While there are potential benefits of generative AI systems in terms of streamlining certain workflows, allowing smaller teams to think bigger, etc. there are immense risks from letting them develop unabated by any regulation. Risks not only to people's livelihood, but also culture itself.

The desire for "progress" in the abstract is obviously commendable, but progress does not need to come at the expense of other broad swathes of society, as it has many, many times throughout history. Progress is not mutually exclusive with a robust framework for compensation and credit. In fact, such a framework is essential to continued progress. With no incentive to continue creating outside of the AI systems developed, the wellspring of human-produced data that AI developers currently rely on to build their systems, will very quickly dry up. in the absence of a well-developed regulatory approach, any work created will be immediately snapped up and iterated on by AI developers with no compensation, and hence no value earned by the work's creator. In this respect, AI challenges the very purpose of copyright - to promote among creators further creativity and innovation. There is already much talk about incentivising the development of AI that augments rather than replaces human workers. While that is obviously beyond the scope of the USCO, copyright will likely play an important part in this process.

The sheer volume of AI output, relative to the output of a human creator, threatens to overwhelm large portions of human culture. I do not believe this is overly dramatic. Already, on particular art platforms I have used for many years (e.g. ArtStation), AI-generated work actually comprises the majority of new images submitted. In the following years it's entirely possible that the vast majority of content on the platform will be AI. This would be a shocking contraction of the scope of works submitted to the platform. Extrapolate this microcosm to the arts as a whole, and the results would be disastrous. The regression of a full and vibrant human culture: people utilising a broad variety of tools and techniques, to a single robotic monoculture. I consider this to be a great risk of AI systems. The sheer volume of output they may produce is something that must be carefully considered by all overseeing parties.

A spectrum of regulatory approaches has already emerged in response to AI. Certain approaches, like that of Singapore, should be avoided, as they give carte blanche to AI developers to ingest anything as training data with no regard for the legal owners of that data. It is a wholly lopsided system that will utterly crater any nascent licensing market for the foreseeable future. Other approaches, (e.g. the UK), are overly permissive with regard to the granting of copyright to artefacts with little or no human involvement. I believe the US currently strikes (roughly) the right balance in applying a de minimis standard to the amount of AI in a work. The de minimis standard incentivises the development of AI tools that augment, rather than replace human work.

As it currently stands, all kinds of copyright-protected training materials are used to train AI, collected by indiscriminate web-scraping, and often also through piracy if certain works are not freely available on the web. Photographs, visual artworks, songs, speeches, films, etc. are all ingested - anything that can be operationalised as data is fed to AI systems.

There is currently almost no licensing market anywhere in the world for AI training materials, beyond a few small pilot programs like the one developed by Shutterstock. Untouched by robust regulation, any licensing markets that do arise will likely be insufficient to actually sustain rightsholders, and realistically just amount to an enormous drop in the value of the works they produce. Early reports on Shutterstock's licensing program explicitly state this is the case. They will also be rife with abuse, e.g. people stealing and uploading another's work fraudulently in order to reap the licensing reward.

Training data, once collected (collected here being a euphemism for "copied" - any collection of works as training data involves copying that work) is stored in perpetuity with AI developers, and may be used or reused in the training of as many specific models as they like. Since the development of larger neural networks earlier last decade, beginning really with models like ImageNet, it has been the norm among AI developers to hoard data, scraping it indiscriminately with no regard copyright, credit, or even often the cleanliness of the dataset itself (Child Sexual Abuse Imagery routinely appears in large datasets, simply because more images has often just meant a better model). While historically this may have been justified, e.g. researchers doing purely non-commercial work, AI development not only now has enormous commercial implications, it threatens often to replace the data they initially scraped. I believe this is an enormous infringement on the exclusive rights of copyright owners.

The explicit training goal of many current generative AI systems is actually to replicate the works in the training set. If trained to "completion" (i.e. if the "training loss" drops to 0), the AI systems would actually perfectly replicate all the works in the training set. When they are interrupted before the loss drops to 0, the system is able to produce works strongly evoking the data in the training set, without explicitly replicating it (though parts of it certainly are replicated, a process known euphemistically by AI developers as "overfitting").

