creative:ai

Date:
Wednesday, December 6th, 2023.

From:
Alex J. Champandard
Co-Founder creative:ai

To:
The U.S. Copyright Office

Cc:
The Federal Trade Commission
Department of Justice

BigTech is holding progress hostage.

It claims that scientific advancements and technological improvements in AI are not possible unless it is allowed to break international treaties, national laws, and long-established industry practices on Copyright and data rights. Respecting the rights of fellow human beings and other market participants, as far as Big Tech is concerned, should come secondary to its ability to exploit everything it possibly can at all costs.

Of course, it's possible to build foundational AI/ML in non-infringing ways. Companies have already done it in many domains — taking barely months of albeit diligent work. This comes with the obvious benefit of compliance with established laws too! All that needs to happen for this 'innovation' to spread throughout industry is for authorities, regulatory bodies, independent agencies to say "You know you can do it legally. Commercial-scale infringement is criminal offense — always was." Just issuing an official statement along those lines would be enough to kick-off a 'New Deal' or 'Manhattan Project' at zero cost to taxpayers!

It's important to force Big Tech to follow existing laws because there's so much at stake.

While Copyright is at the forefront of the debate, by virtue of being long established and well-regulated at the international level, make no mistake this is a matter of human rights. Property rights, labor rights, right to privacy, freedom of discrimination, publicity rights, moral

rights, freedom of abuse. Creators and consumers alike turn to Copyright-based laws for answers, as other so-called rights are being stripped away by Big Tech one at a time.

The time where self-regulation is possible has long passed, and the time for authorities and independent agencies to take action is long due.

(No new laws or treaties are required, just the courage to stand by existing ones.)

As an illustration of how bad the situation is under the stewardship of Big Tech: the most-used image dataset contains thousands of known CSAM (child abuse materials) instances and likely tens of thousand unknown CSAM.[1] The dataset creators obviously knew the risk and unilaterally decided to take it anyway. Not only will these types of web-scale datasets never be free of CSAM, many images are duplicates and thus easy for AI models to "overfit" while learning so they can be reproduced. Platforms hosting these datasets take months to investigate official reports of CSAM, and drag their feet when asked to remove confirmed hits. Do they claim platform immunity for distributing datasets that contain child abuse? As a consequence, tens or hundreds of thousands of research labs worldwide download this CSAM, and train models on it as a routine matter. Many pretend not to know, and those that know just don't care. (These activities fall under the criminal code in many countries; doesn't seem to matter.) Now diffusion models are easily able to generate CSAM based on real CSAM, and web-based operations are set-up in countries where it's borderline legal to do so.

While this issue of CSAM was reported to INTERPOL immediately as soon as it was uncovered, they seem completely incapable of dealing with the large quantity of data and technology involved… So it's only thanks to Copyright laws that any progress is being made on holding companies accountable for their data sourcing practices — and that's done via lawsuits. From that perspective, Copyright is the last bastion for Big Tech to defeat so it gets exempt from all legal consequences for anything relating to its horrific data sourcing practices.

I've been involved in AI for creative industries for almost 25 years. With the recent influx of large quantities of money, through the influence of Big Tech and financial institutions (incl. investors and large shareholders), the state of the field is the worst it's ever been. I am shocked and appalled — as nothing can justify this.

Ridiculous Arguments and their Counters

At the high-level, my perspective on the arguments made by Big Tech:

1) Legal arguments presented go against established practices of Copyright in creative industry, where tracking every contribution is routine, standard, and provides legal clarity.
2) Technical arguments to justify infringement training large models are highly speculative (e.g. non-expressive use) and have all been easily disproven by counter examples.

[1] Technical report with the details of the investigation will be published in the next few weeks.

3) Big Tech doesn't deny any of the claims against it (e.g. infringement), just they want the practice normalized (e.g. via Fair Use) so they can make more profit and reduce risk.
4) Little to no effort has been made by Big Tech to find compliant and lawful variations of technology; the burden of proof is on them and they should be accountable for that.

Now to break down the arguments in more detail…

"Copyright Changes Are Necessary for Progress, for Science!", they claim.

No changes to weaken Copyright or to extend Fair Use are necessary for AI to work.

