Re: Artificial Intelligence and Copyright
(A Notice by the Copyright Office, Library of Congress on 08/30/2023)

I speak as a senior and principal systems architect with 16 years of professional software development experience. My professional journey has given me experience with small startups as well as large companies including Microsoft. I have had the opportunity to act as an artist in the game development industry (for which I have received an internationally recognized IndieCade award), and I currently work for a company that develops and uses "AI" (GAN) technology, where I have trained AI Models on my own data. However, I am submitting this comment in a personal capacity, as a proud American citizen who believes in the power of democratic collectivism over selfish individualism.

I appreciate the opportunity to have my voice heard by the Copyright Office and Library of Congress. It would have been my preference, to contribute answers to a majority of the Questions; however, given the time constraints imposed by the deadline, my very recent discovery of this opportunity, and my work and family obligations, you will find instead a smattering of a few questions that happened to fit within my time constraints. I have arranged my answers not in sequential order, according to the questions' numbers, but rather in the order that allows each answer to build upon supporting ideas established in earlier answers.

Question 3 invites the submission of links to relevant papers and studies. You will find such links embedded contextually throughout the other answers.

Question 1 solicits "views on the potential benefits and risks of this technology". This is well-tread territory, but is also important context for the other questions.

The benefits and risks, of course, vary by subject matter: a chemical-formula-generating AI will have different risks and benefits than a song-generating AI. My comments focus on the subject matter that I have expertise in. These differences arise in part from the inherently different risks and benefits of the modality (chemical formulas are capable of causing very different harms than songs are), but also from the differences in aptitude or relative capability of AI to operate in those modes.

Among the many risks and benefits that I could identify and express my views on, in context of the duties of the Copyright Office, I anticipate that the ones worth discussing are cost-savings (benefit) and unfair economic displacement/exploitation (risk).

It has been my personal observation that media-focused AI (images, videos, audio, 3D models) at this moment in time have dramatically higher degrees of capability than code-generating AI. Studies show that AI images are already indistinguishable from photographs (https://journals.sagepub.com/doi/10.1177/09567976231207095), and there are innumerable examples of AI artwork, generated in mere seconds, that match the utmost quality of works that professional artists spend hours to days creating (https://openai.com/dall-e-3).

On the other hand, my own daily efforts to coax functioning software of an acceptable level of quality from ChatGPT and competing products have as of yet failed every time. The code produced tends to miss the mark entirely, or otherwise exhibits the level of quality I would expect from entry-level software developers. While it is possible to achieve better results by engaging in an ongoing conversation with the AI or providing more robust prompts, I am able to achieve the same or better

output in less time by simply doing it myself, much unlike the seconds-to-days ratio of AI art efficiency. I've found that other senior-level software developers overwhelmingly corroborate my experiences with code-generating AI (https://jarbon.medium.com/chatgpt-sucks-for-test-automation-1c99bcf0c95e), and one formal study show success rates of less than 30% for even simple tasks (https://www.researchgate.net/publication/363267006_Assessing_the_Quality_of_GitHub_Copilot%27s_Code_Generation).

The takeaway is that, for now, media industries are more prone to the risks than other industries are, and require the most urgent attention. This point is further supported by evidence that AI is already more than decimating job opportunities for illustrators in my own industry (https://restofworld.org/2023/ai-image-china-video-game-layoffs/), while the CEO of GitHub has suggested that “the demand for software developers will continue to outweigh the supply,” for at least 10 years.

