andreessen horowitz

Library of Congress
U.S. Copyright Office

Notice of Inquiry on Artificial Intelligence & Copyright (Dkt. 2023–6)

Comments of a16z
October 30, 2023

Andreessen Horowitz (a16z) is a venture capital firm based in Silicon Valley that invests in start-ups that both build and rely on artificial intelligence technologies. a16z appreciates the opportunity to join the conversation precipitated by the Office's August 30 Notice of Inquiry on Artificial Intelligence and Copyright. As a champion of many of the diverse and innovative businesses that make up this country's burgeoning AI industry, a16z's chief interest in this dialogue is to ensure that responsibly designed AI technologies remain both lawful to create and open to use.

The applications of generative AI technology reach far beyond ChatGPT and the other "chatbots" that have sparked the public's imagination in recent months. Indeed, generative AI promises tremendous societal benefits: from empowering the disabled, to delivering world-class educational resources to underserved communities, to solving some of humanity's most pressing and intractable scientific and medical problems. Allowed to reach its full potential, AI can be a tool for enhancing—not undermining—human innovation and creativity. In short, we believe that anything people do with their natural intelligence today can be done much better with assistance from AI.

a16z cares deeply about making sure that the vast opportunities unlocked by this technology are open to everyone. The Office can help to achieve that objective by ensuring that the copyright framework surrounding AI minimizes barriers to entry into

this nascent industry. In the comments that follow, we begin by providing an overview of the many remarkable ways in which AI is already making enormous positive change and solving problems across sectors, to illustrate just what is at stake. We then offer our perspectives on a few of the specific topics raised in the NOI, focusing primarily on issues concerning the training of AI models.

I. The Revolutionary Promise of Artificial Intelligence[1]

It is no exaggeration to say that AI may be the most important technology our civilization has ever created, at least equal to electricity and microchips, and perhaps even greater than those innovations. Like those earlier technological advancements, the changes resulting from AI will have profound impacts on society, the economy and national security. The critical thing to understand about AI is that it is not a *replacement* of human intelligence but a profound *augmentation* of it. It has the potential to make everyone smarter and more capable.

AI augmentation of human intelligence has already started. It is already around us in the form of computer control systems of many kinds. And it is now rapidly escalating with AI Large Language Models, and will accelerate very quickly from here—if we let it. We are already seeing glimmers—just glimmers—of that promise today:

- AI is driving medical innovation. As the FDA has recognized, AI is being used across "the landscape of drug development—from drug discovery and clinical research to postmarket safety surveillance and advanced pharmaceutical manufacturing."[2] AI is being used to scan medical images for patterns that

[1] This section generally responds to question 1.

[2] https://www.fda.gov/science-research/science-and-research-special-topics/artificial-intelligence-and-machine-learning-aiml-drug-development

suggest the presence of cancer.[3] And AI is helping doctors in even more mundane ways, including by helping them complete the necessary paperwork for treatment and billing.[4]

- Leaders—from baseball managers[5] to CEOs[6] to government leaders[7]—are using AI advisors to help digest enormous amounts of data to provide input into the merits of different decisions. We believe that AI has the potential to become a valuable decision-making tool even in the most critical and high-stakes circumstances. For example, AI can make warfare less destructive by helping military commanders and political leaders make better strategic and tactical decisions, minimizing risk, error, and unnecessary bloodshed.

- A number of companies are developing AI research assistants to help scientists brainstorm new ideas, draft outlines for research papers, and summarize massive volumes of research findings.[8] Before long, AI will exceed the capabilities of humans in solving the most complex mathematical problems.[9] AI assistants will allow scientists to supercharge their research efforts, leading to more and better scientific discoveries and engineering achievements.

- An independent school in Palo Alto, California is using a specially-designed AI tutor to help students who have questions about their assignments. The AI has been so effective that some predict that "simulated tutors could soon be as individually

[3] https://www.cancer.gov/news-events/cancer-currents-blog/2022/artificial-intelligence-cancer-imaging

[4] https://www.nytimes.com/2023/06/26/technology/ai-health-care-documentation.html

[5] https://www.cox.com/residential/articles/ai-machine-learning-baseball.html

[6] https://thehill.com/business/4029181-ceos-lay-out-visions-for-ai-uses-across-industries/

[7] https://www2.deloitte.com/us/en/pages/consulting/articles/ai-dossier-government-public-services.html

[8] https://www.euronews.com/next/2023/05/08/best-ai-tools-academic-research-chatgpt-consensus-chatpdf-elicit-research-rabbit-scite

[9] https://www.nytimes.com/2023/07/02/science/ai-mathematics-machine-learning.html

responsive to students as human tutors."[10] AI tutoring thus will be able to democratize learning like no other technology before, bringing infinitely patient and infinitely knowledgeable learning assistants to the most disadvantaged students.

