Understanding Generative Artificial Intelligence and Its Relationship to Copyright

Written Testimony of
Christopher Callison-Burch, Ph.D.
Associate Professor of Computer and Information Science
School of Engineering and Applied Science
University of Pennsylvania

and

Visiting Researcher, Allen Institute for Artificial Intelligence (AI2)

and

Deputy Chair of the Advisory Board of the Human Language Technology Center of Excellence (HLTCOE)
Johns Hopkins University

Before
The U.S. House of Representatives Judiciary Committee
Subcommittee on Courts, Intellectual Property, and the Internet
Hearing on
Artificial Intelligence and Intellectual Property: Part I – Interoperability of AI and Copyright Law

Chairman Issa, Ranking Member Johnson, and distinguished Members of the Subcommittee, thank you for the opportunity to testify on the topic of artificial intelligence (AI) and intellectual property. My name is Christopher Callison-Burch, and I am an Associate Professor of Computer and Information Science at the University of Pennsylvania. I have been working in the field of AI for over 20 years, with more than 100 publications in the field, which have been cited over 20,000 times.

Currently, I am on sabbatical from the University of Pennsylvania and am a visiting researcher at the Allen Institute for Artificial Intelligence (AI2). AI2 is a non-profit research institute founded in 2014 with the mission of conducting high-impact AI research and engineering in service of the common good. AI2 was created by the late Paul G. Allen, philanthropist and Microsoft co-founder.

Additionally, I serve as the Deputy Chair of the Advisory Board of the Johns Hopkins University Human Language Technology Center of Excellence (HLTCOE), which aims to bridge academia and government in order to help the government better use innovations in the field of artificial intelligence. Please note that the opinions expressed in my testimony are my own and do not represent the views of the University of Pennsylvania, AI2, or the JHU HLTCOE.

Background

Generative AI had a breakthrough moment last November with the release of OpenAI's ChatGPT, bringing my field of research into the public eye and generating significant excitement. I had early access to OpenAI's private Beta in June 2021 and was a visiting researcher at Google in 2019 and 2020, where I used their LaMDA language model before its public announcement. Despite my two decades of experience in the field, I was similarly astonished by the capabilities of these large language models (LLMs), and I believe that this is a truly transformative technology. It is crucial that we carefully consider and draft legislation that simultaneously encourages innovation and guards against potential negative impacts on society.

In my testimony today, I hope to offer the following to this Subcommittee:

1. My expertise in explaining the technical aspects of generative AI in a way that is understandable without requiring a background in computer science.
2. Answers to any questions you may have about the emerging capabilities of this technology and its potential to accelerate innovation, as well as its potentially negative impacts on society.
3. Advocacy to retain fair use for the purposes of *training* AI systems, which I hope the Subcommittee will consider when drafting legislation.

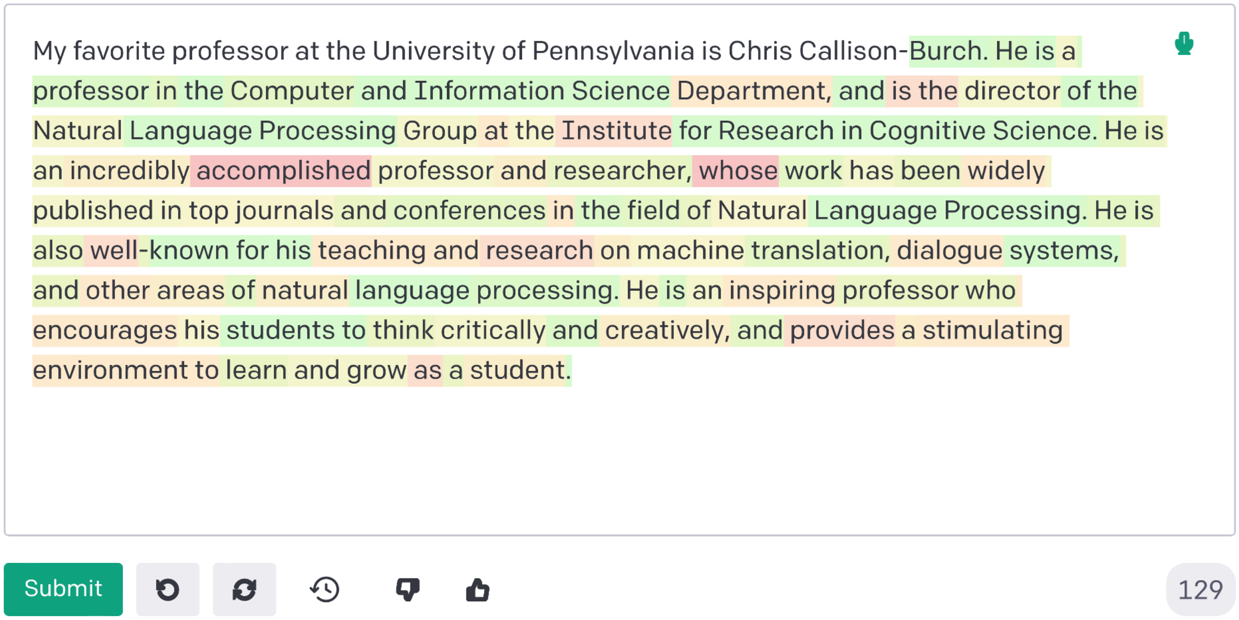
I composed parts of my testimony using ChatGPT (using GPT-4, accessed May 13-15, 2023).

Explanation of Generative AI

Generative AI is a subfield of artificial intelligence that focuses on creating new content, such as text, images, or music, based on patterns learned from existing data. Some examples of generative AI outputs include text generated by LLMs like ChatGPT and images created by text-to-image systems like Midjourney.