Question 7.1 asks how training materials are used in the creation of AI Models - an important question, as the involvement of Training Materials is paramount to the unfairness and exploitative foundation of today's leading AI Models. Matthew Butterick's https://stablediffusionlitigation.com/ offers a wealth of resources that answer this question with impressive technical acuity. I also suggest Kevin Henner's "intuitive introduction to text embeddings". While focused on language models, it provides excellent insight into how machine learning *in general* leverages training material to automatically reverse-engineer the implied truths and associations that artists and authors leveraged internally, in their own minds, to create the works used as that training material (https://stackoverflow.blog/2023/11/09/an-intuitive-introduction-to-text-embeddings/). All of this is essential to understanding why it is an accurate assessment to refer to the majority of generative AI models as being a matter of "copyright laundering" (https://machine-learning-made-simple.medium.com/data-laundering-how-stability-ai-managed-to-get-millions-of-copyrighted-artworks-without-paying-184239bc2d8e).

Section 8 acknowledges the relevance of the Fair Use Doctrine in determining the legality of such training materials, which are rarely licensed. Fair Use is necessarily pivotal in determining how regulation should or shouldn't proceed without further clarification from Congress.

Question 8.3 acknowledges that Fair Use explicitly sanctions noncommercial research. Having been through the college-to-startup pipeline myself, it's no surprise that the engineers behind most generative AI saw fit to continue their university-borne data-scraping habits through and beyond their transition into commercial exploitation. The culture at the intersection of venture capital and university programs effectively disregards legality, actively discourages ethics (students are told - I have been told - to "make up" numbers to make profitability "look good"), and all but ignores the existence of regulation (except where it is an opportunity for monopolizing a market, as in the case of a classmate's avocado delivery business pitch). I can't help but think of the Hawk's Nest Tunnel Disaster, in which regulations that would have saved hundreds of lives were evaded due to the mining project originating as a civil engineering project, and therefore not being subject to the relevant mining regulations (https://www.wvencyclopedia.org/articles/338). Allowing Fair Use's research exemption be "grandfathered in" to commercial activities would be a gross perversion of the law, a "cheat code" to undermine copyright.

Question 8.1 acknowledges that Andy Warhol Foundation v. Goldsmith case may be relevant to questions of Fair Use in AI; in fact, there is a noteworthy parallel between that case and Stability AI's recent developments. Warhol had acquired a license for his first use of Goldsmith's photograph - this action betrays any suggestion that the other use, which the courts ultimately deemed to be a copyright violation, should have been protected by Fair Use. If the Foundation's defense were honest and valid, Warhol would not have bothered to acquire any license for any use. The characterization of Warhol's inconsistency (careless or surreptitious?), is irrelevant.

Upon releasing their latest product, "Stable Audio", Stability AI openly boasts that this product is trained exclusively on data from AudioSparx with whom they entered a licensing agreement (https://stability.ai/research/stable-audio-efficient-timing-latent-diffusion). That Stability AI chose to license the training material for this product should be taken as an admission that it would have been a liability for them to use scraped data sources as they had with their previous image and text products. The difference surely has more to do with the rising lawsuits and public outcry against them, rather than any technical or business differences between music and images or text. Just like with their image generator, their audio generator is based on a research paper that scraped its sources (in this specific case, the preceding research scraped their training data from Spotify: https://arxiv.org/pdf/2301.11757.pdf). In fact, Stability's audio and image AI are more similar than they are different, both rooted in the same denoising research, both using the same fundamental text-conditioned U-net + VAE architecture. It may also be telling that Stability AI deemed it best to acquire a purpose-specific license with a vertically integrated music provider, instead of even purchasing CDs or downloads (e.g. from iTunes, Bandcamp) which would have been a larger and more attractive ("in the style of") data set.

