https://www.federalregister.gov/documents/2023/08/30/2023-18624/artificial-intelligence-and-copyright

1. Generative AI systems have access to a wealth of information not available to any single human, or even groups of hundreds of humans working together. Generative AI's actual access and ability to scan through, analyze, and generate deriviations of known and existing works have already been demonstrated to generate perfect or near-perfect copies of copyrighted source code, and both fictional and non-fictional text documents. As an author of source code (many open and closed-source projects), a technical book, and fictional creative works; generative AI actively threatens my continued ability to work. Further, allowing such generated contents which are near-copies to attain the protection of copyright status flies in the face of all established prior art, fair use, and copyright standards.

2. I have been working in online ads, b2c and b2b online services for 15 years. The content generated from said AI systems is, generally speaking, actively worse in quality and correctness compared to a human editor. As someone who uses search engines, it has become increasingly difficult to find the information I'm looking for, as more and more of the content is either AI generated, or hidden on page 4 after AI generated content. Considering that one of my former spaces was search ads - not getting any relevant search results is actively bad for advertisers, and for the web, as users stop clicking on ads and performing search functions.

3. Fingerprinting generative AI outputs can be done through a variation of https://arxiv.org/abs/2305.20030 . Robots.txt can be modified for opt-in to training / generation via variations of RFC 9309: https://datatracker.ietf.org/doc/html/rfc9309

4. I don't care about what other countries do WRT generative AI. I live in the US, and my copyrights are established in the US. What we decide will lead the world, and should protect my rights worldwide. If we mess up, and somehow decide that a computer should get copyright, I expect the EU to go completely opposite, and for the US to be trailing in protecting its people.

5. Yes. Protect existing copyright holders, no copyrights for generative AI.

Generative AI companies should need to license all content they model from, or

whose content is substantially similar to the source material. Because all generative AI is close to some source material, by the fact that it must be close to all source materials throught the inherent construction of the model. And because all content generated should be associated with some set of closest owners, and those owners should be compensated for their generative works.

6. My source code, and countless other libraries, licensed under GPL 2.0, 3.0 or LGPL 2.0 or 3.0, or MIT, or BSD, etc., all require the inclusion of a license notice with any included portions of written software, to cite the origin of the source code, who wrote it, the license, and your rights to use / redistribute said software. I've not seen a generative AI include the copyright notices from said fully-reproduced software, which makes their production (whether statistical or otherwise) a violation of the license of the software.

Irrespective of those specific reproductions, necessarily, any copyrighted human works should be excluded from any training material, as all underlying generative AI systems have the ability to quote sufficient portions of said works to be in violation of the Fair Use standard. Further, fair use requires a proper references to the source material, and generative AIs (generally) lack the ability to cite the source, making them a system whose whole purpose seems to be to generate content in violation of all existing copyright standards.

6.1. Currently, most modern developers of generative AI systems consume all of the public repositories of information they can get their hands on. From github and bitbucket for source code, to Wikipedia and other sources for general knowledge, to various book archive sites for copyrighted fiction / nonfiction, to all of DeviantArt, Google Photos, Instagram, and more for image-based training. I would wager that less than 1%, perhaps fewer than 1 in 1,000,000 of the authors of said works have agreed to allow them to be consumed for the generation of derivitive works by a computer.

I know I didn't give any of these tools permission to consume and train on my content. I make my content for humans, not for robots. While I also write APIs for robots and humans, my content is strictly for humans.

6.2. Copyrighted works are lifted wholesale without regards to licensing. I've read blog posts from authors who have had entire paragraphs quoted verbatim from their works in ChatGPT and other systems. Works that were published in book form, but which were scanned, uploaded to a piracy site, then used to train

these systems. That violates the original author's copyright, violates US laws with respect to copyrights, and that's just on the reading side of things. To generate verbatim copies, or near-identical with articles swapped (the replaced with an, etc.) shows that these systems were trained on explicitly well-known copyrighted works.

In my personal experience, I've downloaded and run the models for Llama 1, Vicuna, and Alpaca. When asking "factual" questions, 7/10 of the answers I get are either direct quotes (sometimes with bad numbers) or slightly-altered quotes from Wikipedia. For example, asking "How tall is the empire state building?" will lead you to some derivations of the Wikipedia page on the Empire State Building, in all of the generative AI systems I've tested.

6.3. Public domain works are included with copyright works all the same. Variations of Shakespeare are almost as likely to be produced as any other fiction, when prompted properly.

6.4. All training materials are kept by generative AI models after training is complete. Technically speaking, this is because all models are retrained continuously. You start training, that takes some number of days to months to complete. At the same time, you're developing other models, changing the way you think of your problems, possibly even testing smaller models with variations to see if the next training should include more / different content and/or constants / constraints.

But during training, as part of training, you must examine all of your training data. Examining it all is called an epoch, and how "fast" a system learns can be determined by how many "epochs" have passed.

If you throw away that data, you can't train on the same data for another epoch, or a 2nd time to compare two variaitons of the same model, or two different models. So, necessarily, any content that was acquired is either necessarily retained for further retraining, or is re-fetched at further cost to hosting providers / copyright holders.

Some generative AI folks will claim, "we don't keep the data", but they are at best bending the truth. They have transformed the data into "tokens", then they train on that. Don't be confused; if the model trained on a data set X can produce a work contained the the tokens, then the tokens themselves are the original work, or some compressed representation of the work. So they are

retaining substantial portions of copyrighted work without the copyright holder's consent or a proper license.

7.

7.1. My knowledge of this space is somewhat expansive, as I've been involved in the email filtering, search, indexing, and ads space since 2003 (20 years now), which overlaps quite a bit with generative AI, and my interests. I also worked at hCaptcha for about 18 months, learning everything I could about their image analysis (and now image generation) technology.

For processing content to be used in text-based generative AI, these systems process input text / data into a series of tokens (token generation). These tokens may or may not be full words, or individual punctuation characters, but are commonly 1-5 character phonemes or punctuation characters. Upon breaking up the input text into a series of tokens, the text is sometimes overlaid with additional information for some of the tokens (depending on the system). As examples, some will tag when a word was at the beginning of a sentence to note that the beginning of a sentence should have a capital letter, or when a name follows a salutation (Mr, Miss, Dr, etc.). Each system has their own customizations and variations.