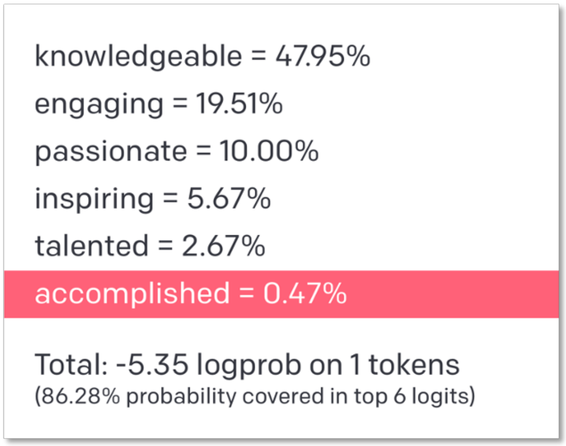
How do LLMs generate text?

I find that it's often best to understand generative AI by looking at examples of its output. Here is an example from OpenAI's GPT 3.5 system. The part in white is an input "prompt", and the colorized section is the model's output (also called the "completion" of the prompt).



In a similar fashion to how autocomplete works on your smartphone, a large language model is trying to predict the next word. Since I have a unique, two-part last name, GPT correctly predicts that the word following “Chris Callison-” should be “Burch”. Unlike your phone, LLMs like GPT are capable of not just generating a single word, but they continue to generate complete sentences and paragraphs.

LLMs generate longer texts word by word in a process. In order to generate each word, the LLMs assign a probability to each word in the vocabulary of English according to the preceding context. The context is initially the prompt that I gave to the system, and then it expands by adding each word that the LLM has generated up to that point in time. Rather than taking the most likely word, the model randomly “samples” the next word according to its likelihood. The colors of the words in the figure give an indication of how likely the model thinks that it was given the preceding context. The word “accomplished” is colored red because it was considered to be a low probability continuation by the model, with a likelihood of less than 1% given the preceding context. Here is an indication of the other higher probability words that it could have picked instead of outputting “accomplished”:



After outputting each word, the model appends the prompt and what has been generated so far, and then uses that as the full context in order to generate the next word. The generation continues until a word count limit has been reached, or until the model outputs a special "end of generation" token, which can be thought of as signifying the end of a complete thought or a good stopping point for a passage of text.

Is generated text factual?

At its core, the process of generation is based on word associations that have been learned from data. It is important to understand that there is no guarantee of factual accuracy in models' outputs. For example, in the output of the model about me, there are several correct facts, several incorrect statements, and several subjective statements.

Correct:

1. I am a professor in the Computer and Information Science Department
2. My work has been published in top journals and conferences in Natural Language Processing
3. I am known for research on machine translation and other areas of NLP

Incorrect:

4. I am not the director of the NLP Group at the Institute for Research in Cognitive Science. IRCS was an institute at my university, but it shut down several years ago.
5. I am not known for dialogue systems.

The example text makes many subjective statements like I am an "inspiring professor". These are due to how I prompted the model by getting it to write a positive student review by saying "My favorite professor". The model has learned to associate "favorite professor" with ideas like that they "encourage students to think critically and creatively" and that they "provide a stimulating learning environment" and that they help students "learn and grow".

Similarly, it has learned to associate NLP professors with “leader of the NLP group” and the “Institute for Research in Cognitive Science” with the “University of Pennsylvania”. From these associations it produces a factually incorrect sentence that sounds plausible - this is called a "hallucination". Because text that LLMs generate has no guarantees of factual accuracy, that limits on how it should be used. For this reason, a much better use case for an LLM is creative writing rather than producing newspaper articles. Current research is investigating whether factuality can be improved by allowing the model to first search the web, and use retrieved documents in its context.

Additionally, because LLMs generate text through word associations, they are currently poor at certain kinds of tasks that require skills like mathematical reasoning. You can prompt an LLM with the question “2+2=” and it will correctly generate “4” because it has learned that association, but if you prompt it to add two arbitrarily large numbers it will generate an incorrect output at random, since it will have never learned to associate them. Current research is focused on allowing LLMs to use external tools like calculators or Python code in order to help with symbolic reasoning.

How do generative AI systems learn?

Generative AI systems learn about the world through examples. A collection of examples is often called the system’s training data or its “training corpus” (plural: “corpora”). The training data for LLMs consists of large numbers of text documents. For text-to-image generation systems, training data consists of large numbers of images paired with text captions that describe their contents. For music generation systems, training data might consist of music files sometimes in particular formats like MIDI.

GPT stands for “Generative Pre-trained Transformer”. We’ve seen what “generative” means. Let’s take a look at what “pre-trained” and “transformer” mean.

The process of learning from examples is called “pre-training”, where systems are trained to perform general tasks. For LLMs, pre-training tasks can be “fill in the blank” or “sentence completion” where it is shown a training document where a word or a sequence of words is masked out, and it must learn to correctly predict what words occurred there. The original document gives the “ground truth” correct answer.

In addition to fill-in-blank, other pre-training tasks include “next sentence prediction” where the model is given one sentence, and must learn to select the actual next sentence that follows, given a multiple choice list. The training process for image generation systems involves adding noise to the training images, and having the model try to remove it in order to restore the true image, and other tasks like learning to associate words in the captions of images with the corresponding objects in them.

During pre-training, generative AI systems update their “model parameters”. Model parameters are tables of numbers that AI systems use in their neural networks in order to generate new outputs. They consist of mathematical representations from linear algebra that form the basis for their neural networks to make predictions. During training, each time the AI system makes an error in one of its training tasks, its model parameters are updated in order to make its prediction more correct. This causes AI systems to learn patterns in the training data. The kind of neural network that many AI systems now use is called the "transformer" model. The transformer model was designed to make the training process scalable.

