Reply to Stability AI's comment on questions 7, 7.1, 7.2 and 8.4.

Format:

> Indented texts are quotes from Stability AI.

Unindented texts are my replies.

7. To the extent that it informs your views, please briefly describe your personal knowledge of the process by which AI models are trained. The Office is particularly interested in:
7.1. How are training materials used and/or reproduced when training an AI model?
7.2. How are inferences gained from the training process stored or represented within an AI model?
8.4. What quantity of training materials do developers of generative AI models use for training?

> "Recent AI models are described as "generative" AI because they can generate complex new content, helping to simplify creative or analytic tasks. These models analyze vast datasets to understand the relationships between words, concepts, and visual, textual or musical features – much like a student visiting a library or an art gallery. Models can then apply this knowledge to help a user produce new content. This learning process is known as training."

Words such as "understand" and "learn" are misleading anthropomorphization of the technology. An AI model approximates the data distribution of training materials and interpolates from it—a process very different from a human experiencing the world, understanding it and forming a creative expression. Companies such as Stability AI deliberately compare data fitting to human intellectual processes in order to promote their products.
"A student visiting a library or an art gallery" is also a misleadingly reductive analogy. The real process is more akin to a student "borrowing" existing artwork from a gallery without permission and attempting to reconstruct them by nudging a random noisy image slightly toward an artwork each time (stochastic gradient descent). The student notes down how this process happened (updates model weights), so that after successful attempts at reconstruction (loss function minimized), the student can use the same process to create vast amounts of other images that look like they can fit

into the same gallery. The end result is a devaluation of the entire gallery, and the student did not learn art.
This process involves the action of copying, creates derivative works that unfairly compete with the original, homogenizes the creative space that used to celebrate individual expressions, and undermines the value of human intellectual activities.

> “For example, during pre-training, an image model such as Stable Diffusion will review billions of pairs of images with associated text captions. Through this process, the model learns to identify fundamental visual structures within images, such as shapes, textures, and patterns. By cross-referencing with known text captions, the model learns to associate these fundamental structures with particular terms. For example, the model may learn to understand the appearance of fur on a “dog”; learn how light interacts with “water”; or capture the visual aesthetic described by words like “bleak” or “Renaissance”. When a user inputs a prompt – such as “a photorealistic astronaut riding a pig” – the model can help to express the desired features in a new image, even if the model has never seen an example of that composition. ”

Again, this is a deliberate anthropomorphization of the technology. The model does not “understand”. It merely notes down the data patterns that humans associate with “dog”, “water” and “Renaissance”. It is a statistical process that works differently from human intelligence.

> “These models learn the unprotectable ideas, facts, and structures within a visual or textual system, and that process does not interfere with the use and enjoyment of original works. “Free learning” of these facts about our world is essential to recent developments in AI, and it is doubtful that these groundbreaking technologies would be possible without it. The U.S. has established global leadership in AI due, in part, to a robust, adaptable, and principles-based fair use doctrine that balances creative rights with open innovation.”

The model would be analyzing unprotectable facts if it were trained on a factual dataset, such as protein structures, traffic patterns, or even a creative commons dataset such as the Open Images Dataset (https://storage.googleapis.com/openimages/web/index.html). However, Stability AI’s products are trained on copyrighted creative expressions. The result is that the models analyze facts intermixed with human artistic expressions, and their outputs are amalgamations of extracted factual information blended with human creative expressions from which the authors have been removed, and the company tries to pass these off as “free for alls". As for what plays a bigger role, the fact that Stable Diffusion has to be trained on LAION-Aesthetics-v2-5+ (images within LAION-2B that have an

aesthetic score of 5 or higher) shows that the appeal of the model's output is highly dependent on the artistic quality of the training materials, and that the company is profiting from the skills and creativity of human artists whose work make up the training dataset more than the analytical nature of the technology itself. If today's commercial image generators such as Stable Diffusion derivatives and Midjourney were trained on the Open Images Dataset, I doubt they can even be advertised as "art generators". However, the companies behind these often use generic statements covering the entire field of AI research to advance their own interests and mask the harms they are causing on others.

In my opinion, the U.S. will truly become a global leader in AI if we find ways to use such analytical technology in fields that will benefit humanity, especially in areas that humans are struggling with, such as disease cure and climate change mitigation, which do not relate to copyright. These require efforts in building the right datasets and channeling research incentives rather than providing easy data-laundering routes for companies to profit off of the labors of productive citizens. In areas where copyright is concerned, it is more important to create a safe environment for humans to freely engage in intellectual pursuits without fear of having the fruits of their labor devoured by opportunistic tech companies.

> "Models learn behaviors, they do not store works. Through training, these models develop an understanding of the relationship between words, concepts, and fundamental visual, textual, or musical features. The model doesn't rely on any single work in the training data, but instead learns by observing recurring patterns over vast datasets (billions of image and caption pairs, and hundreds of billions or trillions of words). The model does not store the material in this training data. They do not "collage" or "stitch" together original works, nor do they operate as a "search engine" for existing content. "

First of all, the model does not form an understanding of anything. It forms statistical associations – what words commonly appear together in what context, and what imagery is labeled by humans with what text. Also, they do store training data to some extent, as shown by Carlini *et al.* (https://arxiv.org/abs/2301.13188) Finally, the models do not "collage" or "stitch" together original works in the conventional sense, but they do so in an incremental and abstract way. This is because the models are trained by an incremental copying process (stochastic gradient descent). If a conventional collage is a salad bowl, then an AI generation is a juice blend.

> "Models apply knowledge to help users produce new works. Models apply this knowledge to help a user generate new and unseen content. That could mean a novel image, passage of text, block of code, series of instructions, or video clip. This knowledge is generalizable,

> which means it can help to develop new content and support new tasks that did not appear in the training data. ”

The model can create new content through remixing what it was exposed to during training, but cannot go beyond this in more fundamental ways. For example, it cannot invent art styles without combining existing ones, and thus will not be able to lead or reshape the art space in an inspiring way without a human artist injecting their own personality and world interpretation. However, human artists are getting displaced from their jobs because AI generators can make cheap remixes of their work. Rather than a tool that aids creatives, AI generators are harming both creatives and the public.

> “Models are components in a creative tool, not independent agents. The model is one part of a creative tool that can help to produce this content, but only operates at the request of a user. The user provides creative instructions by supplying text prompts or reference examples, and adjusting other settings based on a specific desired output. The user ultimately determines how the generated content is shared, displayed, or represented to others downstream.”

AI companies market their products as “creative tools” to give users the illusion of being a “creator”. However, the user only gives instructions like how a client commissions an artist. (Sometimes the instructions are very detailed or even contain rough sketches, but some commissioners also do that.) To me, the real human creator is the set of artists whose work has been used as training materials and plays a role influencing the AI output. The AI model is more akin to an advanced middle agent that curates for references, finds a team of artists with relevant works and gets them to construct something together. Of course, this analogy is not perfect as the human artists do not have to put in hours of labor to construct the art piece, but in terms of contribution to the final output, they play a crucial role. Unfortunately, at the present, the human artists get neither credits nor compensation.