept. 28, 2023), https://blog.google/technology/ai/an-update-on-web-publisher-controls/.
[21] *See* Danielle Romain, *A principled approach to evolving choice and control for web content*, The Keyword (Jul. 6, 2023), https://blog.google/technology/ai/ai-web-publisher-controls-sign-up/.
[22] *See Sony Corp. of Am. v. Universal City Studios, Inc.*, 464 U.S. 417, 429 (1984).
[23] *Id.*
[24] *Id.* (quoting *Fox Film Corp. v. Doyal*, 286 U.S. 123, 127 (1932)).
[25] *Id.*
[26] *Harper & Row Publishers, Inc. v. Nation Enterprises*, 471 U.S. 539, 545 (1985).

principles" because our copyright law "contains built-in First Amendment accommodations."[27] Specifically, while the law protects creative expression, that protection does not extend to the underlying ideas, theories, and facts expressed in a work; those "become[] instantly available for public exploitation at the moment of publication."[28] Additionally, via the fair use doctrine, the law has always recognized "a privilege in others than the owner of the copyright to use the copyrighted material in a reasonable manner without his consent."[29]

With the rapid technological developments of the Information Age, the fair use doctrine has played a crucial role in keeping the copyright system balanced and promoting American innovation. Applying principles of fair use, courts have done the careful work of drawing the boundary in individual cases between the legitimate interests of rightsholders in exploiting markets for their works and the interests of technologists and the public in the development of groundbreaking consumer products and services that operate in copyright-adjacent markets — e.g., photocopiers, audio and video home taping devices, personal computers, Internet search engines, and smartphones.[30] Historically, fair use has provided a workable, flexible, and technology-neutral framework for distinguishing between markets that properly fall within the scope of the copyright monopoly and markets that must be allowed to develop and exist outside it if the copyright system is to remain true to its constitutional purpose.

The boundary-drawing function of fair use is especially relevant to the copyright questions around generative AI. Fundamentally, and as explained above, the training of ML models, including those underlying both generative and non-generative AI systems, captures the statistical relationships among training data, such as, in the case of an LLM, the relationships between words as they are used in writing. If training could be accomplished without the creation of copies, there would be no copyright questions here. Indeed that act of "knowledge harvesting," to use the Court's metaphor from *Harper & Row*,[31] like the act of reading a book and learning the facts and ideas within it, would not only be non-infringing, it would further the very purpose of copyright law. The mere fact that, as a technological matter, copies need to be made to extract those ideas and facts from copyrighted works should not alter that result. That fundamental precept underlies a long line of cases regarding the use of copyrighted works for transformative, fair use purposes.

---

[27] *Eldred v. Ashcroft*, 537 U.S. 186, 219 (2003).
[28] *Id.*; *see also* 17 U.S.C. § 102(b).
[29] *Harper & Row*, 471 U.S. at 549 (quoting H. Ball, Law of Copyright and Literary Property 260 (1944)).
[30] *See Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1198 (2021) ("[Fair use] can focus on the legitimate need to provide incentives to produce copyrighted material while examining the extent to which yet further protection creates unrelated or illegitimate harms in other markets or to the development of other products. In a word, it can carry out its basic purpose of providing a context-based check that can help to keep a copyright monopoly within its lawful bounds.").
[31] *Harper & Row*, 471 U.S. at 545.

For example, in *Authors Guild v. Google, Inc.*,[32] Google's creation of digital scans of more than 20 million books was deemed to be a transformative fair use because, among other things, it enabled the creation of Google Book Search and the ability to engage in statistical analysis of language through text-and-data mining.[33] Similarly, in *A.V. ex rel Vanderhye v. iParadigms, LLC*,[34] the court found that it was fair use to make copies of entire works for the purpose of building a system to identify plagiarism in student essays. The court noted in particular that the use of the copied works "as part of a digitized database from which to compare the similarity of typewritten characters used in other student works" was "unrelated to any creative component."[35]