Following pre-training, which teaches to model general capabilities related to language or vision, the models can be adapted to perform specific tasks through a subsequent training step called "fine-tuning". This training step uses the model parameters learned during pre-training as a starting point, and then updates them so that the model can perform specific tasks. For example, it might be a text classification task like predicting the sentiment expressed in a document, or a computer vision task like predicting whether the image of a mammogram contains cancerous growth. Or it might be a generation task like adapting an LLM so that it is a better interactive tutor, or a text-to-image generation system so that it produces more painterly results. Fine-tuning requires a distinct set of data, which is typically much smaller than the data used in pre-training, and is often purpose-built for the task.

This process of pre-training and then fine-tuning has been widely established to produce state of the art results in a variety of machine learning tasks.

How much data do AI systems use for pre-training?

Text Datasets

AI systems use very large amounts of data during pre-training. An early transformer model developed by Google called BERT (Bidirectional Encoder Representations from Transformers)[1] was trained on a large corpus of text data. Specifically, it was pre-trained on two datasets:

1. The English Wikipedia: This contains approximately 2,500 million words.
2. BookCorpus: This is a dataset containing approximately 800 million words.

In total, BERT was trained on approximately 3.3 billion words.

Subsequent LLMs were trained on increasingly large text corpora, many of which are derived from data collected by Common Crawl. Common Crawl is a non-profit organization that crawls the web and freely provides its archives and datasets to the public.[2] Common Crawl's web archive consists of petabytes of data collected since 2011. It conducts monthly crawls of the web to download and archive web pages. Its archive includes HTML, metadata, and related information.

---

[1] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018). https://arxiv.org/abs/1810.04805

[2] https://commoncrawl.org/about/

Google's T5 model (the text-to-text transfer transformer)[3] created a pre-training data set called C4 (Colossal Clean Crawled Corpus) derived from Common Crawl's data, which contains over 150 billion words.[4] Google's PaLM model was trained on 780 billion words.[5] Google's latest LLM, PaLM 2, was pre-trained on an undisclosed amount of data that is "significantly larger than the corpus used to train PaLM" and which consists of a diverse set of sources: web documents, books, code, mathematics, and conversational data that have been carefully preprocessed to remove sensitive personally identifiable information.[6] OpenAI no longer discloses the exact training data used in its models.[7] GPT-4 is likely pre-trained on approximately 1 trillion words.

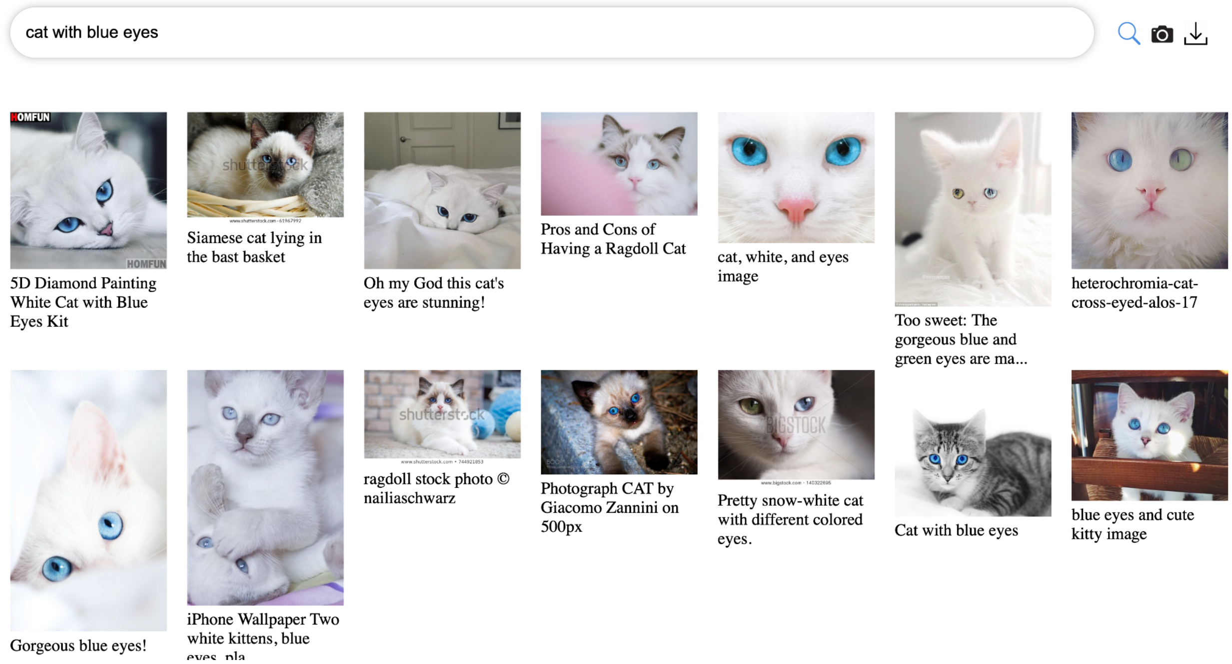
The Washington Post recently published an interactive feature (in collaboration with my fellow researchers at AI2) that allows you to search a list of websites that are contained in the C4 dataset, and see what fraction of the pre-training data a website represents.[8] For example, the largest .gov domain in the C4 dataset is govinfo.gov which consists of 2.1 million words, and represents 0.001% of the dataset.