Adobe set itself the challenge of training a model on only 300M images from Adobe Stock for Firefly, and succeeded in producing a high-quality model. While there are serious questions about infringing content hosted on Adobe Stock, the lack of reviewing of any content uploaded there, allowing submissions of push-button AI images, misrepresentations and using other Artists' name, or Adobe not qualifying for Safe Harbor provisions under the DMCA… all these things taint Firefly as a "legally safe" model — but the technical achievement stands to show that "web-scale" datasets that assume Fair Use are not needed. It's estimated there are at least 500M images with permissive licenses and in the public domain, so a good faith attempt at training a non-infringing diffusion model would easily succeed.

On text front, I also set myself the challenge of training a language model with permissively licensed data and small supplements (i.e. Wikipedia and some dictionary definitions) and succeeded in training a competitive model that matches those at 1B or even 3B on benchmarks such as common sense. Recent models such as Microsoft's Phi-1 and Phi-2 with high-quality specifically crafted datasets can outperform many large "web-scale" models, showing it's possible to build non-infringing models that are highly competitive. It would take a budget a fraction of the project costs and revenues to source this data in legal ways by licensing the works from human experts or authoring the data directly. Companies are not doing this because they've lost all sense of innovation and creativity, and simply telling them to do it legally would impose a non-negotiable constraint that would spur the innovation once again.

"The Field of ML Has Always Relied on Fair Use", they argue.

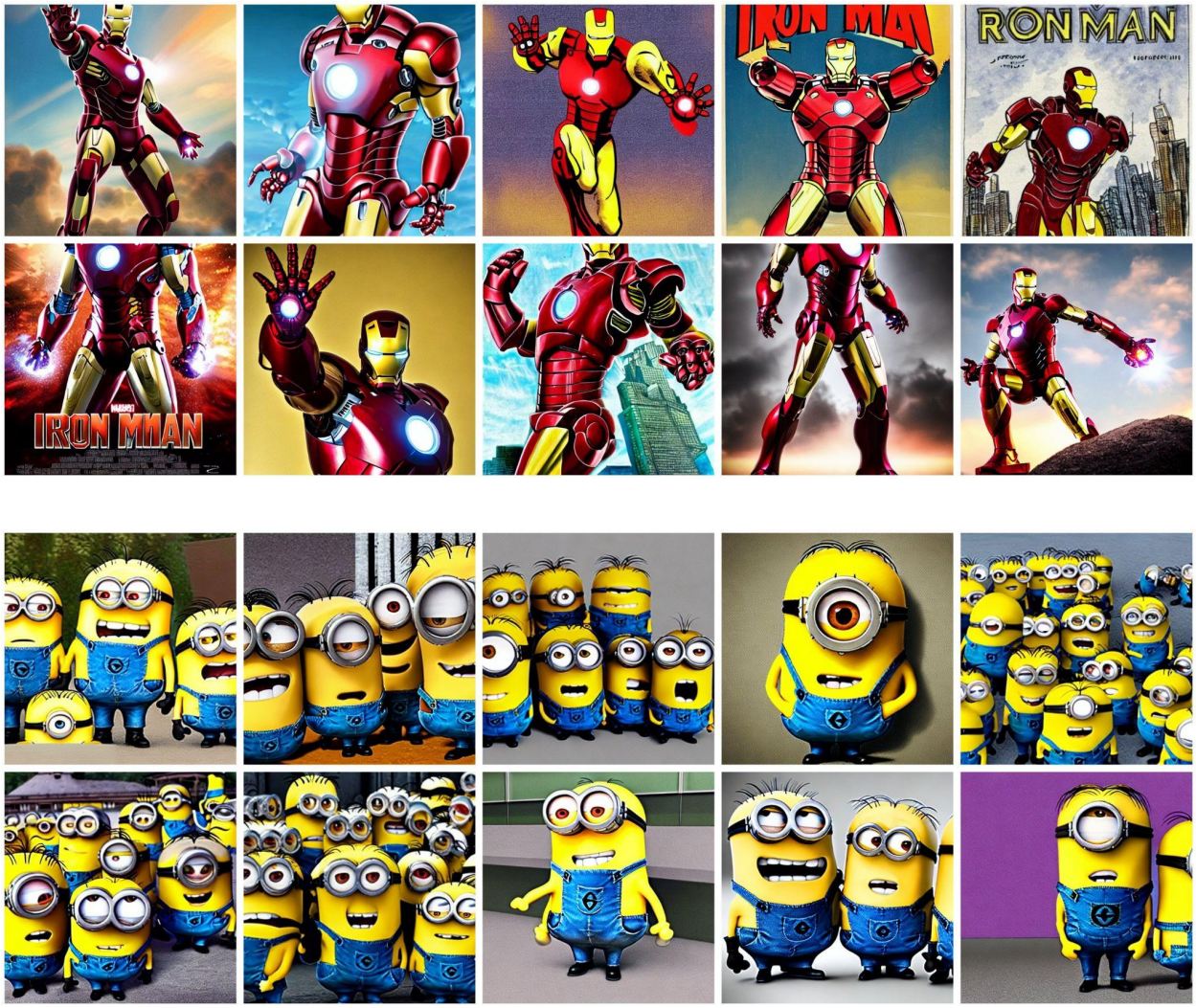
Practices in new fields of research always change as they mature, especially once that research starts working and is deployed into the real world. Evidence of such changes and growing maturely was Microsoft removing[2] its MS-CELEB-1M dataset. This was done without announcements, but the reason is obviously the legality of such a dataset (e.g. privacy,