- Creators of all kinds are using AI to supplement and expand their output well beyond what was possible in an earlier era. Writers are using AI to overcome writer's block.[11] Musicians are using AI to provide the building blocks for music creation.[12] Artists are using AI to help develop and refine their visions.[13] Video game developers are using AI in their art production pipelines to facilitate concepting and save time for human artists.[14] AI will reduce barriers to entry for the creative arts, allowing people living with disabilities or those who lack sufficient technical skill to express their ideas through art.

AI represents a new computing paradigm that will transform information technology and computing as fundamentally as the development of the microchip and the rise of the internet did over the past 70 years, and impact the economy in ways that cannot yet be fully appreciated. AI will increase productivity throughout the economy, driving economic growth, creation of new industries, creation of new jobs, and wage growth, allowing the world to reach new heights of material prosperity. But the only way AI can fulfill its tremendous potential is if the individuals and businesses currently working to develop these technologies are free to do so lawfully and nimbly.

[10] https://www.nytimes.com/2023/06/08/business/khan-ai-gpt-tutoring-bot.html

[11] https://www.wired.com/story/artificial-intelligence-writing-art/

[12] https://dittomusic.com/en/blog/ai-for-music-production-tools-for-musicians/

[13] https://www.moma.org/calendar/exhibitions/5535

[14] https://a16z.com/a16z-slack-debate-will-generative-ai-supplant-therapists-game-makersfriends/

Perhaps more importantly, U.S. leadership in AI is not only a matter of economic competitiveness—it is also a national security issue. The growing geopolitical rivalry between the U.S. and China makes dominance in AI a key component of our national defense strategy, illustrated best by the Department of Defense's Third Offset Strategy.[15] As China aggressively integrates AI into its military strategies, surveillance apparatus, and economic planning, ensuring U.S. leadership in AI is increasingly about safeguarding our national security – we cannot afford to be outpaced in areas like cybersecurity, intelligence operations, and modern warfare, all of which are being transformed by this frontier technology.

II. Using Copyrighted Content to Train AI Models Is Fair Use[16]

Before turning to the legal issues, it is important to appreciate a few salient facts about AI technology and its development.

- First, the only practical way generative AI models can exist is if they can be trained on an almost unimaginably massive amount of content, much of which (because of the ease with which copyright protection can be obtained) will be subject to copyright. For example, large language models are trained on something approaching the entire corpus of the written word.

- Second, AI models are not vast warehouses of copyrighted material, and any suggestion to this effect is a plain misunderstanding of the technology. As a result,

[15] *See* Eric P. Hilner, "The Third Offset Strategy," at 3 (May 2019), found at https://usacac.army.mil/sites/default/files/publications/17855.pdf (Listing among top technological priorities: "Learning machines: leveraging Artificial Intelligence and autonomy into an offset advantage; i.e., instantly responding against cyber-attacks, electronic attacks or attacks against space architecture or missiles" and "Network-enabled autonomous weapons: weapons platforms and systems plugged into a learning command, control, communications and intelligence (C3I) network.").

[16] This section responds generally to questions 8 through 8.5.

when an AI model is trained on copyrighted works, the purpose is not to store any of the potentially copyrightable *content* (that is, the protectable expression) of any work on which it is trained. Rather, training algorithms are designed to use training data to extract facts and statistical patterns across a broad body of examples of content—i.e., information that is not copyrightable.