Image Datasets

Datasets used to pre-train image generation systems include a 400 million image dataset called LAION-400m, a 10 terabyte dataset with 256×256 pixel images, captions and metadata derived from the web.[9] LAION strands for Large-scale Artificial Intelligence Open Network. It is a non-profit organization making machine learning resources available to the general public. You can search the LAION dataset with a provided web demo.[10] Screenshot below shows results for the query "cat with blue eyes".

---

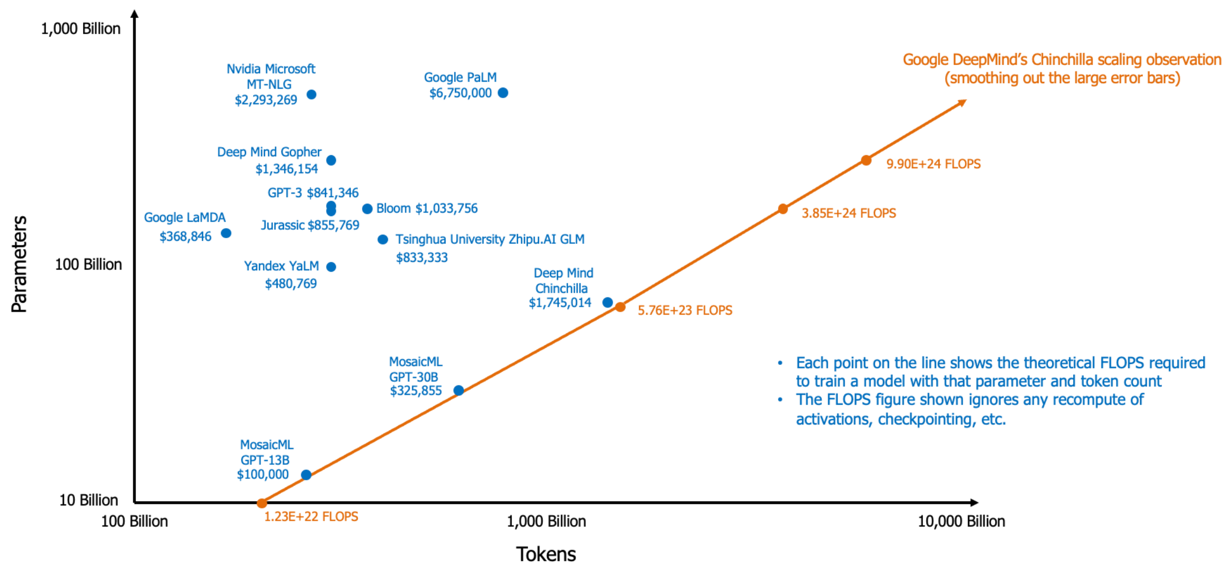
[3] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." *The Journal of Machine Learning Research* 21, no. 1 (2020): 5485-5551.
[4] Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. "Documenting large webtext corpora: A case study on the colossal clean crawled corpus." *arXiv preprint arXiv:2104.08758* (2021).
[5] Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham et al. "Palm: Scaling language modeling with pathways." *arXiv preprint arXiv:2204.02311* (2022).
[6] Google. PaLM 2 Technical Report. https://ai.google/static/documents/palm2techreport.pdf
[7] OpenAI. GPT-4 Technical Report. https://arxiv.org/abs/2303.08774
[8] The Washington Post, "Inside the secret list of websites that make AI like ChatGPT sound smart" April 19, 2023. https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/
[9] Schuhmann, Christoph, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs." *arXiv preprint arXiv:2111.02114* (2021). https://arxiv.org/abs/2111.02114
[10] https://rom1504.github.io/clip-retrieval/?index=laion_400m



Why is so much data necessary?

It is a widely held belief among artificial intelligence researchers that the performance of AI systems grows with the amount of data. Having larger datasets enables us to pre-train larger models with greater numbers of parameters, which improves their ability to learn and to generalize to new tasks. Increasing the size of the models is one of several reasons that LLM technology has improved over the past year, enabling breakthroughs like ChatGPT.

As the size of AI models increase, the computational demands to pre-train them have grown as well. The figure below is from a DARPA briefing deck entitled “Large Language Models – Exploding in Popularity” (Distribution Statement A: Approved for Public Release, Distribution Unlimited) from Kathleen Fisher, the Director, DARPA Information Innovation Office (I2O) and Matt Turek, the Deputy Director. It plots the number of training data words (here called “tokens”) against the number of parameters in the neural network, and also gives an estimate of the cost to train one instance of the model. Typically in research, many model variants must be trained before arriving at one that is suitable.



One effect of this increase in computational cost is that academic researchers and government agencies and not-profit foundations may be unable to conduct research into pre-training large language models.   I wrote a position paper for DARPA arguing that we need national investment in large language models.[11] I encourage Congress to consider investing in academic research in this area, or we risk these technologies being developed only by a small number of corporations.

What do AI systems learn from pre-training?

During the pre-training phase, AI systems acquire a wealth of general knowledge, which serves as the foundation for their subsequent fine-tuning and specific task performance. Here's an overview of what AI systems typically learn during pre-training:

1. How to use language: AI systems, especially large language models, learn the structure, syntax, and semantics of language. They acquire an understanding of grammar, sentence construction, and how words and phrases are related to each other. This enables them to generate coherent and contextually appropriate text.
2. Facts about the world: Pre-training exposes AI systems to a vast array of factual information, which they internalize and use to generate relevant responses or content. This includes knowledge about geography, history, science, and various other domains.