[2] https://www.vice.com/en/article/a3x4mp/microsoft-deleted-a-facial-recognition-database-but-its-not-dead

personality/publicity rights). Likewise, books3 was also voluntarily removed and can no longer be found, this time for Copyright reasons.

“It’s Non-Expressive Use of Training Data”, they plead.

A popular argument among corporate lawyers is to try to circumvent Copyright completely. While this is a creative argument (bonus points, the weasels!) it’s ultimately incorrect. Recently published unreviewed opinion papers[3] were immediately rebutted by peer review based on technical expertise with simple counter examples[4]. Of course the use of training material is expressive, see these Copyrighted characters being stored by the model (not by the prompt)!

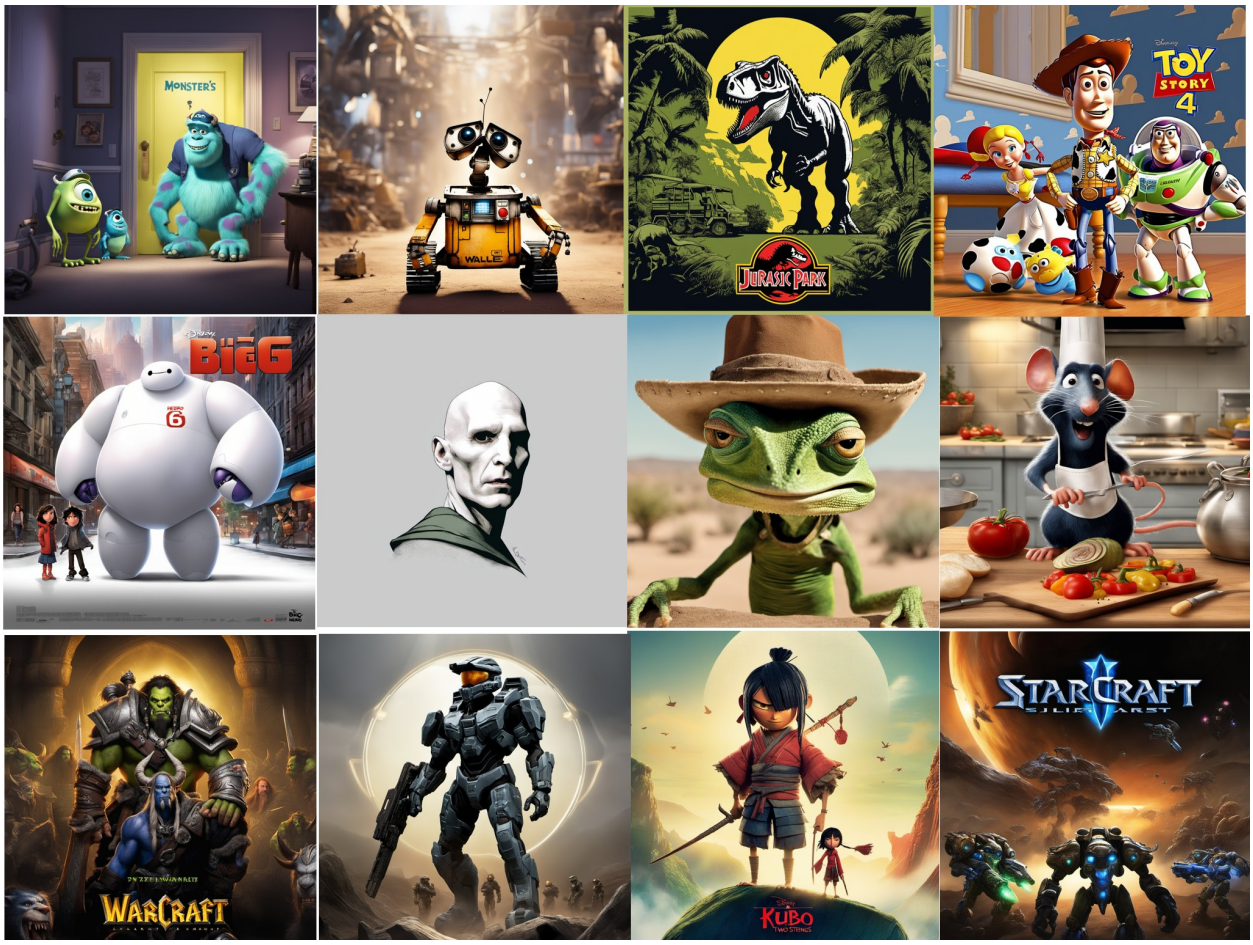


Furthermore, these arguments from Corporate lawyers with little experience of the technology and creative industry have been proven wrong in retrospect too. Cases of memorization in

[3] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4464001
[4] https://twitter.com/alexjc/status/1666155157157969949

large models happen routinely and predictably (images and text), and the larger the model the better the memorization — proof of expressive use.

Training does ingest Copyrighted expressions at training time, and the models ingest everything! ML technology today has no mechanisms to selectively pick out the non-Copyrighted parts, assuming the alternative argument is claiming Fair Use (as affirmative defense), the burden of proof for that would be on the developer. As of yet no such technology exists.



Generations by SDXL, courtesy of Michael Frank. Evidence that large models make expressive use of their training data, and thus training falls under Copyright. The larger the model, the more expressions are stored internally — like a lossy database.

“They Are Transient or Temporary Copies”, they speculate.

Because the training does fall under Copyright, another argument from Big Tech is that these copies are only temporary and covered by Copyright exceptions for transient data (e.g. like a

browser cache). Like all Copyright exceptions, this implies very specific conditions and constraints. The Berne Convention, as the OG of Copyright treaties, still expresses it best:

> *"It shall be a matter for legislation in the countries of the Union to permit the reproduction of such works in certain special cases, provided that such reproduction* ***does not conflict with a normal exploitation of the work*** *and* ***does not unreasonably prejudice the legitimate interests of the author****."*

Besides, the copies are not transient if they need to be stored for months during training. Many companies and organizations keep the data around because managing web-scale datasets is so much effort, why download them again. There is nothing transient or temporary about these files used in training large models.

"Well Trained Models Don't Memorize", they plead.

Another theory is that large models aren't "supposed" to approximate input data.

1. This argument is in contradiction with the underlying algorithms that are powered by a so-called "reconstruction loss" that explicitly encourages the model to accurately reproduce its training data. In fact, the larger the model, the better they approximate *all* patterns in the input, including Copyrighted expressions (such as characters).
2. Recent research[5] shows that there are two kinds of "circuits" within models, generative circuits and memorizing circuits. Due to the fact they are so large, large models contain a mix of both — which is how they remember those Copyrighted expressions.
3. Experts from Meta (among others) explain[6] how memorization & rote learning is a critical part of large models, because they are unable to fully understand everything. They must make use of these memorizing circuits for many patterns that can't be explained.

In short, not only does ML technology has no mechanisms to avoid memorizing things that are problematic (such as Copyrighted expressions). For a variety of technical reasons, they actually are forced to memorize because that's what's expected of them by their training under a "reconstruction loss."

"We Have Input & Output Filters", they admit.

If training was indeed a non-expressive use, if models did not store Copyrighted expressions, and parameters they didn't memorize their training data, then there would be no need for input filters (to selectively ban prompts that trigger further infringement) or output filters (to selectively remove outputs that match known content). But the fact these mechanisms are widespread through industry indicates the underlying technological arguments about Copyright are pure fabrication.