Regardless of customizations, tokens are passed into a "training" system whose purpose is to gather metrics on a given token, the tokens preceeding it, and one or more tokens following it. For every token, there is a multi-dimensional space of the tokens that comes before and after. Sometimes this is represented directly as numeric metrics, sometimes the language model itself is embedded in a geometry for easier translation between written languages (see the last 10 years of research on the geometric structure of human language).

In the case of generating content, after training, these systems take whatever "prompts" are provided by the runtime and the user, looks up the symbols in its metrics database, then asks the simple question "what is the most likely token that comes next" based on its metrics. Sometimes the systems include a "seed", which gives you a consistent way to make a choice about what is "most likely", when the choices are simliar, or to make "less likely" choices to make things more interesting.

For example, if a prompt was, "finish this sentence: hello". If the system had been over-trained on computer science work, it would almost certainly return

"world" as the next full word. But with a seed, it might not choose "world", it
might choose "mom".

For processing content into image-based generative AI systems; individual images
or groups of images from videos are chopped into sub-blocks of the image,
transformed using the jpeg pixel order, followed by discrete cosine transform
over the pixel data. Variations of these DCT-pixel blocks with full terms, and
incrementally truncated with fewer terms, are fed into more classical NxM neural
network architecture, for training on inverting the truncated transformed DCTs
into the original full-pixel resolution. Sometimes these systems are made
multi-step; first for generating a low-resolution version of the image, then a
2nd step for generating a high-resolution version, both trained on
DCT-transformed blocks from the input images.

For systems where text is used to generate images, like Stable Diffusion and
others; as part of the training process, every image has an associated text
block description. Part of training is an attempt to associate those words or
tokens with the image training data. This is done, sometimes, by using a second
NxM (different N, M) neural network trained on original description words
mapping to the low / high resolution DCT sub-blocks.

Generally speaking, the copyrighted works are transformed as part of their
tokenization and ingestion process as part of the feed-through into the NxM
neural networks, and/or insertion into metrics / inference databases. Whether a
DCT of transformed pixels, or aggregates over text, data is transformed and used
for training.

7.2. Inferences are (technically speaking) mostly represented by how common
specific groups of tokens / features are associated relative to one another. In
math speak, there is a cluster of points that live close to each other in a
high-dimensional space, and related topics are closer than non-related topics.
One example metric on text that could lead to said clustering is "how many
tokens away do we see these two tokens, on average?"

7.3. For text-based generative AI models, there isn't a good way to unlearn a
concept without removing the concept entirely from the metrics database, or
explicitly including rules that prevent the generation of specific words /
groups of words.

For example, I am sure that most generative AI systems being run on the open web

include profanity filters that prevent either the generation or display of words

that the company considers profane. While this *can* work for individual words,

you sometimes run into issues *because* inferfence isn't actually inferrence.

Like, maybe you can get the computer to not talk about "rape", but then you run

into a potential problem with the computer not knowing what "rapeseed" (canola)

is.

In NxM neural-network categorization models (what is this a picture of?, what is

this text about?), literature has suggested that some concepts or ideas inferred

are the result of "lottery tickets", or some set of Y activation paths through

the N layers, where Y is 1-10% N. That these Y activation paths are dominant in

that they are what actually decide among the concepts.

If it is the case with image generation AIs like Stable Diffusion, that there

are "lottery tickets" that result in the vast majority of outputs, then there

may be some subset of the Y paths that can be explicitly deleted from the model,

which would substantially if not entirely eliminate those specific outputs from

the model.

However, given the importance of the "lottery ticket" activations, and the

literature saying that lottery tickets are important for the "positive" output

of the models results, it would suggest to me that eliminating these specific

lottery tickets would likely have impact beyond the specific topics, which would

not be known until well after.

Shorter version: you may be able to do some text filtering in / out, maybe even

some image filtering; but in practice, a full retrain without problem topics is

likely required.

7.4. It can be possible to determine if a model was trained with a specific

piece of training material. The easiest way is to ask the model questions or

request it generate content related to said specific piece of training material.

Just now, I asked ChatGPT, "Can you quote me some Shakespeare?" It returned,

Â Of course! Here's a quote from William Shakespeare's play "Hamlet":

Â "To be, or not to be: that is the question."

So clearly the platform has been trained with Shakespeare, or at least someone who wrote about Shakespeare. Google Bard takes some convincing, but it quotes just the same.

I just asked ChatGPT about myself, asking "who is <my name>", and it quoted my biography from a website hosting one of my talks.

Google refused to answer about me, but did answer about Bill Clinton.

Llama 2 from Facebook answered similarly to Google.

8. There would be exactly one scenario in which AI models could generate according to fair use, and never be questioned: factual information.

If the model is reproducing factual information about the topic, specific to one
given question, and is not producing language or other "interpreted" concepts, then the information could be considered fair use. For example, if I asked, "how tall is the empire state building?" and it answered:

"roof: 1,250 feet; total with antenna: 1,454 feet"

... that would be fair use.

However, if it replies with:

"The building has a roof height of XXX feet (XXX m) and stands a total of YYY feet (YYY m) tall, including its antenna. "

... with bad numbers, that is not fair use. FWIW, ChatGPT says:

"The Empire State Building, located in New York City, has a total height, including its antenna, of 1,454 feet (443.2 meters). Excluding the antenna, the roof height of the Empire State Building is 1,250 feet (381 meters). It was the tallest building in the world upon its completion in 1931 and remained so until the completion of the North Tower of the World Trade Center in 1970."

Which seems like a variation of the Wikipedia quote:

"The building has a roof height of 1,250 feet (380 m) and stands a total of 1,454 feet (443.2 m) tall, including its antenna. The Empire State Building was the world's tallest building until the first tower of the World Trade Center was topped out in 1970;"

Right now, Google and others are doing similar auto-summarization as part of search results. This prevents the original content owner from receiving that traffic, potentially gaining that client, customer, or even ad impression. This
is actively bad for the web search space, and website advertising outside of search engines, benefitting Google and other entrenched parties at the cost of every existing website owner.