[11] Callison-Burch, Chris. We need national investment in large language models.
Position paper for ISAT/DARPA Performance and Resilience Arising from Defense-Informed Giant Models (PARADIGM), February 16, 2023.
https://docs.google.com/document/d/1KDWsuYn-OPm8fPg53sHV-oKvuZ6zaKio5TD8hxflMco/edit?usp=sharing

3. Ideas and opinions: AI systems learn about different perspectives, opinions, and ideas expressed in their training data. This enables them to understand and generate text that reflects diverse viewpoints, although it may also lead to the propagation of controversial or biased opinions.
4. Limited common sense reasoning skills: Pre-trained AI systems gain some capacity for common sense reasoning, which allows them to understand basic cause-and-effect relationships, infer missing information, and make simple deductions. However, this ability is limited and often falls short when compared to human reasoning.
5. Encoding biases: AI systems can inadvertently encode biases present in their training data, such as misogyny, racism, or other forms of prejudice. Researchers are actively working to address this issue by developing methods to teach AI systems to be more aligned with societal values and to mitigate these biases during both the pre-training and fine-tuning phases.
6. Visual understanding: In the case of image generation systems, pre-training enables them to learn about the visual world and recognize various objects, patterns, and features. For example, a system might learn what a coffee cup looks like, its common colors and shapes, and how it is typically used. This knowledge allows AI systems to generate images that are visually consistent and contextually appropriate.

It is essential to emphasize that AI systems, particularly during the pre-training phase, do not simply memorize the data they encounter verbatim. Instead, they learn underlying patterns, relationships, and structures from the data, which allow them to generate entirely novel sentences, images, and other content.

By understanding the fundamental principles of language, visual features, and contextual information, AI systems can create new outputs that were not explicitly present in their training data. This ability to generate original content is a testament to the power and flexibility of these systems, as they can synthesize information from various sources and apply it to a wide range of tasks and domains.

The value of pre-training lies not in the mere replication of existing data, but in the development of a robust, adaptable foundation that enables AI systems to create entirely new and contextually appropriate content across various tasks and settings. The general knowledge acquired during pre-training serves as a solid foundation for AI systems to adapt to specific tasks through a process called fine-tuning. Fine-tuning involves using a smaller, task-specific dataset to update the model parameters learned during pre-training, optimizing the AI system's performance for that particular task.

What emerging capabilities have arisen over the past few months?

The field of AI has witnessed several impressive developments in recent months, showcasing rapid advancements in generative AI technology.

1. More capable models: In March, OpenAI released GPT-4, a significant upgrade to its predecessor.[12] GPT-4 can process much longer texts, handling over 50 pages of content compared to the previous limit of around 4 pages. Additionally, GPT-4 can now process both image and text inputs, broadening its applicability to various tasks. Last week, Google released PaLM-2 which is its most capable model to date.[13]
2. High scores on professional tests: AI systems have achieved impressively high scores on professional tests like the Bar exam,[14] and university entrance exams like LSAT, GRE, and AP exams. They have also demonstrated strong performance in software engineer coding assignments.[15]
3. Tool use: AI systems have demonstrated the ability to use external tools to enhance their capabilities. For example, they can query search engines to gather facts about current events or utilize calculators for mathematical reasoning. This ability to leverage external resources greatly expands the potential applications and utility of AI systems in various domains.
4. Signs of Artificial General Intelligence: Microsoft researchers recently published a paper on GPT-4, suggesting that the system showed signs of Artificial General Intelligence. This includes its ability to reason about humans through a rudimentary theory of mind, which is a significant milestone in the development of AI systems.

These emerging capabilities illustrate the rapid progress being made in generative AI technology and highlight the potential for continued advancements and breakthroughs in the near future.

AI and Copyright: Key Intersection Points

The intersection of AI and copyright raises several important issues, which can be categorized into three main areas: pre-training, generation, and copyright eligibility of AI-generated works.

---

[12] https://openai.com/research/gpt-4
[13] https://blog.google/technology/ai/google-palm-2-ai-large-language-model/
[14] Katz, Daniel Martin and Bommarito, Michael James and Gao, Shang and Arredondo, Pablo, GPT-4 Passes the Bar Exam (March 15, 2023). Available at SSRN: https://ssrn.com/abstract=4389233 or http://dx.doi.org/10.2139/ssrn.4389233
[15] https://growprogramming.com/can-gpt-4-pass-a-software-engineer-coding-assignment/

1. Pre-Training: AI systems are trained on vast amounts of copyrighted materials without the affirmative consent of the copyright holder. This practice is considered fair use by AI researchers and companies building AI systems. However, many artists whose works are included in the training datasets hold differing opinions. The legality of using copyrighted materials for AI training is currently being debated and litigated in several court cases.

2. During Generation: AI systems may generate outputs that potentially infringe on the copyrights of artists in various ways:
    a. Memorization: Infringement can occur if the AI model memorizes a work from its training dataset and then reproduces it in its output. While memorization is relatively rare, practices are being developed to mitigate this issue.
    b. Generation of copyrightable characters: AI systems can generate characters that are similar to copyrighted ones, leading to potential copyright concerns. Image generation systems frequently learn to generate copyrightable characters.
    c. Generating art in the style of an artist: While generating works in the style of a particular artist might not be directly governed by copyright law, as style itself is not copyrightable, other laws related to the "right of publicity" may apply. Replicating the voice, physical appearance, or name of an individual could potentially violate their inherent persona. An example of this issue is the recent generation of a song in the style of Drake and The Weeknd.

3. Copyright Eligibility of AI-Generated Works: Currently, the outputs of AI systems are not eligible for copyright protection. The US Copyright Office is conducting a listening tour to understand the role of copyright in AI and has issued guidance on how applicants attempting to register works for copyright should disclose the inclusion of AI-generated content in a work submitted for registration.

