Comment in re: DoJ and Stanford Workshop on “Promoting Competition in Artificial Intelligence”

To whom it may concern,

Thank you for hosting your recent workshop on “Promoting Competition in AI” and inviting further comment. While the session covered a wide array of topics, AI is a complex and rapidly evolving area, and it is crucial to continue to build a thorough evidentiary record.

With that in mind, this brief submission encourages further focus on one crucial point: existing or new barriers to accessing training data for AI can and do impact competition. This is true not only in the markets for AI tools but also competition in the markets in which AI users, including artists and content creators, participate.

As the Workshop discussions noted, training data is a crucial ingredient in the development of foundation models and other AI tools. Large, well-established, or well-resourced companies have access to significantly more training data than competitors and new entrants. As such, erecting barriers to collection of training data can act as a barrier to competition, cementing the market power of incumbents.

Consider, for instance, access to books as training data. Books are uniquely valuable for purposes of training Large Language Models (LLMs); they can significantly impact the performance and quality of the models, and help mitigate bias. But as my coauthors and I note in a recent paper exploring access to books,

> “In the status quo, large swaths of knowledge contained in books are effectively locked up and inaccessible to most everyone. Google is an exception — it can reap the benefits of their 40 million books dataset for research, development, and deployment of AI models [which required hundreds of millions of dollars in investment]. Large, well-resourced entities could theoretically try to replicate Google’s digitization efforts, although it would be incredibly expensive, impractical, and largely duplicative for each entity to individually pursue their own efforts. Even then, it isn’t clear how everyone else — independent researchers, entrepreneurs, and smaller entities — will have access.”[1]

Barriers to collection of training data cement the power not only of big tech incumbents, but also large, established media companies. They will be most able to reap the rewards of licensing their data. After all, the value of any individual artist’s data to a model trained on billions of pieces of information will be minimal; large aggregators are much more well-positioned to license information and capture value in this context.[2]

[1] https://openfuture.pubpub.org/pub/towards-a-book-data-commons-for-ai-training/release/1

[2] Whether meaningful revenue then trickles down to individual artists depends on contractual relationships and their bargaining power in the face of concentrated industry players – for explication of this point, see e.g. Cory Doctorow, “Copyright won't solve creators' Generative AI problem,” https://pluralistic.net/2023/02/09/ai-monkeys-paw/#bullied-schoolkids.

Data as a barrier-to-entry impacts more than AI tool builders, of course. It also impacts the people who rely on these tools – including creators.

Unfortunately, the workshop's "Creator Rights' Spotlight" effectively ignored the rights and interests of creators who rely on these tools to create new works. For instance, in the US Copyright Office's proceeding on generative AI, dozens of artists wrote in to explain,

> "Artists like us – and the communities for whom we create – benefit tremendously from generative AI tools. We use the tools to create new works; while generative AI can be used to exploitatively replicate existing works, such uses do not interest us. Our art is a craft – rather than simply automating the production of art, we use generative AI tools to assist our creative processes in a diverse, often complex, range of ways in order to create new works."[3]

In addition to engaging with a broader swath of creators with varying interests, the Department should engage with a more fulsome list of users overall. There are myriad other uses of generative AI – in healthcare, education, scientific discovery, business productivity and more – that have little if anything to do with producing commercial art.[4] Consider, for instance, ClimateGPT, "a model family of domain-specific large language models that synthesize interdisciplinary research on climate change."[5]

The Department's workshop considered the interests of artists in controlling use of their works and inhibiting competition with their existing works, but did not consider how giving them control over generative AI model training would in effect make them gatekeepers in industries far afield from the arts and the creative sector. Assistant Attorney General Jonathan Kanter correctly noted in his remarks that generative AI is about so much more than art and content creation; indeed, much of the controversy and litigation around generative AI has centered on a tiny percentage of the uses of this technology. Yet these other uses – and other people making uses in a wide array of other fields – were not considered.

People, including creators, certainly have a diverse range of views about these tools. The Department should work to engage further with that diversity as it evaluates how to promote competition in artificial intelligence. To that end, the Department should analyze the ways that erecting barriers to training data will impact not only builders of AI tools, but also users of these tools - many of whom are themselves competitors.

Sincerely,
Derek Slater

[3] https://www.regulations.gov/comment/COLC-2023-0006-8426
[4] See e.g. Comment from Developers and Users, Ad Hoc Group
Posted by the U.S. Copyright Office on Oct 31, 2023 ,
https://www.regulations.gov/comment/COLC-2023-0006-8427
[5] https://arxiv.org/abs/2401.09646