8.1.

Re: Google v. Oracle America

- this case is unrelated and not relevant - Google implemented an interface that emulated the JVM, based on public APIs. This is a computer program, that does what other computer programs do, based on a specification. Even if you were to make the claim that the API specification was copyrighted, you cannot claim that an arbitrary implementation of said specification is also copyrighted. So Google was fine on fair use, because they re-implemented a whole programming language runtime based on a specification of the API.

This was a waste of the court's resources by Oracle from the start. I don't recall speaking with a single software engineer who thought Oracle was in the right here.

Google implemented a spec.

Generative AI producing text or image content based on a prompt is not the same, in any sense.

Re: Andy Warhol Foundation for Visual Arts Inc. v. Goldsmith

- This is the only case with any relevance. Conde Nast violated Goldsmith's copyright by requesting the production of a derivative work without Goldsmith's permission. Both Conde Nast and Andy Warhol violated Goldsmith's copyright, and clearly robbed Goldsmith from the long-term profits owed to Goldsmith that ultimately Conde Nast and Andy Warhol Foundation profited from. This was a problem with the original Warhol derivative work requested in 1984, which persisted until 2023 when the case was decided.

Goldsmith explicitly licensed their work for specific intent, purpose, and cost.

Andy Warhol and Conde Nast violated those copyrights in spirit, and in license.

Many web pages, Wikipedia, personal blogs (like my own), open source software, and more, have explicit and specific licenses about how said works of humans are to be licensed, regardless of the commercial intent. While webpages, etc., are "giving away" their content for "free", there are known and explicit copyrights involved, and proper uses defined.

It's my simple legal opinion that any party using website content, from any source, without the explicit written permission and a license from the copyright holder of said content, for the use of training a generative AI model, is, by definition, a violation of fair use, and a violation of their copyright.

This is simply because the models have the ability to significantly quote verbatim from sufficiently unique creative works, without the ability to cite the source or provide copyrights. Without copyright notices on said duplicated works, or citations of sources from said substantially duplicated works, said generations are simply computer versions of plagarism. Plagarism is either fraud or theft or copyright violation when produced by humans, and it makes zero to negative legal sense to allow a computer to violate copyright and steal from people.

You might say, "but search blurbs!" - when referencing a real webpage, will lead you to the source. Generative AIs generally claim to be the origin source. When I have asked Google Bard factual questions about objects or things in the real world, it sometimes does provide a claimed link or reference to the source material. But of those links, about half have lead to a live webpage (wikipedia), and the rest were dead links (websites that were not available from a regular home internet connection).

Generative AI blurbs, however, are remixing content with the intent of not leading to the destination link, and are a likely violation of fair use in that they prevent the author of the original work from gaining benefit from said work through the "standard" method of web discovery: internet search.

8.2. Information brokers need to abide by copyright terms of the holders. Failure to abide by copyright terms of the holders should result in fines and/or jail time. Anything else is madness.

8.3. Any commercial use must have commercial licenses for all content. If an algorithm worked well on copyrighted data that was unlicensed, surely it should work for data that does not require a license, or is properly licensed? If not, what is so special about copyrighted content that makes it so special for training? And why should we (the people, society, copyright holders in general) continually bear the burden of supplying commercial entities with unlimited

content to remix and resell back to us?

Historically, creative companies paid human beings to produce content to be consumed by human beings. Modern AI companies are suggesting that they should be allowed to take arbitrary creative works, and remix them arbitrarily with computers, again to make money.

What's wrong with paying people to create content? Seems to me that some companies want to take content, not pay for the content, get arbitrary benefit from the content, and ... profit off of human efforts and not pay the humans who did the effort? This is both morally bankrupt, and clearly a violation of copyright holders rights.

We can see in Hollywood right now as executives have been remixing videos of individuals for years. How they want to be able to 3d-scan an actor and reproduce that actor forever. They want to clone a voice and use it forever.

This is what generative AI represents to creative people, artists, actors, writers, software developers, animators, ... someone literally wants to take their job and outsource it to a computer.

8.4. Modern generative AI models consume trillions of tokens, and millions to billions of unique images. This is literally thousands to millions of times more information than one or a hundred creative people can ever process, in order to generate information to "remix" what humans have already produced.

To give context to the scale of this...

In any given year, there are approximately 31.5 million seconds. In a human lifetime, we get around 80 years of life. A single human being, carefully studying, and consuming perhaps 5 tokens per second for their whole lifetime, could consume at most 12.6 billion tokens if they never slept, never stopped to eat food, blink, or say good morning to a coworker.

Training on 1 trillion tokens is the rough equivalent of 79 whole lifetimes of human beings continuously reading 5 tokens each second for 80 years.

While there are a collection of "open" datasets, I suspect few to no modern commercial generative AI system uses only open datasets, as all generative AI models get significantly better with more information. This "more information" comes at the cost of all existing copyright holders on the open web, whether producing source code, fan fiction, or an afternoon doodle. These are consumed wholesale by the web and content spiders feeding the generative AI models, where

generally the only concerns are "can we show this to our users?" (which limits consumption of pornography and similar content by the AIs).

For companies such as Facebook with Llama, Llama 2, etc., their terms of service for Facebook, Instagram, and Threads allows them to directly use public and private postings, comments, pictures, direct messages, and any other content posted to their platform as source material in their training. So they generally do.

In terms of whether volume can / should affect fair use? Yes it should. The more unlicensed copyrighted material you consume as part of your training process, the less claim to fair use you have in any of your platform. Further, the more copyrighted content you have that you didn't license, the more you are saying that this content is actively special, and necessary, for the production of new content. And if copyrighted content is so necessary for the production of new content, then that new content must be a derivative work - because without the copyrighted content, it couldn't have been generated.

If a generative AI company wants to make the claim that their generative AI is truly amazing, they should train it on works that they have rights to. That is a true test. Anything else is marketing bullshit wrapped up with copyright infringement and content theft.

8.5. Let's examine 3 scenarios:

a. Let's imagine that a generative AI produced some new piece of media that is incomparable to what any human has ever created before - like the the collected works of shakespeare - only better.