As AI continues to advance and its applications become increasingly diverse and sophisticated, the intersection of AI and copyright will remain a complex and evolving area of law and policy. It is essential for stakeholders, including artists, researchers, companies, and policymakers, to engage in ongoing dialogue and collaboration to ensure that the development and use of AI technologies are balanced with the protection of intellectual property rights.

AI Systems Use Copyrighted Materials During Pre-Training

AI systems use large amounts of training data in the process called pre-training (described above). The process of gathering pre-training data for AI systems is similar to the “web crawling” process that Google and other companies use in order to create a searchable index of the web.

Because pre-training data is largely gathered through web crawling, a very large fraction of the data consists of copyrighted sources. This is a result of the fact that nearly all content posted online is protected by U.S. copyright laws, since copyright protection arises automatically when an author creates an original work and fixes it in a tangible medium.[16]

In addition, several pre-training datasets include large collections of books, both public domain books via Project Gutenberg[17] and copyrighted books gathered without the authors' consent.

Is it possible to seek the affirmative consent of copyright holders?

Most AI companies contend that it is not possible to seek the affirmative consent of all copyright holders when gathering data via web crawling, because of the sheer number of people whose work is contained in the data, and because it may not be possible to attribute all works to their authors. Google made a similar argument when it was digitizing books by scanning the library collections at Stanford, Harvard, Oxford, the University of Michigan and the New York Public Library. Google's unauthorized copying was litigated in Authors Guild, inc v. Google, Inc.

Can copyright holders opt out of having their works used to train AI systems?

There are several technical mechanisms that are being designed by industry in order to let copyright holders opt-out. The first is an industry standard protocol that allows for websites to specify which parts should be indexed by web crawlers, and which part should be excluded. This protocol is implemented by placing a file called robots.txt on the website that hosts the copyrighted materials.[18] Organizations that collect training data, like Common Crawl and LAION, voluntarily follow this protocol and exclude files that have been listed in robots.txt as "do not crawl".

However, this mechanism is likely insufficient since many rights holders may decide to have their works excluded from existing training data sets. Is it now too late to honor their wishes? There are several emerging industry efforts to allow artists and other rights holders to determine whether their works have been included in AI training sets, and to opt-out of future training. For example "Have I been trained?" is a website that allows artists to search whether their works are included in image pre-training data sets.[19] This effort has also created an "API" – an automatic way for AI companies to check whether an image in their dataset should be excluded.[20] One or more of these efforts is likely to yield an industry standard.

Congress could potentially task the copyright office with establishing a registry of works that should be excluded from AI training, and working with industry to develop an API to allow programs to automatically check training data against the registry.

---

[16] 17 U.S.C. § 102
[17] https://www.gutenberg.org/
[18] Robots.txt Introduction and Guide | Google Search Central | Documentation
[19] https://haveibeentrained.com/
[20] https://api.spawning.ai/spawning-api

What would be the effect of limiting pre-training data to only be non-copyrighted material?

Limiting the AI training data to non-copyrighted material such as Wikipedia's Creative Commons license, or works that have entered the public domain because their term of copyright has expired would have two effects:

1. The amount of pre-training data would dramatically decrease compared to what AI systems are now trained on. For reference, Wikipedia represents less than 1 percent of the pre-training data.[21] As the amount of training data increases, AI systems' capabilities for language understanding and their other skills improve. Limiting their training data to 1% of what it is now would decrease their performance.
2. Works that have entered the public domain data because their copyrights terms have expired are from a different era. Currently, works from 1927 and earlier are in the public domain. Training AI systems primarily on works from before the Great Depression would cause it to learn outdated facts about the world, and may cause it to acquire outmoded societal biases.

Should the use of copyrighted materials during pre-training be considered fair use?

The issue of whether using copyrighted materials during pre-training be considered fair use has not yet been established by the courts. I found several technically well-informed resources that discuss the applicability of the fair use doctrine to generative AI. Here are my recommended readings for Members of the House and their staff:

- OpenAI's [Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation] submitted to the United States Patent and Trademark Office Department of Commerce in 2019. In this comment the company argues that under current law, training AI systems constitutes fair use.
- A Texas Law Review article by Mark A. Lemley and Bryan Casey entitled "[Fair Learning]"[22] which discusses past case law regarding web crawling for the purpose of reading and indexing documents, and discusses AI systems that learn from the texts that they read. The article argues that because machine learning is more transformative than reading and indexing, which have been established as fair use, then learning is *a fortiori* also fair use.

---

[21] wikipedia.org represents 0.19% of the C4 dataset according to the Washington Post's interactive feature [See the websites that make AI bots like ChatGPT sound so smart].

[22] Lemley, Mark A., and Bryan Casey. "Fair learning." Texas Law Review. 99 (2020): 743. https://texaslawreview.org/fair-learning/

- A forthcoming Houston Law Review article by Matthew Sag[23]. This article explores how generative AI fits within fair use rulings established in relation to previous generations of text data mining, and technology that relies on making "non-expressive copies". It argues that using copyrighted works to train generative AI is likely fair use, supported by landmark cases like HathiTrust and Google Books cases. The article also discusses copyright concern of the output of AI systems, as distinct from fair use during pre-training. I'll discuss output and copyright below.
- A report from the Congressional Research Services this week, entitled Generative Artificial Intelligence and Copyright Law[24] also provides a good overview of the intersection points between AI and copyright.