Some might object to this logic in the context of generative AI systems, arguing that even if such a system produces content that is not substantially similar to any of the content it was trained on, the output of that model might compete in the marketplace with works used for training or, more broadly, with the *authors* of those works with respect to any *future* works they might create. This argument misunderstands both the nature of the fair use inquiry and the creative markets that copyright is intended to protect. Even if generative-AI-assisted outputs do compete with existing works that were used in training, or with future works by the authors of those works, the pro-competitive nature of copying for the purpose of "knowledge harvesting"[36] has traditionally been a reason to *favor* a holding of fair use, not a reason to reject it. For example, in *Sega Enterprises Ltd. v. Accolade, Inc.*,[37] the court held that it was fair use for Accolade to make copies of Sega's computer code for the purpose of reverse engineering — to study how to make new video games that could be played on Sega's console — even though Accolade's new games would thus be able to compete with those made and authorized by Sega.

The fair use analysis *is* concerned with market harm to a rightsholder resulting from a secondary use, but the relevant market for purposes of that analysis is "the potential market for or value of *the copyrighted work*."[38] As the Second Circuit explained in *Authors Guild, Inc. v. HathiTrust*,[39] the fourth fair use factor "is concerned with only one type of economic injury… the harm that results because the secondary use serves as a *substitute for the original work*."[40] The fact that a use "enables [a user of an AI tool] to enter the market for *works of the same type* as the copied work," *i.e.,* by using the tool to create new visual works that compete in that

[32] 804 F.3d 202 (2d Cir. 2015).
[33] *Id.* at 209 ("The search engine also makes possible new forms of research, known as 'text mining' and 'data mining.' Google's 'ngrams' research tool draws on the Google Library Project corpus to furnish statistical information to Internet users about the frequency of word and phrase usage over centuries.").
[34] 562 F.3d 630 (4th Cir. 2009).
[35] *Id.* at 641-42.
[36] *Harper & Row*, 471 U.S. at 545.
[37] 977 F.2d 1510 (9th. Cir. 1992).
[38] 17 U.S.C. § 107 (emphasis added).
[39] 755 F.3d 87 (2d Cir. 2014).
[40] *Id.* at 99 (emphasis added).

market, is not the kind of market harm that is relevant to fair use.[41] If it were otherwise, copyright would become a tool for suppressing rather than incentivizing the creation of new works. While there is no dispute that a regurgitated replica of an original work still under copyright would infringe, simply exposing models to millions, billions, or trillions of data inputs to derive model weights does not.

Importantly, innovation in AI fundamentally depends on the ability of LLMs to learn in the computational sense from the widest possible variety of publicly available material. The process of training an ML model is analogous to reverse engineering, which was held to be fair use in both *Sega* and *Sony Computer Ent. v. Connectix Corp.*[42] As described above in Section II, AI training is a computational process of deconstructing existing works for the purpose of modeling mathematically how language works. By taking existing works apart, the algorithm develops a capacity to infer how new ones should be put together. This deconstructive, computational use of creative works in model training is fundamentally different from the communicative, aesthetic purpose for which those works were created. Under the Supreme Court's recent decision in *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*,[43] making a copy of a work to train a computer algorithm "uses the work to serve a distinct end"[44] and is thus transformative for fair use purposes, even where the resulting model is generative rather than predictive or classificatory.

Fair use and similar doctrines around the world support innovation by ensuring that developers are able to assemble the building blocks needed for the development of AI. These provisions further the purpose of copyright law by purposefully and carefully balancing protections for creators with the need for innovation and cumulative creativity. Any prohibition or limitation on the use of copyrighted materials for purposes of AI training would therefore undermine the purpose of copyright and foreclose the many opportunities that come with this technology.

B. Transparency and Recordkeeping

There has been significant discussion about compelling AI developers to disclose the datasets that they have trained on. Most LLMs are trained on a wide variety of publicly available online data, including web-crawled data, rather than on 'offline' datasets that are prospectively compiled and documented. Given that fact, a disclosure requirement would be unsound policy for several reasons: First, the source of much of the training, validation, testing, and input data is the massive volume of content available on the entire open World Wide Web — in contrast with models that use a limited number of well-defined, readily identifiable sources. Second, identifying the datasets used to train particular systems would expose competitively sensitive (and potentially trade-secret-protected) information. And third, AI developers do not