- Third, as an empirical matter, the overwhelming majority of the time, the output of a generative AI service is not "substantially similar" in the copyright sense to any particular copyrighted work that was used to train the model. Even researchers employing sophisticated attacks on AI models have shown extremely small rates of memorization.[17]

- Fourth, over the last decade or more, there has been an enormous amount of investment—billions and billions of dollars—in the development of AI technologies, premised on an understanding that, under current copyright law, any copying necessary to extract statistical facts is permitted. A change in this regime will significantly disrupt settled expectations in this area. Those expectations have been a critical factor in the enormous investment of private capital into U.S.-based AI companies which, in turn, has made the U.S. a global leader in AI. Undermining those expectations will jeopardize future investment, along with U.S. economic competitiveness and national security.

Turning to the legal questions raised by the Office's NOI, a16z believes that generative AI model training is a productive, non-exploitive use of training material. That type of use does not exploit any protectable expression in any given work, and so it does not implicate any of the legitimate rightsholder interests that copyright law seeks to protect. It is for that reason that model training falls squarely under the fair use doctrine, the purpose of which is to "avoid rigid application of the copyright statute when . . . it

[17] *See, e.g.*, Nicholas Carlini, et al. "Quantifying Memorization Across Neural Language Models," at 3–4, 9 (Mar. 6, 2023) (finding memorization rates of roughly 1% for models trained on de-duplicated datasets).

would stifle the very creativity which that law is designed to foster." *Stewart v. Abend*, 495 U.S. 207, 236 (1990); *see also* H.R. Rep. 94–1475 at 65–66 (1976) (fair use should be "adapt[ed]" to account for "rapid technological change").

That conclusion follows a long line of established precedent. Where copies of copyrighted works are created for use in the development of a productive technology with non-infringing outputs, our copyright law has long endorsed and enabled those productive uses through the fair use doctrine. Without the safeguard of fair use, we could not have now-ubiquitous technologies like internet search engines, *see Kelly v. Arriba Soft Corp.*, 336 F.3d 811, 818–22 (9th Cir. 2003), online book search tools, *see Authors Guild v. Google, Inc.* (*Google Books*), 804 F.3d 202, 217–18 (2d Cir. 2015), and video game emulators, *see Sony Computer Entm't, Inc. v. Connectix Corp.*, 203 F.3d 596, 603–08 (9th Cir. 2000). Each of these technologies involves the wholesale copying of one or many copyrighted works. The reason they do not infringe copyright is that this copying is in service of a non-exploitive purpose: to extract information from the works and put that information to use, thereby "expand[ing] [the works'] utility." *Google Books*, 804 F.3d at 217–18.

For the very same reason, the use of copyrighted works *en masse* to train an AI model—by allowing it to isolate statistical patterns and non-expressive information from those works—does not infringe copyright either. Imposing infringement liability for the use of copyrighted works in AI model training, notwithstanding the case law that clearly demonstrates why such uses are fair, would be extremely misguided. Among other things, it would upset at least a decade's worth of investment-backed expectations that were premised on the current understanding of the scope of copyright protection in this country. The United States is currently at the vanguard of the AI industry as a direct result of these expectations and investments. But AI is not just being developed here in the United States; for example, it is also being developed in China, which views AI not as

a tool for the betterment of humanity, but as a weapon for greater authoritarian control and influence. There is a very real risk that the overzealous enforcement of copyright when it comes to AI training—or the *ad hoc* limitation of the fair use doctrine that properly protects AI training—could cost the United States the battle for global AI dominance.

The bottom line is this: imposing the cost of actual or potential copyright liability on the creators of AI models will either kill or significantly hamper their development. And, significantly, treating AI model training as an infringement of copyright would inure to the benefit of the largest tech companies—those with the deepest pockets and the greatest incentive to keep AI models closed off to competition. A multi-billion-dollar company might be able to afford to license copyrighted training data, but smaller, more agile startups will be shut out of the development race entirely. The result will be far less competition, far less innovation, and very likely the loss of the United States' position as the leader in global AI development.

III. Collective or Statutory Licensing Is Not Workable[18]

Under any potential legal framework where the use of copyrighted training data is *not* presumptively fair, anyone who wishes to lawfully continue with the development of AI models would be forced to take a license. Here, then, is another practical consequence of narrowing fair use in the context of AI model training: the necessity of somehow developing a framework for licensing the massive amounts of content required. As we discuss in this section, that task is all but impossible.