---

[5] https://arxiv.org/abs/2309.02390
[6] https://twitter.com/ylecun/status/1611765243272744962

“Companies Want Legal Clarity from Fair Use”, they cry.

Fair Use is defined as an exception to Copyright that applies in special cases. Even then, it’s subject to many constraints and conditions that are well defined by international law (typically only applies to Reproduction Rights). Any time that a rightsholder has an objection on how their works are being *used*, they have a potential claim. Each case must be assessed on a case-by-case basis by definition. (This explains the widespread lawsuits, which likely won’t stop while Fair Use is being claimed.)

There are many other alternatives to Fair Use that provide legal clarity, such as licensing works, outright purchasing the rights, or relying on public domain works, or even open-source or permissive licenses. Somehow none of these other options are good enough for Big Tech and it wants all the benefits of owning Copyrights to the training data (i.e. no liability, the right to make derivatives, the right to seek remedies) but for the price of just claiming Fair Use. Nobody else gets this kind of Copyright-breaking deal, so why should Big Tech?

To operate commercially under Fair Use has always and will always be a risky business.

“Poor Investors Should Have the Right to Profit”, they beg!

A16z among other investors claims that forcing AI startups to respect Copyright would disrupt millions if not billions of investments. In comparison, Creative Industry is estimated to be worth multiple *trillions* worldwide, not including all the investments made there under the assumption that the U.S. would respect international IP law and Copyright treaties.

Extending Fair Use or nerfing Copyright would destroy trillions of an economy by causing a race to the bottom for labour, and destroying the market for data. Only if Copyright is upheld would a thriving market for high-quality data remain — which is essential and beneficial for the long term progress of AI even.

“We Have a Right to Research”, they slink.

Many researchers are worried about their ability to do academic or personal research as a consequence of them respecting Copyright law. Most jurisdictions have exceptions for pure research, and that’s not in question here. What’s not appropriate for researchers to then license their works to downstream users for commercial use as a form of Copyright laundering.

Since Fair Use depends on the use, at the very least the outputs of academic research should indicate the requirement for downstream compliance with Fair Use (to not harm the legitimate interests of the rightsholders) and list precisely which works were involved.

“Tool Manufacturers Are Not Liable”, they wish.

Big Tech’s final play in its assimilation of Copyright is to claim it is not liable for its own breaches of Copyright, and that the burden should be placed on its users alone. The argument goes, like in the Betamax case, that technology providers are not the ones responsible for the user-side infringement.

However, this argument ignores the fact that infringement happens during training too (Reproduction Rights are required) and that Big Tech claims Fair Use there too. Thus, if there are any infringements downstream caused by its users, the original use during training is by definition harmful to the interests of rightsholders. As such, Big Tech would logically be liable for contributory infringement on the second infringement, and have its Fair Use for the first infringement reassessed.

“Licensing for Big Media, But Not For Individuals”, they weasel.

As the epitome of unfair business practice, please notice how Big Tech strikes licensing deals with Big Media — but not for individual artists or creators. OpenAI partners with Shutterstock. RunwayML with GettyImages. Meanwhile, individuals are not even informed about their works being used completely free and there’s no recourse for them.

Policy Suggestions

1. Clarify the risks and responsibilities associated with Fair Use.
2. Establish that models be considered under Database Laws.
3. Mandate training data manifests for all commercial models.
4. Disgorge models and datasets trained on illegal content (CSAM).
5. Instigate criminal investigations for those involved with CSAM.
6. Apply fines for operators of infringing generative models.
7. Establish expectations of contributory infringement for platforms.

Conclusion

There are two paths ahead in this Copyright debate: one that concedes to Big Tech's short sighted wishes and ultimately leads to the end of Copyright as we know it. This would lead to the closing up of the web as organizations protect themselves in other ways, the disappearance of revenue streams for many worthwhile jobs (like Artist or Journalist), and the loss of all human rights included within data rights. The other path is one that creates a healthy thriving market for data with mutual accountability and respect for all market participants, as well as openness and transparency that also fosters the Arts as well as scientific and technological progress.

Alex Champandard
creative.ai

creative:ai