Question 8.1 also asks about The Google v. Oracle case, which is indeed important, but also nuanced, and messy. In a sense, it should be seen as a precautionary tale: Although the outcome of the case was the correct one due to Oracle's licensing of the Java API, Breyer's attempts at applying the Fair Use factors to software are worryingly off-the-mark. My professional opinion is that Breyer did not consult with software developers to the degree necessary to sufficiently understand the relevant aspects of the subject matter: Presenting the Dewey Decimal System as an analogue for APIs is a gross mis-characterization of them. The truth is plainly in the name: "interface". Suggesting that APIs serve only an "organization function" is like suggesting that the particular shapes, layout, tolerances, and communication sequence of Tesla's NACS interface serve only an "organization function" for the plastic, metal, and electric signals. Software developers deserve to copyright and choose how they license their APIs just as much as car companies to copyright and license charging interfaces (https://electrek.co/2023/08/10/tesla-issues-license-volex-build-nacs-connector/). APIs are explicitly called out in every employment agreement I've ever seen in my years in the software development industry, in the section that assigns rights for the employee's or contractor's copyright-able work product to the company. Oracle had chosen, as is their right, to license the Java API and one of its implementations using a highly permissive Classpath Exception GPL, under which Google and others are granted the right to use it under the condition that they share any improvements upon request (which they do) (https://arstechnica.com/tech-policy/2016/01/android-n-switches-to-openjdk-google-tells-oracle-it-is-protected-by-the-gpl/). However, the written opinions of the majority in this case stray unnecessarily from that point, causing “distortion” that, as the dissenting opinion put it, “makes it

difficult to imagine any circumstance in which declaring code will remain protected by copyright.”. This serves as a grim warning that without better checks and balances from the other branches of government, without better software-specific legislation and executive guidance, a simple mis-characterization could be disastrous for considerable aspects of an entire competitive industry.

The Copyright Office's Notice stands in contrast to the justices' written opinion in Oracle v. Google, on the other hand, as remarkably well researched and considerate. Upon reading through the Notice, I brought it to my colleague's attention to share in this appreciation. It gives me hope to see such diligent progress in software regulation from at least one branch of our government, and I hope to see those efforts prevail in ensuring true and effective justice.

Question 5 asks if new legislation is warranted.

Just as I have illustrated that the Warhol Foundation v Goldsmith case implies a requirement for training data to be licensed for its use case, opponents of AI regulation will surely hand-pick a few examples that, in isolation, seem to support their unbounded exploitation of scraped data. As important as case law is, taken holistically, all the nation's Fair Use proceedings are a mess of contradictions when considered as a source for heuristics. Summarily, our Supreme Court has made it clear that Fair Use "is a context-sensitive inquiry that does not lend itself to simple bright-line rules." The reality is that "Generative AI" is unprecedented as context. It is a new thing that didn't exist before, and therefore must be considered in light of itself, in anticipation of the future that we want to build. It is important, in doing so, to weigh egalitarian principles of justice and The American Dream over pedantic dogmas rooted in outmoded laws. It is the sworn duty of our federal offices to manifest these new principles, as they inevitably will in the course of their business, under the guidance of President Biden's executive orders.

Section 9 asks about opt-in versus opt-out. From my perspective, this is the most powerful and important decision that our government can make to tip the economics of Generative AI one way or the other, towards fairness or injustice. To ensure that media-generating "AI" systems are ethical, economically equitable, and possible to regulate at scale, I believe the United States Government must do everything in its power to establish and enforce a strict opt-in policy for the training data of such systems. The opt-out and no-option scraping that have been used to create products like DALL-E are examples of "abusive data practices" that the AI Bill of Rights strives to protect against.

The reasoning is simple and undeniable: the legislative branch of our government has a finite amount of "bandwidth". Tackling these problems as close to the source as possible is necessary to prevent the greatest amount of injustice. Scrutinizing Fair Use principles on a case-by-case basis applied to individual users of the technology, and individual artists whose work was used to build it, is simply not viable.

Question 34: Misc. I will conclude on an observation from economist Yanis Varoufakis: The difference between feudalism and capitalism is that feudalism values owning things, while capitalism values doing things. In a sort of paradox, in order to protect those who stand to gain by doing (art), from those who stand to exploit by owning (massive data-centers), there must be a way for the former to own their doings. In America, there has been a way to own what you do - we call it copyright. I implore you to keep it that way.