Joseph Coco
josephcoco@protonmail.ch
225.892.5041

1. I believe generative AI can make basic tasks much more convenient. I've been using large language models to generate copy for promotional purposes for more than 3 years. I currently use large language models to answer a variety of questions both technical and non-technical in nature, to assist me in writing creative stories, to assist me in writing compelling copy to promote things, and to assist me in quickly understanding large bodies of text, to understand deeper relationships within text I may have missed, to write code in a variety of languages, and to errors produced by software. I imagine many people are doing the same thing. I've also personally used diffusion models to create visual reference for myself and others and to create placeholder art for my video game. I think it's currently affecting copyrighters directly as I imagine they are getting less work. I think stock visual artists / musicians are being affected as well as lower quality or fast turnaround products are simply generating content now rather than paying for it. I think low-end software developers are being affected as most of their work can be done by a large language model now. I think copyright holders of large datasets such as Getty is likely losing money as well for a similar reason. Technology developers are likely pivoting their services to accommodate for generative AI. Researchers may be investing more funds and energy into finding a new technique since diffusion models and BERT aren't scaling much further than they already have. I think the public is mostly exciting for all of the new tools companies are releasing but are also a little scared what might happen to creative jobs.

2. There's some concern that code written by AI may not be very secure. But people just copy code off the internet without checking it regularly which also is not a secure practice and modules are used without being inspected. I think these large language models will ultimately improve code security.

5. I believe legislation defining a clear stance on how copyrighted material can or can't be used to train artificial intelligence models would help.

7. I've trained some smaller models myself. As I understand it, large language models are first trained by processing the input to remove stop words and using word2vec to transform the input into embeddings. Those embeddings are fed together into a deep neural network with some ReLU layers. Positional information for the words is then added. Then the embeddings are fed through an attention layer which requires all outputs to be dependent only on the embeddings positionally before it. Subsequent layers operate over all embeddings regardless of position. After the data is fed through the model, it is compared to a desired output and weights are updated in the model.
I'm less familiar with diffusion models, but I know they start with the source image and inject noise into it. Then the model attempts to restore the original image from that noise, updating the weights based on the results.

7.1. I have a decent understanding of entropy and compression and can tell you that if the model size is significantly smaller than a compressed version of the input, then it's impossible for the model to be able to reproduce the input exactly.

7.2. Inferences in all generative models are stored very abstractly. It would be difficult if not impossible to point to individual layers and say how that layer specifically contributed to the resulting content.

7.3. I can't say for sure as I haven't done this, but it's possible to update the weights of a model to discourage it from producing particular output.

7.4. It's impossible to say for sure, a large language models or diffusion networks can likely be used to reproduce original input whereas that's very unlikely to do that if it wasn't part of the training data set.

8. I don't believe training models on copyrighted material for a commercial generative model is covered by fair use. It's taking the totality of the content and boiling it down to its fundamental identity. Similar to the platonic ideal of what the content is. That means extraneous details are lost, but the defining characteristics are still represented. I'm not sure how that would qualify as fair use, but I don't have a legal argument for this. The Apple Corps v. Dhanjal case in 2010 comes to mind, but that was in Britain.

8.3. If for-profit entities want to fine-tune a noncommercial generative model trained on copyrighted materials with their own data which they have the rights to, then that's fine. I don't think it would make a difference if a commercial entity funded a noncommercial model unless that noncommercial model were not open to the public or other competing businesses.

8.4. Effective generative models need to be trained on large amounts of data, generally hundreds of millions of pieces of content. I don't believe the volume has anything to do with fair use.

8.5. I think it makes sense to consider how it competes with the particular copyrighted work used to train the model.

9. I believe in an ideal world, copyright owners would have to opt in to their works being used. However, I think this would cripple the generative models beyond usability, which would just create a black market for them. Given this, I think adopting an opt out policy would be acceptable.

9.1. Opt in for commercial models and opt out for noncommercial models would be wonderful.

9.2. For images, audio, and some text, metadata would be a great way to indicate an opt-out status. However, a lot of copyrighted text exists on the internet without metadata. I believe

adding data elements to HTML would be a good means to opt-out for any text which does not contain metadata.

9.3. Many tools automatically scrub metadata, so altering content or copying it in certain circumstances may remove those opt-out or opt-in metadata fields. Further, adding metadata to a ton of content is going to take up disk space. This could be significant in terms of text content.

9.5. No, if the original human creator doesn't own the copyright, they don't have say over how the work is used unless the right to dictate how the work might be used by generative models is explicitly held by the creator in the contract.