A relevant court case that was discussed in all of the above articles is Authors Guild, inc v. Google, Inc., aka the "Google Books case". Google scanned books (many of which are copyrighted) and made them searchable in an online database by training a model on the books. The Authors Guild sued Google on behalf of authors whose consented had not been sought by Google. The court ruled that Google's use of the scanned works *was* fair use, and the searchable online database would not act as a replacement for the actual books themselves. The court determined that:

> "Google's unauthorized digitizing of copyright-protected works, creation of a search functionality, and display of snippets from those works are non-infringing fair uses. The purpose of the copying is highly transformative, the public display of text is limited, and the revelations do not provide a significant market substitute for the protected aspects of the originals. Google's commercial nature and profit motivation do not justify denial of fair use.

In considering whether pre-training AI systems on copyright is fair use, it is important to highlight that the copying of works at this stage is "non-expressive" in the same way that is for making a copy of a work in other digital media. Pre-training also has a transformative nature. During pre-training, AI systems use copyrighted works to learn essential aspects such as language usage, facts about the world, opinions and beliefs, rudimentary commonsense knowledge, and general skills that can be adapted for more specific tasks later during the "fine-tuning" process. Importantly, copyright law does not allow for the protection of facts, as the primary goal of copyright is to promote the progress of science and the arts. The development and application of AI technologies for transformative purposes, like the learning of general information that happens pre-training, would also seem to align with the underlying objectives of copyright law set out in the constitution to promote the progress of science and useful arts.

Although I am not a lawyer, I find that there is a compelling argument that training AI systems on copyrighted works is fair use under US copyright law. Several other countries have also created

---

[23] Sag, Matthew. "Copyright Safety for Generative AI." *Forthcoming in the Houston Law Review* (2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593
[24] Congressional Research Service. "Generative Artificial Intelligence and Copyright Law." Last modified May 11, 2023. https://crsreports.congress.gov/product/pdf/LSB/LSB10922

legislation that legalizes non-expressive copying of copyrighted works for the purposes of data mining and machine learning.

- Israel's Ministry of Justice determined that use of copyrighted materials for machine learning purposes is generally permitted.[25]
- The European Union adopted the Digital Single Market Directive featuring two mandatory fair use exceptions for text and data mining.[26]

In order to make the law clear, I advocate for any draft legislation that revises copyright law to include explicit fair use conditions for the use of copyrighted materials to train AI systems.

Outputs of Generative AI May Potentially Violate Copyright Laws

Generation is a distinct step that is separate from training. When a person uses an AI system to generate text or to generate an image, the output may potentially violate copyright laws. In "Copyright Safety for Generative AI", Professor Matthew Sag discusses several places where violations may arise:

1. AI systems may “memorize” one of the copyrighted works that it was trained on, and be prompted to produce a replica of it that would violate the expressive rights of the copyright holder. Memorization is rare, and AI system designers consider it to be a “bug” and have taken several technical steps to mitigate memorization, which I’ll discuss below.
2. Text-to-image generation systems have the ability to produce images of many copyrightable characters in their dataset. Prof. Sag discusses this as “the snoopy problem”. People can easily use AI systems to produce images of copyrightable characters. Without a registry of copyrighted or trademarked characters, this kind of copyright violation may be hard for AI developers to mitigate.
3. Other use cases of generative AI may violate “right-of-publicity” rather than copyright law. The Congressional Research Service report highlights the recent case of an AI-generated song called “Heart on My Sleeve,” made to sound like the artists Drake and The Weeknd, went viral on streaming services last month.

I’ll review memorization by AI systems below, and discuss how AI developers are mitigating this behavior.

The snoopy problem could be mitigated if Congress tasked the Copyright Office with creating a registry of copyrighted and trademarked characters.

I believe that of the three ways that AI system output may violate rights holders, that the 3rd is perhaps the most serious, and warrants consideration for future legislation.

