I'm a software developer at a tech company that's currently working on developing AI models. I have a master's degree in computer science and have worked with generative image models since 2017.

1. Allowing copyright protection to AI generated images would cause significant harm to the creative ecosystem and the copyright system. The sheer volume of AI generated images would easily overwhelm the public space, allowing AI to claim rights to large swaths of potential creative output as a new form of copyright trolling.
   Even in its current form, generative AI is being used to displace artists, writers, and other copyright owners by exploiting the creative's own works. This form of unfair competition will serve to erode the public space just in sheer spam and volume. Even if the original author's work is better, how can they compete against someone who pushes out cheap knock-offs at a fraction of the speed and effort? The creative field is already too undervalued and it should not have to compete with its own exploitation.
   Generative AI is a great danger to the market for entry level jobs and positions since experienced workers are needed to correct its errors. This job loss is a great danger to the long term as the entry level jobs are critical for cultivating the next generation of experienced workers. Those who use generative AI as a crutch will never learn the skills for being able to fix its errors.
   In the long term, this raises a huge concern for the future of human creativity as the generative AI "... still merely regurgitate stuff they've been trained on." according to one of the most preeminent AI researchers (https://twitter.com/ylecun/status/1718263147591573949). This poses great damage to the future of creative works as those who are capable of going beyond merely regurgitating existing work will be pushed out of the market.
   From the technology and research side, generative AI's benefits are heavily limited due to its tendency to hallucinate. While some scientific value can be extracted, it's minimal compared to the damage it would cause to the creative fields.
2. I am a professional software developer. Copyright in software has always been a misfit for protecting code. Code isn't quite a purely creative work that copyright commonly protects, nor is it fit to be protected under a patent. As such, code should have its own unique discussion when it comes to copyright in the age of generative AI. Regulations that work for other forms of creative works aren't quite applicable to code and vice versa.
3. https://dl.acm.org/doi/fullHtml/10.1145/3600211.3604681
4. The EU once proposed that use of copyrighted works for the purpose of research should not be allowed to be commercialized. I'm not aware of the current status of this proposal, but a similar distinction should be made to avoid the abuse of data laundering from research into commercial products.
5. Legislation that explicitly extends copyright protection for creative works from being trained on AI would greatly address the issues related to exploitation. This should mean that copyrighted works should not be allowed to be used to train generative AI unless there is explicit opt-in agreement from the copyright owner. Opt out is not viable since it is impossible for every individual copyright owner to possibly be aware of every model

being trained. Nor should it be the copyright owner's job to notify every company and start-up. Frameworks for transparency and auditing of existing models will be necessary. However, this will not be enough as large companies hold too much power when it comes to negotiating with individual creatives to coerce them into allowing their work to be trained on. Adobe famously paid pennies per image for training their Firefly generative AI. In addition, long term negative effects on the market will still be present as works fall into the public domain. Further legislation to increase individual protection would be beneficial for the long term.

6. Since there is little protection for copyright in training, large foundational models in both the image and text domain utilize every piece of copyrighted work that's available, legally or not. Both private companies and the open source community have both completely ignored copyright when training their models. Civitai recently publicly posted a competition for the open source community to train models on specific artists.
    6.1. Data for AI models are acquired through a variety of means. Web crawlers can crawl the internet for any publicly available image, including images being rehosted on sites that don't have copyright to distribute the image. https://danbooru.donmai.us/ is a website that hosts images from artists without permission, and this has been crawled to train systems like novel.ai. Books3 is another famous example of copyrighted works being distributed and used for training.
    Prior to the announcement of the google-extended useragent, Google's AI training datasets often leveraged search crawling to get data. Data scraped from websites that trusted Google for driving traffic was then used to train AI systems without the website owner's knowledge. The data crawled from before the google-extended useragent can still be reused to train models.
    6.2. Copyrighted works are used in its entirety to train models. Images are sometimes minimally cropped or rescaled due to technological limitations.
    6.3. AI companies do not differentiate between copyrighted works and public domain works for training. Both are being extensively used. Minimal commissioning is done by developers for writing text during RLHF, but these are often outsourced to the lowest paying countries.
    6.4. Most of these models are in continual development. Therefore, AI developers retain training material to apply to any new technological developments. There is little pressure to not retain training material since it might be difficult to reacquire these materials in the future. Retransferring the data across the web would be more difficult than retaining it.
7. I've done research with generative AI models since 2016 and I work as a developer at a tech company that trains AI models.
    7.1. Data is fed to AI through a variety of means during training depending on the type of work. Modern generative AI models are unsupervised. This forces the model to be purely graded on how well they manage to reproduce the training data. For example, diffusion models are trained on the basis of how well it can denoise an image to reproduce the training data.
    7.2. The inferences from the training process are stored as weights within the model.