10. An email or a letter would be fine to get permissions from content rights holders. Though having users interact with a blockchain mechanism through a web interface would be a smart way to do things as well if infrastructure were created to support this.

10.1. I think direct voluntary licensing is possible in all creative sectors, but it's not feasible to quickly have diffusion models and large language models require licensing and expect them to continue to perform to the level that they are now. It would create a black market if that were to happen.

10.3. I think a compulsory licensing regime wouldn't necessarily solve the problem, since the logistics of paying license holders is so complex. Additional infrastructure would have to be developed to support that. If that infrastructure were created, it would be trivial to allow for opting out. Royalty rates should be set on a quarterly basis. Allocation, reporting, and distribution would require an extensive online infrastructure. This infrastructure should be funded by the AI companies, though it should likely be built and maintained by a third party.

10.5. Licensing regimes fees should possibly vary by work since different types of work have different median values.

11. This sounds like it could be hairy. I think it makes sense for the group who is training the model to pay the license holder. Additionally, the group using the model should have to pay per the company which trained the model per usage, and that money would also go to the license holder.

12. It would be technically infeasible to determine exactly how much of a particular work contributed to a work generated from a model. However, with some additional complexity added, a model could give an estimate of which works contributed the most to the output by essentially searching the original dataset space for embeddings like the output. This would be effective to different degrees with different types of models.

13. I think if licensing requirements were enforced in the near future (within 9 months), then this wouldn't give time for model developers to adapt their models and retrain them. Again, this would create an immediate black market for generative AI. Now that they've built these models,

model developers can find content distributing across the latent spaces of embeddings and simply pay for usage for that content which don't have quality non-copyrighted counterparts which contribute substantially to the model. This would be a way to cut down on the fees paid by model creators and still allow for quality outputs from the models.

14. I believe having a blockchain based infrastructure would be an effective way to track license holders and potentially to pay users as well. Like the Basic Attention Token the Brave browser users paired with something like the service Uphold (https://uphold.com)

15. Requiring model developers to maintain a full copy of the data used to train their models wouldn't be ideal from a technical or financial perspectives but may be necessary for paying license holders.

16. Again, the logistics of notifying users their work is being used by AI model is great. This would be ideal but would require some extensive infrastructure.

18. I do believe generative models could be tuned to the extent that the tuning of the model represents the human creation of a work. That's difficult to define, but there's an entire genre of music created with West coast synthesizers or logic programming. Many respected artists use technology in a fundamentally creative and deep way. The GDC talk about the creation of the Doom 2016 soundtrack comes to mind. I think the human-centricity of a material should be decided on a case-by-case basis based on an explanation of the process from the creator.

20. No, I don't believe we need legal protections for AI generated materials. That would encourage corporations to use AI materials for research and pre-production services but have some meaningful restrictions on it being a final product. Yes, existing protections for computer code that operates generative AI is sufficient. It will likely continue operating as trade secrets.

21. No, raw AI-generated materials are not created by humans, and should thus not be copyrightable, like how a photograph taken by a monkey is not copyrightable to the monkey.

22. Yes, AI-generated materials can violate other copyrighted works when they are substantially similar to another one of those works.

23. Yes, substantially similar test is adequate.

24. If the copyright owner knows the source material inputted into the model, then they can plug that into the model with low stylization or variability and that will give a decent indication of if their material was used in training.

25. I think the developer of the generative AI model should be held responsible if the model was trained on copyright material. Though if users of the model were held responsible, they would probably just use contracts to still make the developers of the generative AI model responsible, so either works in the end. However, if the model wasn't trained on copyrighted material, but

still generates a work substantially similar to an existing copyrighted work, I believe the user should be responsible for any accidental copyright infringements. My hope is technology will make this less of an issue.

25.1. It depends on the license of the open-source AI model. If they're insisting that any AI model created from it, that should probably be fine. If they're insisting any material generated from the AI model is open source, that could be an issue.

28. I'm not sure it would be of much help to label content as AI generated. I think people who would want to abuse it would abuse it anyway. It likely won't cause much harm to require labeling though.

28.1. The problem is, no one can currently identify if a work is AI generated.

28.2. Yes, the models are so advanced currently, they are virtually indistinguishable from human-generated works.

28.3. I'm not sure. I think enforcement should just be done by online communities. I don't think there should be a monetary fine for an institution or an individual mislabeling.

29. I haven't used the tools, but I've read a few articles on arstechnica and elsewhere which state that AI/human differentiation models are terribly inaccurate. They have both low sensitivity and low precision.

32. I agree that the copyright office shouldn't be involved in protecting styles. Just because it's easier for a computer to emulate styles now doesn't mean the copyright office should change its procedures / policies.