[41] *Sega*, 977 F.2d at 1523.
[42] 203 F.3d 596 (9th Cir. 2000).
[43] 143 S. Ct. 1258 (2023).
[44] *Id.* at 1274.

have access to detailed or accurate information about copyright status, ownership, or licensing terms for the content available on the public web. In fact, there is no such source of truth anywhere in the world. In addition, recently Google and other AI developers announced improved web publisher controls for training of generative AI models.[45] These controls, and others like them soon to follow, make disclosure requirements unnecessary because they enable rightsholders to know and control ex ante whether their online content may be used for training of future models. Further, giving web publishers the ability to choose whether or not their content may be used for training may also facilitate new, market-based solutions.

C. Generative AI Outputs

1. Copyrightability

As the Office has recognized, works that are generated by AI without cognizable human creative intervention are not copyright-eligible because they do not meet the constitutional requirement of authorship. AI systems do not need an incentive to create, and so there is no sound public policy reason to extend copyright protection to AI-generated works.[46] That said, the presence or absence of sufficient human intervention is a nuance that will need to be addressed on a case-by-case basis. In particular, it is likely that most commercial uses of AI will entail at least some amount of human creativity. There may also be many cases where creators use these tools integratively as part of their creative process. In that circumstance, as the Office recognized in its "Zarya of the Dawn" registration decision, the final work product would be protected by copyright.[47] The question of what the scope of that copyright would be is a matter properly handled by the courts in the context of specific disputes.

2. Infringement

Under the established rules of copyright law, a work is not infringing unless its creator has improperly copied expressive content from a copyright-protected work — that is, the allegedly infringing work must be "substantially similar" to the allegedly infringed work.[48] Some have offered more novel infringement theories, challenging replication of an artist's style or arguing that all output of an AI system is an infringing derivative work of the content the system was

---

[45] Google-Extended is an example of an approach that complies with the European Union Digital Single Market Copyright Directive, and specifically Article 4's reference to machine-readable opt-out tools. *See* Danielle Romain, *An update on web publisher controls*, The Keyword (Sept. 28, 2023), https://blog.google/technology/ai/an-update-on-web-publisher-controls/.

[46] *See Thaler v. Perlmutter*, No. 22-cv-1564, 2023 U.S. Dist. LEXIS 145823, at *13 (D.D.C. Aug. 18, 2023) ("Non-human actors need no incentivization with the promise of exclusive rights under United States law, and copyright was therefore not designed to reach them.").

[47] *See* U.S. Copyright Office, Letter to Van Lindberg (Feb. 21, 2023), https://www.copyright.gov/docs/zarya-of-the-dawn.pdf.

[48] *See* 4 Nimmer on Copyright § 13.03[A] (2023).

trained on.[49] These theories are not founded in any cognizable copyright principles, and in fact run contrary to accepted copyright canons.

Styles and creative methods are not copyrightable.[50] Extending copyright protection to styles would impede the creation of wide swathes of original works and would run headlong into core First Amendment protections. Overextending the scope of the derivative work right in the context of generative AI would be equally problematic. To infringe the right to prepare derivative works, a secondary work "must substantially incorporate protected material from the preexisting work."[51] This rule is critical to ensuring that copyright remains within its proper bounds; as the court recognized long ago in *Emerson v. Davies*, every new work "borrows and must necessarily borrow, and use much which was well known and used before."[52] The competing rule would treat every instance of mere inspiration as a basis for a claim of infringement.

We are aware of concerns that AI systems can output content that is substantially similar to individual pieces of content on which they were trained. According to the flagship research paper in the field, this is an exceedingly rare occurrence, even under adversarial prompting.[53] The possibility that AI models can occasionally, despite the best efforts of their developers, output content that replicates existing expression is a bug not a feature, and developers are taking a range of measures to limit that occurrence even further, including deduplication of training data. This problem is well understood to be an open research challenge in the AI developer community and is on that basis a focus of significant attention that is expected to lead to effective interventions. It is a problem more effectively addressed technically than legislatively.