The impact of said media would be incredible. It would quickly attain best seller status at the bookstore, or best movie, or best musical, or ... whatever. People would be bending over backwards to produce, consume, and jump on this one new bit of computer media.

Then what? That one product has now caused all other human creations to be paler in comparison. That we, the collection of humanity, cannot produce something as good as a computer just ... literally mixing up some bits? You'll crush third a generation of writers, with a third not caring, and the other third trying to

burn themselves out trying to beat the computer.

I'm not interested in this scenario, regardless of how unlikely it is.

b. More likely is that generative AI produces okay to good content, for web pages, and some types of images.

The impact of this is actually hugely substantial. Small, medium, and large businesses are already contracting with generative AI firms to generate web page content (replacing human writers). Entire writing teams of websites have been removed in favor of generative AI articles. I'm looking at you, buzzfeed.com . As we see this move forward, content gets generated better and better, and we won't have kids writing essays anymore. Heck, the tools already can't differentiate between someone who wrote their work from scratch, and someone who fed a document through a rewriting AI.

As the market for, and the technology for generating images and videos get better, we will watch as graphic artists, designers, animators, and others lose their jobs as content writers already have. We'll watch as some companies with money choose to spend it on technology, replacing people, and putting creative people out of work.

This is where we are at now, and it's getting worse. More people are being laid off or not hired due to the current and upcoming AIs being developed.

c. Least likely is that generative AI products don't ever do anything useful, and fail in the market because they are too expensive for what you get.

This is okay. The vast majority of work being done by AI by these AI companies are not going towards curing disease, solving world hunger, or improving the quality of life of people. It's about creating content to be sold to people.

The companies that are working on the types of data that do lead to drug discovery, materials discovery, etc., are generally working from open datasets or genetic data with express permission of the participants (I have worked in one such company).

In practice, for content / media / "entertainment" purposes, providing copyright to a generative AI or the generative AI's creators basically AI steals opportunities from human beings for the sake of cheaper, computer-generated content that generally is only ever, at best, a rewriting or remixing of trained content.

Only when we can talk about the practicalities of generative AI at reducing human misery - curing disease, making it easier to make food, etc., rather than replacing the work that people *want* and *enjoy* doing, like creating artwork, writing software, writing prose - fiction, nonfiction, satire, writing poetry, music, and every other creative endeavor ...

Let's let humans do the fun things, and let computers do the hard and not fun things. Yeah?

Philosophically, that's the right answer, as it reduces human misery, and keeps money in the hands of people who create beautiful things for other people; in any of the arts I described earlier, and ones unlisted and uncountable that are threatened by generative AI.

Further, when we see generative AIs "hallucinate", we are really just seeing that while the metric models that try to define language or DCT inversions are quite good, after training on trillions of tokens, the generation is primarily a random walk through some tokens. When seeing outputs of Stable Diffusion and other image generators (I also have Stable Diffusion running on my local development machine), we can see how outputs can be morphed into one another through the changing of adjacent seed values.

I see no gain to allowing the copyrighting of a computer taking a random walk through a metric space of tokens, or to invert the DCT of some set of pretrained images. That seems ridiculous from a creative perspective. As then I should just generate all X-token phrases, all X-pixel blocks, copyright all of them, and sue everyone for violation of my copyrights.

What makes copyright special is that human beings aren't all taking random walks through metric spaces in the production of their creative works, most if not all of us have intent and purpose in our works.

A computer's only intent and purpose is to follow the instructions of the programmed software as it interacts with data. While we human beings may interact in data processing or game playing on said computer, and many of us are programmers; the computer itself can only follow the instructions provided by the software and data on it. (modulo unintended consequences of software

bugs, or hardware malfunction)

In this sense, the computer is acting, generally, on instructions by a human being to ingest data, process said data, and produce remixed data based on previous human inputs. I don't believe that said remixed content could / should be copyrightable. And so far, of everything that I've heard from Sam Altman, or anyone else in the AI space, not one can provide a good reason why generative AIs should should be able to get a copyright with their productions.

9. Copyright owners should have to opt-in, as this is a new space for nearly everyone. Few, if any, copyright holders have clauses in their licensing contracts about machine-based AI derivations. Historically, anything not covered under an existing contract requires a new, updated, or supplementary contract.

If this system were to start by opt-out, the government would be establishing a precident saying that any rights not previously covered under contract can be exploited by literally any party for profit, until said contracts are updated.

This is well in violation of spirit and letter of the law as it is written and enforced today. The only reasonable solution is requiring copyright holders to opt-into the system, either through licensing their works through a broker, or directly through their own content portals as I describe in 9.2 .

9.1. Let's examine the options and their outcomes.

If we allow researchers to use content arbitrarily, but only require permission when they go commercial, we will observe the non-licensed material leaking through. Always. This is what OpenAI has done with ChatGPT. They have consumed everything to produce the best generative AI they could, but did not license everything, and now want to make a lot of money off of this.

I personally find this morally bankrupt, which is one of the many reasons why I don't use the platform - aside from testing like the above.

Or, we can stick with what the typical standard is: you license the things you want to use, and you stop stealing things and remixing things you don't have the license to. Even music allows remixes, but you always pay the source artist or rights holder, and you commonly get permission first. There are various websites you can go to in order to explicitly license single or multiple songs to do so, for movies, television, video games, playing as a cover band, public

performances, etc.

If we want the AI industry to be as serious as the music business is, we are going to need to build similar websites for licensing content, so that content creators and owners (like myself) are properly compensated for our works. The generative AI tools can wait until they've properly licensed to release. And if their business models can't keep up with that; maybe their business models were not on very good financial basis to begin with.

9.2. robots.txt exists on most websites. But I'd say that we shouldn't opt-out, this should be explicitly opt-in; so that you had to add your robots.txt, include a line that says "you can use content X for purpose Y", perhaps:

User-agent: ai*
Allow: /

To allow content explictly to be used as training material.

You might argue, why not:

User-agent: ai*
Disallow: *

And I would say that the average copyright holder does not want their content to be consumed by these models, so the burden should be on the people who DO want this to opt-in to the consumption by the models, not for the average person to need to opt-out of this thing that:

1. uses website resources to spider / mirror the website
2. then remixes the content of said website at the detriment of website's owner,

as we've already established in 8.1

9.3. Robots.txt was expected to be used by every website owner in history to tell Google and other robots to go away. This was default opt-out because Google was bringing people to your webpage because they indexed it.