7.3. It is currently not possible to unlearn without retraining from scratch. Current proposals for merely blocking the offending keywords or overwriting through additional training is not actual unlearning.

7.4. There are ways to confirm if a model has trained on copyrighted works. Such as if the image can be nearly identically reproduced, or if the author's name is recognized. However, there is no way of confirming if an image is not used for training.

8. Fair use should only be allowed for discriminative AI that don't directly compete in the same market as the training data.

8.1. Both the stages of pretraining and fine-tuning should be treated the same when it comes to fair use. For the purpose of generative AI, neither are fair.

8.2. Distribution of copyrighted works without permission should be illegal regardless of whether it's for training or other purposes.

8.3. AI trained for research should not be later used for commercial purposes. If there's a need to commercialize an AI, it should be retrained from scratch with data that's not copyrighted.

As long as there's no barrier between the research and commercialization process, then funding for research should be treated as if it was done for commercialization.

8.4. LAION-5B has 5 billion urls to images in its dataset. Smaller models have been trained on fewer, but the exact number is not publicly available. Training on a larger volume of copyrighted data should be penalized more heavily as more rights are infringed.

8.5. The fourth factor for generative AI should consider the competition of all of the listed forms of competition. Since the AI model competes against all 3 types. In addition, it should consider the value of future works created by the author.

9. Copyright owners should have to affirmatively opt in since they can't be the ones who carry the burden of finding every company or individual attempting to train on their work. On the flip side, those who are doing the training have a much easier time of confirming the work is opt-in since they must at least have the work available in order to train.

9.1. Consent should be required for commercial use, or uses that can later be repurposed for commercial work. This includes open source and research usage.

9.2. Technical flags and metadata can be easily stripped and rehosted, making this infeasible. Furthermore this does not provide protection to copyrighted works created before the technical tools were available. Opt out is not a feasible approach.

9.3. If it's not feasible to acquire consent, then the training of these models should not be allowed.

9.4. Objections that aren't honored should trigger algorithmic disgorgement and retraining of the models. There's no true way for AI models to unlearn.

10. Licensing should be done in a fair manner where companies can't unfairly pressure the copyright owners to give up their copyright. The pricing should be clearly stated and no agreement should give perpetual agreements for future technology that's not yet developed.

10.1. Voluntary licensing is feasible for artworks. I can't speak for other sectors
10.2. Collective management only gives power to those managing the rights. Congress should not give antitrust exceptions for collective agreements.
10.3. Compulsory licensing regimes should not be considered

11. The developer for the AI model should be the ones responsible for securing licenses.
12. The only true way of determining the impact of a particular work is to train two models from scratch. One with the work in question, and one without. Then comparing the output image of the two models. This is infeasible and there's no known approximation of this process yet.
13. This would depend on the exact nature of the legal landscape and protections available.
14.
15. Both should be made available for auditing by a government agency or an auditing company. Low resolution images can be made available to the public for potential infringement.
16.
17. US laws in regards to the right to publicity should be considered in regards to how AI is trained. At the moment, AI can generate materials based on someone's name and likeness and infringe upon the right to publicity.
18. Most examples of human feedback are very akin to a commissioner or editor giving a human feedback. In all these scenarios with a human, the commissioner, art director, or editor is not considered the author. Instead, authorship is granted to the person producing the piece of creative work. As such, I see no reason why a human using a generative AI system should be granted authorship of the work.
    On the contrary, I'd argue that a human using generative AI should have a higher standard to cross than individual humans due to the AI's sheer output volume.
19. Explicit legal clarifications to the Copyright Act in regards to AI should be added.
20. Copyright protection for the code itself should be sufficient, its outputs do not need copyright protection.
21. Copyright Clause does not protect AI generated images. The outputs are neither useful for science nor arts.
22. AI generated outputs can sometimes reproduce near exact materials that it's been trained on. In addition, diffusion model's img2img techniques can output near identical images as its input as an attempt to copyright launder an image.
23. Further standards are necessary since AI models are reliant on copyrighted data, but don't always generate substantial similarity.
24.
25. The developer of the generative AI model should be held directly responsible. The end user should be held secondarily liable if they can be proven to use it to purposefully infringe
    25.1. Open source should be similarly held liable
26. The output should be treated as if it removed or altered copyright information.
27.
28. Labeling AI generated content would be beneficial, but this is difficult since the technology is rapidly advancing.

28.1.

28.2. While some AI generated content have distinct tells, identifying all AI generated content is impractical without documentation of the process. Future technological developments may make it impossible to identify.

29. Existing tools for detecting AI generated text are akin to flipping a coin. AI image detectors are slightly better, but can be trivially easy by cropping an image or adding some text.
30. Right to publicity should be considered for name or likeness
31.
32. There exists protections from the University of Chicago's Glaze. However, this is only a temporary measure as we wait for further legislative action.
33. Sound recordings should be protected from training without permission