The possibility that a generative AI system can, through "prompt engineering," be made to replicate content from its training data does raise questions around the proper boundary between direct and secondary infringement. When an AI system is prompted by a user to produce an infringing output, any resulting liability should attach to the user as the party

---

[49] *See e.g., Amended* Complaint at 46-47,, *Doe 1, et al. v. GitHub*, No. 22-cv-06823, 2023 WL 3449131 (N.D. Cal. June 08, 2023); Complaint at 11, *Tremblay v. OpenAI, Inc.* No. 23-cv-03223 (N.D. Cal. June 28, 2023); Complaint at 11, *Silverman v. OpenAI, Inc.*, No. 23-cv-03416, 2023 WL 4448007 (N.D. Cal. July 7, 2023).

[50] *Hartford House, Ltd. v. Hallmark Cards, Inc.*, 846 F.2d 1268, 1274 (10th Cir. 1988) (holding that copyright does not confer "exclusive rights in an artistic style or in some concept, idea, or theme of expression. Rather, it is ... specific artistic expression ... that is being protected").

[51] *Micro Star v. Formgen Inc.*, 154 F.3d 1107, 1110 (9th Cir. 1998).

[52] *Emerson v. Davies*, 8 F. Cas. 615, 619 (C.C.D. Mass. 1845).

[53] The researchers reported that 94 near duplicate images were created, out of 175,000,000 attempts. These attempts were focused on the 350,000 images with the largest number of duplicates in the training set (500 attempts per image). The researchers needed to have not only information about which of the 160,000,000 images in the training set to duplicate, but also the caption data used to identify those most-duplicated images. *See* Nicholas Carlini et al., *Extracting Training Data from Diffusion Models* at 4-7 (Jan. 30 2023), https://arxiv.org/pdf/2301.13188.pdf (included in USENIX Security Symposium, 2023).

whose volitional conduct proximately caused the infringement.[54] The AI developer can be liable (or not) under settled doctrines of secondary copyright liability applicable to device manufacturers and online service providers.[55] A rule that would hold AI developers directly (and strictly) liable for any infringing outputs users create would impose crushing liability on AI developers, even if they have undertaken reasonable measures to prevent infringing activity by users. Had that standard applied in the past, we would not have legal access to photocopiers, personal audio and video recording devices, or personal computers — all of which are capable of being used for infringement as well as for substantial beneficial purposes.

Generative AI is a technology engineered to create new works, not to copy or facilitate the copying of existing works. It is capable of substantial noninfringing uses, and the law has long been wary of permitting rightsholders to hold up such technologies merely because they *could potentially* be used for infringing purposes. In *Sony Corp. of Am. v. Universal City Studios, Inc.*,[56] the Supreme Court held that the sale of a product that may be used to infringe "does not constitute contributory infringement if the product is widely used for legitimate, unobjectionable purposes. Indeed, it need merely be capable of substantial noninfringing uses."[57] This rule exists to limit the copyright monopoly to its proper scope so that new technologies and the markets for them are allowed to develop.[58] Excluding developers of generative AI systems from the *Sony* safe harbor would put all innovation in the field of machine learning at risk.

3. Labeling

We believe that users should be able to know when they interact with AI-generated content, where appropriate and technically feasible. Building on our long track record of providing context about the information people find online, we're adding new tools to help people evaluate information produced by our models. For example, we've added "About this result" to generative AI in Search to help people evaluate the output produced in response to their

[54] *See Perfect 10, Inc. v. Giganews, Inc.*, 847 F.3d 657, 666 (9th Cir. 2017) (stating that the requirement of volitional conduct "simply stands for the unremarkable proposition that proximate causation historically underlines copyright infringement liability no less than other torts").
[55] *See, e.g.*, *id.* at 671 ("A computer system operator is liable under a material contribution theory of infringement if it has *actual* knowledge that *specific* infringing material is available using its system, and can take simple measures to prevent further damage to copyrighted works, yet continues to provide access to infringing works.") (internal quotations omitted).
[56] 464 U.S. 417 (1984).
[57] *Id.* at 442.
[58] *See Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd.*, 545 U.S. 913, 933 (2005) (explaining that the staple article of commerce doctrine "leaves breathing room for innovation and a vigorous commerce").

queries.[59] We also introduced new ways to help people double check the responses they see in Bard.[60]

Context is especially important with images, and we're committed to building tools to enable people to identify images that have been generated using our AI systems. We recently launched a beta version of SynthID, an industry-leading tool to enable watermarking and identification, to a limited number of Google Cloud customers using our Imagen text-to-image model.[61] Similarly, we recently updated our election advertising policies to require advertisers to disclose when their election ads include material that's been digitally altered or generated. This will help provide additional context to people seeing election advertising on our platforms.