However, in this case, a generative AI visiting your website doesn't mean that a human being will ever visit it again. I've spoken with another technical founder of a generative AI + search company who wants to never return an organic search result again; instead generating a web page that answers your question, and keeping any ad money for themselves. Basically what Google is already doing.

For an opt-in robots.txt system for generative AI free access, we can do something simple like:

User-agent: ai*

Allow: /

If its content is to be consumed by AI. Paths can be anything as defined in RFC9309 section 5.1 .

9.4. In the examples I've seen generative AI being used, it is to supplant one or more humans that would have been hired or contracted to do the work instead.

Copyright holders should be entitled to 100% of the revenue earned by AI companies used in the production of their derivative works, plus damages for having their copyrights violated.

If the RIAA can get $1000 for a song being copied by a kid over Napster, I see no reason why any average copyright holder can't get $1000 per work consumed, multiplied by the number of times it was remixed and redistributed to users. So if someone asks a question to an AI that generates a content that is obviously derived from work of fiction X, and that was redistributed 100 times, then the copyright holder would be entitled to whatever the AI company received for producing the content, plus 100 x $1000 redistribution penalty.

Because honestly, generative AI has the power to remix music, images, text, and video to the point where you can ask the computer to generate pleasing disco music like the Bee Gees, and it will. Should that disco music be copyrightable? Should the source that trained it be eligible for licenses if it is in the style of a specific artist?

Without explicit artist permission for source materials, I don't see how you can claim that generative AIs aren't merely a very complicated Napster. One that can actually be tasked with generating fairly convincing duplicates of just about anything.

9.5. Human creators should always be able to opt out of their creations being used by AI, unless they've been found incompetent by a court of law. Regardless of copyright, a person's creations is their voice - it is a part of them forever. Such a system would work as it always has - the rights holder always knows the creator, so they communicate, negotiate any further NEW rights on generative AI and licensing for consumption / reproduction / etc., and we are done.

This is simply contract law, with the understanding that historically, no media

contracts included clauses regarding AI reproduction, as few (if any) foresaw

how quickly generative AI would advance. So if no existing contract has clauses

regarding generative AI use, then that requires a supplemental contract and / or

whole contract renegotiation.

10. Consent can be through robots.txt, compensation can be through bitcoin, eth,

or the new federal token. An address is provided, a cost per page / kilo token

for training is provided, and if the AI agrees, they pay.

One for the cost of training, one for the cost of generating content based on

these tokens. The platform needs to be able to recognize when certain tokens or

token-arrangements are somewhat unique, and when they are used in the production

of content. Any reasonable plagarism detector could work.

User-agent: ai*

Train: 0.01 / ktoken @ COIN_ADDRESS@FEDNOW

Generate: 0.01 / ktoken @ COIN_ADDRESS@FEDNOW

Allow: /ai-trainable/

Which also allows for different train / generation addresses based on url path.

I can think of several algorithms for choosing "closest related works", though I

do have a Phd in theoretical computer science, am a former maintainer of the

Python programming language, have written a book on the Redis database, have

written a language transpiler, and countless other works in the last 16 years of

my career, some secret, some not so secret. If I can come up with several ideas,

I'm sure the generative AI companies can do at least as well.

10.1. Anyone can get a new fednow address, robots.txt is supported by basically

everything, and bad headers are already ignored. The burden of adding clauses to

existing contracts should be on the people who want to use this to generate

content. They will contact people with the content they want to use, they will

properly license it, and we will move on.

Doing anything else is, once again, putting the burden of saying "don't steal my

stuff" on the average copyright holder, instead of on the commercial

organization to just not steal and license the creative works properly in advance.

That we're having this discussion is a bit ridiculous, to be honest. If generative AI companies don't want to license the content that they are basing their entire business on, then they don't need to have a business. That's bullying millions of content creators into being stolen from for the interests of generative AI companies. That's clearly morally bankrupt, and in violation of nearly every single licensing contract in existence

10.2. Collective licensing benefits the generative AI companies more than it does the copyright holders, and it certainly benefits the collective organization structure more than it does the members of said collective.

Which is to say - all management structures necessarily have overhead, which are necessarily inefficiencies compared to direct bargaining with target copyright holders, especially ones providing robots.txt options as I have described above.

If generative AI companies need to understand where to get data, or how to get data, then perhaps they should offer a place where users can submit urls to be tokenized / used, like Google, Yahoo, Infoseek, and just about every other web search company has done in the history of the web. The companies themselves need to properly filter for someone linking Disney's back catalog ripped to a Google drive, and similarly for Shakespear's works.

In the case of public domain media, I believe that there should be a base licensing price provided to any *commercial* entity. Research-only entities get to use the data for free; but upon needing to sell, must pay the train and generation cost, etc.

I believe a distributed licensing scheme that is technically oriented is likely to be sufficient in terms of ability to license, and any laws to be passed should be directed towards generative AI companies on penalties for failing to properly license content (theft), and on individuals or companies who claim to hold the rights to sell without having them (fraud).

If we try to compel copyright holders to license their works, we fall into the pit of Warhol v. Goldsmith again, and the copyright holders will eventually win again. Either individually, or as a class-action.

10.3. No. I don't believe that the profit motivation of generative AI

companies
is more important than the social good of continuing to protect human-made copyrighted works.

I also believe that any compulsive licensing should be compelling generative AI
companies to license in one of the distributed methods described, or through a broker who collects links to said works and rates their "quality", and is paid for references, or directly through collective bargaining (if 500 websites all have the same price, then you could say they are collectively setting a price, even if it is in a distributed manner).

I believe it is simply a matter of providing pricing, and allowing the generative AI companies to pay the price if they are willing.

10.4. No. This primarily benefits the generative AI companies and the licensing
organizations at the cost, once again, to the copyright holders. It also seems to force non-members to have their works remixed and exploited without their knowledge or consent.

For the sake of argument, that's like saying that every 13-year old artistic prodigy needs to also know the laws and how to opt-out of their work being turned into something commercial when they post a picture of their artwork to their Facebook wall.