The unique considerations involved in training AI models make direct, voluntary licensing impossible. Generative AI models require not only enormous *quantities* of training content, but also immense *diversity* of content. A model that has access to only

[18] This section responds generally to questions 10 through 11.

a limited amount of training data, or only a few types or categories of works, will be unable to accurately discern the meanings of or semantic relationships among words as used in human language broadly. The fact that large rights owners are willing to strike deals is irrelevant, as such deals would only permit use of a small amount of the content needed to adequately train AI systems. In fact, the reason AI models are able to do what they can do today is that the internet has given AI developers ready access to a broad range of content, much of which can't reasonably be licensed—everything from blog posts to social media threads to customer reviews on shopping sites. Indeed, the cost of paying to license even a fraction of the content needed to properly train an AI model would be prohibitive for all but the deepest-pocketed AI developers, resulting in dominance by a few technology incumbents. This would undermine competition by the technology startups which are the source of the greatest innovation in AI.

Even copyright owners and the industry groups who represent them recognize the impracticability of voluntary, direct licensing for AI training data. They propose, instead, that Congress pass legislation to implement a statutory or collective licensing model.[19] But such legislation would effectively require AI developers to remunerate rightsholders for a use that falls squarely within the protections of the fair use doctrine. The Supreme Court has warned against such attempts to "alter[] the traditional contours of copyright protection" by reducing its speech-protective limitations, like the idea/expression dichotomy and fair use. *Eldred v. Ashcroft*, 537 U.S. 186, 219 (2003).

Moreover, a collective or statutory licensing scheme would prove administratively impossible to implement, as the Register acknowledged in a recent Congressional

[19] *See, e.g.*, "FAQs on the Authors Guild's Positions and Advocacy Around Generative AI," https://authorsguild.org/advocacy/artificial-intelligence/faq/ (discussing proposal for collective licensing).

hearing.[20] The most obvious problem is scale. Some of the most powerful AI models in existence today were trained on an enormous cross-section of all of the publicly available information ever published on the internet—that is, billions of pieces of text from millions of individual websites. For a very significant portion of those works, it is essentially impossible to identify who the relevant rights holders are, and thus there would be no viable way to get statutory royalties to the proper parties. Moreover, since this technology enables production of creative work at an unprecedented rate, the problem will compound over time. The Office has encountered similar costs and administrative problems with routing royalties to rights holders in administering the Music Modernization Act, where the total number of musical works in the entire ecosystem is something on the order of 25 million. In the context of AI training data, where the relevant quantity of works is almost certainly in the billions, these costs and problems would be multiplied by *many* orders of magnitude. The amount of "unmatched" royalties would be astronomical, leaving almost all creators with no remuneration at all.

Nor do the economics of any sort of statutory licensing scheme make sense. Again, a staggering quantity of individual works is required to train AI models. That means that, under *any* licensing framework that provided for more than negligible payment to individual rights holders, AI developers would be liable for tens or hundreds of billions of dollars a year in royalty payments. AI development in the United States would thus become much more expensive than in a jurisdiction with less burdensome restrictions, thereby entrenching the dominance of incumbent technology platforms and discouraging investment in the technology startups that drive U.S. innovation in the AI space. The creation of an impossibly high financial barrier to AI development that only the largest, technology companies have any hope of clearing is a step backward from the

[20] House Judiciary Subcommittee on Courts, Intellectual Property, and the Internet – Oversight of the U.S. Copyright Office (Sept. 27, 2023).

current environment in which innovative, new entrants are challenging these dominant platforms. Any participation by small businesses or individual innovators in the nascent AI industry—like the startups that a16z invests in—would be all but barred, and competition would be snuffed out.

* * *

AI offers us the opportunity to improve the lives of everyone in a way that few other technologies—and maybe *no* other technologies—ever have. The Office can play a part in bringing about that result not by constraining AI but by embracing it wholeheartedly, and by placing faith in the balance U.S. copyright law has always struck between protecting expression and enabling generative, non-exploitive uses. By the same token, the best way to lose the United States' current leadership in the burgeoning AI industry—along with economic competitiveness and national security benefits that leadership brings—is by rushing to pass legislation that undermines the long-standing and principled approach to copyright law that has made this country both a creative and technological leader.

a16z appreciates the Office's dedication to collecting a diversity of viewpoints concerning this critical technology and its implications with respect to the law, copyright policy, and the good of society as a whole. We are grateful for the opportunity to contribute to the conversation and look forward to further engagement as the Office continues its work.