I believe fair use only applies in very specific instances to the ingestion of works for training, e.g. research for a non-commercial purpose. Fair use here is very important.

However, when that non-commercial research is later adapted for commercial use, fair use should no longer apply. This is actually already to predominant way current AI systems are built, as a sort of "data laundering" allowing large commercial companies to potentially dodge thorny legal questions. Case in point: Stability AI, the company behind widely used generative art model Stable Diffusion, used a dataset called LAION as the basis for Stable Diffusion. LAION was assembled for specifically for research purposes by the LAION organisation, a non-profit. But the creation of the LAION dataset was funded in no small part by Stability AI, nominally just as a generous pledge. In reality, this was a simple investment in the assembling of a dataset to be later adapted for commercial purposes, with the intent of washing their hands of any legal scrutiny that may arise in the future. This practice would be going on at an increasingly large scale, as large companies realise there are many unsettled legal questions about the nature of AI training. Unfortunately, there is not even a reliable way to know the extent of this, as companies are not required to make their datasets/the nature of their datasets public.

The volume of training data should have no bearing on what constitutes fair use. Even systems like Stable Diffusion, which are trained on vast swathes of images, can be prompted to produce works e.g. in the style of Monet. When prompted to produce a work "in the style of Monet", the specific images in the training set by Monet will have a disproportionate impact on the resultant AI image, relative to all other works in the training set. Monet's works were absolutely essential for the creation of this specific AI output, where other images were not necessarily. If you remove all the Monet works from the training set, and ONLY the Monet works from the training set, then the model would no longer be able to output works in the style of Monet. So the rest of the images, for this specific AI output, were relatively inconsequential. It would be similar to training an AI system only on Monet images, and then having that system output an image. There is functionally very little difference between these two scenarios. So no, fair use should not become more applicable the more works are scraped.

Copyright owners should affirmatively have to "opt in" to AI training, rather than "opt out". Opt out as a system is pre-emptively making a decision on behalf of the rightsholder, saddling them with the unnecessary burden of then opting themselves out of training. It once again has great potential for abuse, with rightsholders becoming responsible for actively hunting down and removing their own works that may have been fraudulently submitted under another name, or uploaded by mistake from another source, etc. It offloads the responsibility for the quality of the dataset onto the originators of the data, rather than the curators of the dataset. Without greater transparency, these curators may not even sufficiently respect the opt out process. Consent is an affirmative process, not a retroactive one. It must be a process of opt in.

AI developers should be required by law to retain records pertaining to the nature of their training sets. This will be a fundamental part of ANY robust regulatory approach with regard to AI and copyright. It would also be highly pertinent to the safety of datasets. Transparency is key.

I believe transparency is also key in ensuring AI-generated outputs are properly labelled as AI-generated. This is important for copyright, for safety, and even for the continued progress of AI, as unlabelled AI outputs entering into the training sets of new systems can cause problems down the road. There are very, very few positive use cases for deliberately unlabelled AI-generated material. Transparency is once again key.

There may be certain instances where a human using generative AI can be considered the author of the work. I do not believe that either selecting training data or iterating a prompt is sufficient to claim authorship over the resultant work. Selecting training data may potentially allow an authorship claim over the resultant MODEL, but not that model's outputs. Iterating a prompt is simply akin to commissioning a work from a creator, and thereby should not be the basis for affording authorship over the work. No copyright regime, without the express transfer of copyright, gives ownership of the work to the commissioner of a work over the actual author. As AI is not a human, and thereby cannot itself be afforded ownership, the work should default to the public domain. This would also enrich the commons.

The responsibility for infringing AI-generated material is a complex question, and I believe will have to be settled on a case-by-case basis. Certain cases may point to the end user of the system being responsible, and others to the developer. The boundary between developer and end user may be increasingly blurred as systems become easier to train. Responsibility may even sometimes lie with the originator of the dataset.