Why are we talking about letting companies do this to kids, adults, everyone?

At best the discussions here are exploitive of the hard work of millions of hard working writers, artists, software engineers, musicians, speakers, and just
about every other creative person out there.

10.5. No, the licensing scheme works for any data that can be downloaded from an
url from the internet, and fetch an associated robots.txt file. There is one small rights + cost check that needs to be done, but this is an afternoon to a week of work for companies that are already doing generative AI.

There is limited to no burden to this if some simple cost analysis to be done after fetching a robots.txt as defined in my answer to #10 above.

11. In the model you describe, the curator would be responsible for getting copyright holders to agree to let their content be trained on, would collect, and would distribute proceeds from the trainer at the training price, and distribute the proceeds from the user employing the AI system at the generation
price.

If these prices are too high for either the training company or the generation company, then they need to work on their business models to raise more money, or find lower-cost training and generation sources.

Similarly, if there is one party that is the broker, trainer, and generator, then the one company is responsible for directly paying for their use when the content is used.

12. For commonly remixed and re-performed works, it can be difficult to say if one source or if all said remixed topics / works combined to produce the output. Occasionally, there are topics that are unique, and generally produce the same outputs with different seeds, or produce unique groupings of words related to a given topic.

For example, if I were to invent some new english-sounding words and use them in a creative work, then later, if a system were to be asked about those unique words from that creative work, or if the system could be induced to produce those unique words, then you would know where the source material came from.

For images, videos, and audio, this sort of fingerprinting of the source can be substantially more difficult, though I suspect something like the reverse of https://arxiv.org/abs/2305.20030 (fingerprinting the image outputs of generative AI models), where we could watermark images, audio, or text inputs in such a way that anything generative AI outputs on those sources would produce at least one of the unique tokens with medium to high probability. I'm sure this research has not yet been done.

However, image similarity metrics already exist, and image segmentation for both manipulating and finding the "seed" for Stable Diffusion to re-generate your image reliably, already exists. I'm sure the other models have it too, as image-morphing based on these methods rely on taking an image, finding the source seed, then finding adjacent embeddings with nearby seeds with points where the person wants.

In terms of text; you can use any of the text similarity metrics already used for plagarism detection, or use a cosine similarity, or even TF-IDF for for the whole document (generate a vector, embed the vector in a space, dot product).

Turns out that video cards are really good at these things, and text similarity

is a problem that is somewhat solved. I usually find good results with picking some of the least common words from a document and their synonyms.

13. Assuming copyright holders are able to set their prices and derive proceeds directly without overhead; copyright holders would have more money to produce more content, generative AI companies will make less top-line revenue.

If a generative AI company can't find a way to make money off of the data processing of someone else's creation, while also paying licensing rights for those creations, then they are strictly parasites to the existing system, and aren't actually adding value. Because adding value would allow them to charge enough for their services to make the licensing costs worthwhile. Failure to add sufficient value in what they are producing, IE producing CHEAP products, does not help Americans or copyright holders, generally.

14. I think that it's only time before someone invents some words, and the AI produces those words without the author's permission. If those words happened to be coded in such a way to include intentionally-secret content, that could be bad for some company / government.

15. For the sake of repeatability, every company and researcher I've spoken and worked with has kept raw source data, references to where they got it, when the data is likely to be updated again (if ever), any special rules, licenses, etc.

Companies are further required by law to keep references to every software license for all commercial and open source software used, as part of business insurance. This is normally done by IT and/or asset management groups, and has been on several business insurance application forms that I've personally filled out.

I don't see how generative AI companies who are already required by insurance companes to keep track of this, by law to properly license software, that can come back and simply say... this is too hard to do for content. When they've been doing it with software already.

If they can write software to process trillions of tokens into a model, they can write software to pull a robots.txt file, do some math to decide if they should pay the price for the training / generation, then either pay and spider or don't pay and not spider, and keep that record and progress for the perhaps millions

of websites or copyright holders they need to handle.

If this isn't sufficient, and storing payment information for copyright holders
is outside of the realm of reason, then they can talk with a broker, pay them to keep a record of "x people in this group with cost y are ...", and use the broker's service to say "pay the people who were part of the cost Y group at time ...".

This seems like a very simple technical and legal solution, and I'm not sure I've heard any reason from any of these so-called "very smart" generative AI companies as to how or why they can't keep a simple database of rights holders,
prices, urls, and similar. They can get a computer to tell you how tall the empire state building is, can keep a customer database, can spin up potentially
hundreds of machines in a moment... but can't keep basic rights information? I don't buy it.

Creators of training datasets could / should be link farms with price collections and token counts. Something like:

```
. .10 @ F6 # cost of using this domain list to get content, which is shared
           # evenly among all subcontent domains, and this host, equally
https://domain1.com/subcontent .01 @ FC1 | .01 @ FC1 | 3728 ktok | 45 mb
https://domain2.com/subcontent .02 @ FC2 | .01 @ FC3 | ...
https://domain3.com/subcontent .02 @ FC4 | .02 @ FC4 |
https://domain4.com/subcontent .01 @ FC5 | .02 @ FC5 |
```

But hey, I don't want to tell people how to create a marketplace that is:

1. distributed
2. effective
3. profitable
4. likely to get some content producers to dump their stuff

15.1. If software downloads a single image, block of text, snippet of audio, or
any other segment of data with >0 total bytes of space used, that single identifiable piece of information should be able to be paid out to the single copyright holder of that data, if it can be seen that the author's work contributed to the output.

If no one author can be specified, then the group of authors of said works closest in the space (a metric can be determined) should share earnings based on
normalized cosine similarity to the original works. Ambiguity in source should not stop the paying of rights on the generation.

More formally; all source documents (text, image, audio, video) live in a high dimensional space in all of these models. If you produce a document, you can (with work) find out what input documents are closest to the output document, and how close they are. If you pick the closest 10, or closest X where the farthest is at most 2x as far as the closest, or some similar metric, then you can have a selection of "these are relevant close documents likely contributed highly to the generation of this final output, so should get paid / credited."

I've done this exact thing for text documents, and I know the algorithms exist for images and audio, which then applies to video.