Industry stakeholders are already working collaboratively to develop voluntary provenance mitigations, such as metadata inclusion, watermarking, and other techniques for audiovisual content, which may be helpful to determine if a particular piece of content was created with their system. And we take part in the Partnership for AI's Responsible Practices for Synthetic Media, which helps offer recommendations for developing, creating, and sharing synthetic media responsibly. However, this is still an evolving area of research and product development, and any proposed regulation must be considered carefully so as not impede or stifle research in this dynamic field.

VI. Other Issues

Google understands and appreciates the concern artists have about misleading commercial use of their name or likeness. But Congress should be extremely cautious before enacting a federal right of publicity or anti-impersonation law. First, at their core, right of publicity laws restrict speech rights and can be justified only when they are narrowly tailored to serve compelling state interests.[62] A federal law that harmonizes existing state laws and incorporates their exceptions may survive First Amendment scrutiny.[63]

---

[59] *See* Hema Budaraju, *How we're responsibly expanding access to generative AI in Search*, The Keyword (Sept. 28, 2023), https://blog.google/products/search/google-generative-ai-search-expansion/.
[60] *See* Yury Pinsky, *Bard can now connect to your Google apps and services*, The Keyword (Sept. 19, 2023), https://blog.google/products/bard/google-bard-new-features-update-sept-2023/.
[61] *See* Sven Gowal & Pushmeet Kohli, *Identifying AI-generated images with SynthID*, Google DeepMind (Aug. 29, 2023), https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/.
[62] *See Sarver v. Chartier*, 813 F.3d 891, 903 (9th Cir. 2016) ("By its terms, California's right of publicity law clearly restricts speech based upon its content."); *Reed v. Town of Gilbert, Ariz.*, 576 U.S. 155, 163 (2015) ("Content-based laws—those that target speech based on its communicative content—are presumptively unconstitutional and may be justified only if the government proves that they are narrowly tailored to serve compelling state interests.").
[63] *See Sarver*, 813 F.3d at 905 (explaining that a right of publicity claim that "appropriates the economic value of a performance or persona or seeks to capitalize off a celebrity's image in commercial advertisements is unprotected by the First Amendment" but that "the First Amendment … safeguards the storytellers and artists who take the raw materials of life—including the stories of real individuals, ordinary or extraordinary—and transform them into art, be it articles, books, movies, or plays").

But any law that expands publicity rights — for instance, to prohibit the creation of works "in the style of" a particular artist — would raise serious First Amendment concerns. Indeed, a law protective of artistic style would inevitably inhibit creativity and impoverish the public domain, which is the cultural commons from which all artists can freely take inspiration. "Intellectual (and artistic) progress is possible only if each author builds on the work of others.... Every work uses scraps of thought from thousands of predecessors."[64] Moreover, any federal right of publicity would have to ensure that it does not inadvertently impose liability on platforms merely for hosting user-provided content. For instance, it would be manifestly unworkable to impose on YouTube an obligation to monitor videos posted by users to determine whether any of them might violate publicity rights, and subject it to penalties if it failed to do so.

* * *

We are still in the early days of understanding the full capabilities of AI systems, including generative AI. New applications are being developed every day, and the competitive landscape is robust. Moreover, industry practices around AI training are rapidly evolving. In addition, existing copyright doctrines are sufficiently flexible to handle many of the scenarios that are likely to arise with respect to AI. Courts informed with the facts of specific cases are thus the appropriate first venues for exploring and deciding how those doctrines should apply. In light of these facts, premature legislative action or an unduly narrow interpretation of fair use could do more harm than good — limiting new opportunities for creators, consumers, and society.

[64] *Nash v. CBS, Inc.*, 899 F.2d 1537, 1540 (7th Cir. 1990) (Easterbrook, J.).