If AI companies need help with this, I know a lot of nerds that can help.

15.2. Disclosures should be made to an organization whose purpose is to verify compliance of generative AI companies in their legal duty to compensate rights holders of the content they are generating from. These disclosures should also include proof of payment of copyright holders from whom they have modeled or generated from.

15.3. Users of generative AI should be paying the licensing fees for the content
that is being modeled and generated. Licensing fees and costs should be a part of the model being transferred.

15.4. The cost of keeping the records of the source of the data should be no more than 1/1000 of the cost of training and generating based on that content, in practice. Keeping a source url (maybe 100 bytes) along with destination payment information (perhaps 1 kilobyte) should be substantially less than any meaningful document. Hamlet is about 70kbytes, and any source will include all of Shakespeare's works if it includes one. So if we are to consider sources as at least as prolific as Shakespeare as a baseline, then the cost of keeping a reference to that material, in practice is less than 1/100 of the cost of storing the source material, which 1/100 or less of the the cost of training over the source material, which gets us to about 1/10000 for keeping the rights
on an author's materials about as prolific as Shakespeare, relative to the cost
of training over the source material.

I'd love to hear an explanation as to how or why this cost necessarily needs to
be higher. Becuase I have a real technical solution that would be around this cost for storing this data, and if generative AI companies cannot come up with a
similar or better solution, then perhaps they shouldn't be in the space, or should hire me to build it for them. I'd be willing to rent my Friday's for

the

right price.

16. This should be opt-in, so the notifications would be easily in the form of web logs on the web server that a specific AI spider has consumed from and accepted terms, by continuing to spider, and the user should receive a deposit in the amount relative to the amount of data that was spidered.

For any system, every URL spidered and modeled should cause the author and url to be searchable on their website as a citation. If Google, Yahoo, Infoseek, and

literally every other company that has ever done search can reference source materials, then so can generative AI companies.

There are commercial packages that allow this off the shelf, and several companies will be happy to do this for any generative AI company for a reasonable price.

17. Any information involving a person should likely respect a non-public person's right to privacy. Basically no one in my family is a public figure, except sort-of me, as a result of writing, and my other works. My family should

absolutely be private, I don't know about me. Bill Clinton should be public.

Also, see my answer to 20.1.

18. No. This is information extraction from a metrics database with a random walk. Giving a starting point for the random walk doesn't make it not a random walk. It just starts in a good starting place, to generate the random walk that

the human wants. Starting from specific training content to generate specific output content, again, doesn't make this not a random walk. And in my reading, I'm not seeing any reason why telling a computer to take a random walk through a metrics dataset should be copyrightable.

Re-training a dataset to do specific types of walks just adjusts the metrics, and is not materially transformative of either the source works, or the retrained model.

I don't believe I've been shown or there has been evidence shown to me to make me believe that a generative AI produced work should be copyrightable.

19. I think copyright should be explcitily amended to require that all works to

be copyrighted should be substantially the works of human minds, and not the work of a computer remixing human or other computer creations. That the act of copyright - of creativity itself - necessitates a creative being. That a computer, literally by construction, is not creative, as it can only follow

the
software and data provided by its operator or programmer (in some cases).

Software itself can be written to appear creative, beautiful, interesting, and more. I've written such software. However, the human being is what is truly creative in their efforts, in their creation of both the software, and the art and information that goes into feeding these digital metric systems.

Philosophically, it is a useful appliance, but an appliance none the less. And to give an appliance the ability to create copyrighted material seems out of scope in 2023. AIs don't have names yet, don't have personalities, don't have feelings. People do. People live and die. And people live and die on licensing of their creations, their voices, their faces, their acting, their souls for money to pay rent, buy food, and sustain lifestyles.

Generative AIs don't need to exist, they are an optional creations of companies
with too much money and compute time on their hands. So far, everyrhing I've seen a generative AI do, a human could or should be doing, and being paid for it. But someone wants to have a computer do it instead, for cheaper.

I'm not seeing how copyright holders should be bullied into supporting such a system.

20. I don't think AI generated material should have any real legal protections,
aside from needing to be marked with "generated by AI" so as not to be confused
with legitimate human creation.

No, I don't think that additional protections are necessary. On the contrary, I
think that generative AI companies have been abusing the collective creativity of the world for their own benefit, and that needs to stop.

I think that if we were to say that "if you violate others copyrights, then you
lose your own copyrights to your own software", that would be a very intersting
"tit for tat". I described a penalty scheme in 9.4, but forcing a generative AI
company to go 100% open-source for violating copyright would be a very interesting legal precident. One that I don't think would hold up.

Damages as described in 9.4 would could / should hold up in court.

20.1. Arguably people should have protections *from* generative AI. Let's say, for example, that someone created a generative AI to place images of a person X
inside of pornographic videos. Note that this is possible now, and such tools

already exist. I believe person X should have the ability to take any such fake

imagery down, for the same reasons that you can get taken down any revenge porn.

Because it all seeks to explot an individual person, without their consent, for

the purposes of shame and profit (on behalf of the generative AI companies and

or hosting sites).

Expanding on this, it seems that people truly need to be protected from

generative AI. Not only can it produce fake images, video, text, etc.,

potentially ruining someone's life, but it can steal your job by generating

content like the content you've already written, faster and without sleep.

This is where we are *right now*. It's not going to get better unless we do

something about it.

21. No. The constition applies to people. And as the saying goes, "necessity is

the mother of invention". If people are forced to pay for what they use, then

they will get *better* at using it, more efficient in their methods, and

ultimately innovate better and faster. Sam Altman himself has said that the age

of the giant model is over, and I recently read research showing that training

past the "chinchilla optimal" level for smaller models (7b) for the same amount

of total compute time as a larger model (60b) can end up with a better result.

So if they reduce their dataset, optimize their models so they don't need to

spend as much compute, there should be plenty of money for licensing.

22. It absolutely violates the exlcusive rights clause of copyright, especially

in the case where the generative AIs can be prompted to generate whole

paragraphs and pages from copyrighted works. Or can be prompted to generate an

image nearly indistinguishable from an original with a 64 bit seed. These types

of reproductions violate the copyright holder's exclusive rights to reproduce or

explicitly license. And, to the point, we are discussing how we can license this

to generative AI companies for training and similar generation, and how people

can get paid for their works being generated.

23. I think it is sufficient to assign rights and pay people. So yes, actually

quite easy to say that some work exists or is VERY similar if the language

generated is close enough.

24. Web logs in cases where such are available, and the ability to reproduce

substantial portions of the work with prompting. Sometimes, saying, "please continue this quote", then quoting some of the original document is sufficient to generate the source text.

I think that generative AI companies should be keeping records, and likely already do. If they aren't, then they are actively bad data scientists, or data modelers. As any decent data scientist keeps records, for the sake of updates, or any other of a dozen reasons already mentioned.

25. End user has paid for the service, but doesn't own anything generated, and so has only stolen the consumption of the media, and not (necessarily) the content of the creators unless they try to pass it off as their own.

Anyone who has been paid for a service to train, model, generate, etc., should be liable to pay any copyright holder at the terms when the data was downloaded and trained. Updates to licenses should be checked quarterly.

25.1. At best, "open source" models could be considered research models. Not infringing and okay as long as the any upstream data is also not copyrighted, or as long as the outputs are not being used in a commercial manner.

Being that an open source model cannot guarantee that their models are not being used commercially, they must find sources that are free or cheap enough for their downstream consumers to pay, as they had previously been expecting to pay "0" for the outputs.

Also, they could offer models pretrained on data, for pricing sufficient to pay the upstream data sources for the training itself. This may require initial investment on behalf of the original project owners, but could be made up by additional licensing on generation (we get 10% of what you send upstream to the content generated licensing).

There are many licensing models available for firms willing to pay, even for self-hosted models.

26. It would seem to me that failure to cite the original works, and/or pay for the generating license of the original works that were remixed into the output of the generative AI, is a willful act of removing copyright and/or copyright notice from the work being reproduced. Pretty far in violation of 1202(b).

27. There are always rumors around Hollywood about script X having been stolen from a screenwriter by a shitty friend, or a bad agent. A generative AI is

like

the worst friend you didn't think to worry about. It comes into your website, reads all of your content from whereever it can find it, then it tells other people what you said. Maybe directly, maybe more elegantly. But it tells your stories. And you may not know.

28. Yes. All AI generated content (images, video, text) should be labeled as being AI generated. And individuals should have the chance to choose to consume human-produced media if there is a choice.

Images can be fingerprinted with several different methods (I provided one using tree rings earlier, the arxiv link), and text can be required to include the citations. Including these will be left to the generator + user of these tools, and failure to cite will result in the same $1000 per occurrence as is the case with not paying copyright holders.

Because if it is being played off as being human, but it's AI, then that is fraud.

28.1. Anyone should be able to identify AI-generated works, and submit them to an authority for verification and penalty acquisition. Anyone who finds AI generated works claiming to be human-generated, whose report is verified, and infringing individuals are fined, will be eligible for 10% of the proceeds.

28.2. If the AI can generate it, the AI can associate the generated content with existing documents... like we have been doing for 20+ years with search, cosine similarity, etc.

28.3. The same penalties as not licensing; $1000 per occurrence, plus proceeds.

29. There are several companies that offer tools; some better than others. Some are image-only, others are text-only. I think it ultimately depends on the data that they have to work with, but over time, I believe companies will get better at doing these things, especially if they only need to pay the training cost for identification (you are not generating, you're just comparing the output with a model or direct lookup of all known inputs to you).

30. As far as I know, unless you license your voice explicitly for reproduction or provide it to a model for reproduction explicitly, the generative AI company

doesn't have any rights to your voice. Similarly for your likeness. If a person signs away their rights, then that is a different thing and perhaps should be protected by default (especially with voice signatures being an authentication method).

In terms of actors, actresses, and the exploitation of the creative community; we really do need better protections of all people. For example, the actor who played "Zordon" in the TV series "Power Rangers" was purportedly paid less than $200 for an afternoon of work, and his likeness was broadcast for years for no more compensation.

I believe that we have an obligation to ensure this doesn't happen to actors, actresses, or any member of the public.

For images uploaded prior to this technology (perhaps anything uploaded before January 1, 2024), maybe humans should have the default right that it should not be training material. And anything used going forward must be opt-in.

31. Arguably yes. People should be protected from AI-generated content, especially content intent to harm a person. Satire is one thing, especially when directed towards a public person. But private people need all the protections they can get, as malicious people (1 out of 1000 or less) tend to ruin things for the rest of us, all the time.

I think laws regarding revenge porn should extend to any AI-derived works intended to deride a private individual.

32. Training the AI itself in only a given style can force an AI to generate the style of a given artist, so whether it is language as inputs, or specifically tuned training materials, AI can generate in arbitrary styles. This has been the case since the early 2000's, as undergraduates were making AIs that generate music.

I think that the protections should be in the form of identifying after the fact the closest works, and just directly paying the licensing, or charging for violating licenses if it was not paid, and/or not properly licensed in the first place. You know, that $1000 per occurrence I keep mentioning. That if the RIAA can get $1k for a song, then reproducing a piece of artwork for generative AI should get the same.

33. A generative AI has the mathematical model of the sounds based on

literally

the time + frequency decomposition of the signal. It can reproduce with as high

fidelity as is chosen by the author of said generative AI, and the level of

training done. Arguably, this needs an extended licensing scheme with respect to

existing audio licensing terms, which are already pretty extensive.

Also realize that if we don't do it now, the RIAA is going to have a lot to say

about it in the coming years.

34. Ultimately, I think this is a philosophical question about whether the

copyright office wants to spend its time wading through uncountable masses of

computer and so-called AI generated content in order to defraud the millions of

human copyright holders. Or if we should just protect the citizens and existing

copyright holders of the United States.

I'd love to hear how AI is going to transform anything except content creators

into other careers. How AI is working to take away some of the most important

and human-loved jobs - art, writing, music, movies, software, and more. Why?

Because someone with a computer thinks that they should be able to take the sum

human creative product and remix it, on a computer, for less than the cost of a

human.

And I think that is tragic that the copyright office is considering allowing

that.