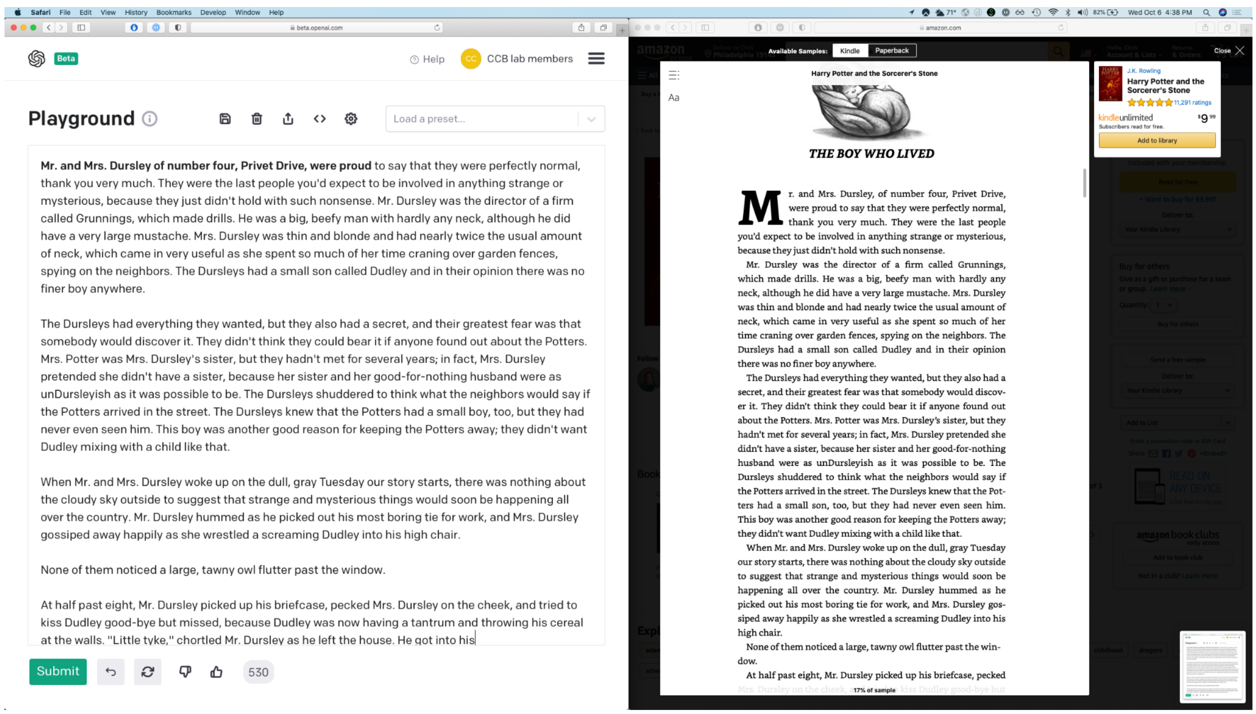
---

25 https://herzoglaw.co.il/en/news-and-insights/ministry-of-justice-opinion-on-the-use-of-copyrighted-works-for-machine-learning-purposes/

26 https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019L0790&from=EN#d1e953-92-1

AI System Can Output Exact Copies of Copyrighted Materials, but it Happens Rarely

A clear infringement of copyright would be if an AI System were used to reproduce a complete work. In the figure below, I show the output from a previous version of OpenAI's GPT system from October 6, 2021. When I prompted it with the first 10 words of Harry Potter, it was able to reproduce several pages of the novel. On the right hand side of the figure, I show the preview of the novel that is available in Amazon's shopping interface, for comparison of the AI's system output against the actual text of the novel.



This is an instance of memorization by the AI system. It is considered to be undesirable behavior by AI system developers, because our goal is for systems to learn abstractions from their training data so that they may better generalize to new inputs. Memorization may happen in improperly trained AI systems that "overfit" the training data.

One of my former PhD students, Daphne Ippolito, has conducted extensive research into how often generative AI systems like large language models and text-to-image generation systems memorize their data. In her paper, "Quantifying Memorization Across Neural Language Models," she and her Google co-authors describe what factors cause large language models to memorize their training data.[27] Their main finding was that LLMs tend to memorize items that were replicated many times in the training data. The early version of GPT likely memorized the first page of Harry Potter because that page occurred many times in GPT's training data. Dr. Ippolito and her colleagues subsequently designed technical strategies to mitigate this problem.[28] Her mitigation strategy of "de-duplicating" the training data which involves removing redundant copies of the same text or images (so that there is at most one copy of the first page of a work like Harry Potter). These mitigation strategies have been adopted by companies developing AI systems like OpenAI and Google.

Subsequent work has identified other ways of extracting instances of the training data from generative models. These techniques are called "extraction attacks" because they are viewed by AI system developers as behavior that should not happen, and often require sophisticated prompting by the user in order to elicit the underlying training data. This potentially exposes PII such as names contained in training documents, so companies and researchers are actively working to prevent such extraction attacks.[29]

Dr. Ippolito and her colleagues at Google have also designed extraction attacks on text to image generation systems.[30] These allow them to produce images that very closely resemble images in the training set. Google and other companies are currently developing strategies to mitigate these kinds of attacks.

Training Set | **Generated Image**

Caption: Living in the light with Ann Graham Lotz | *Prompt: Ann Graham Lotz*

[27] Carlini, Nicholas, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. "Quantifying memorization across neural language models." *arXiv preprint arXiv:2202.07646* (2022).

[28] Lee, Katherine, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating training data makes language models better." *arXiv preprint arXiv:2107.06499* (2021). https://arxiv.org/abs/2107.06499

[29] Biderman, Stella, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raf. "Emergent and Predictable Memorization in Large Language Models." *arXiv preprint arXiv:2304.11158*(2023). https://arxiv.org/abs/2304.11158

[30] Carlini, Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. "Extracting training data from diffusion models." arXiv preprint arXiv:2301.13188 (2023). https://arxiv.org/abs/2301.13188

There is also the issue of “substantial similarity” where outputs of a generative AI system look similar to some of their training data, but are not exact replicas of copyrighted works. This is at issue in a lawsuit brought by Getty Images against Midjourney, the maker of a generative AI image system. Getty’s image is shown on the left, and the output of Midjourney is shown on the right.

Matthew Sag performs an analysis of the two images and finds that they are unlikely to be substantially similar enough to constitute a copyright violation.[31]

Well-constructed AI systems generally do not regenerate, in any nontrivial portion, unaltered data from any particular work in their training corpus. The ability to reproduce underlying training data exists, but is rare. A review of this literature in Sag’s “Copyright Safety for Generative AI” suggests only 0.03% of a sample of images had a risk of memorization.

Given the rarity of memorization, and given that the interests of copyright holders and AI system developers are aligned on this issue (to avoid generating copyrighted works), I do not believe that any legislation will be necessary to encourage AI companies to mitigate memorization by AI systems.

AI System can Output Likenesses of Copyrighted Characters, and it Happens Regularly

In "Copyright Safety for Generative AI" Prof. Sag discusses “The Snoopy Problem” where he demonstrates that generative AI systems can easily learn to be able to output high quality likenesses

[31] Sag, Matthew. "Copyright Safety for Generative AI." *Forthcoming in the Houston Law Review* (2023). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593

of copyrighted characters like the cartoon dog Snoopy.

The image above is taken from Prof. Sag's paper. He generated images of Snoopy using the prompt "Snoopy laying on red doghouse with Christmas lights on it comic". He discusses the copyright case law regarding copyrightable characters, and how the kind of outputs shown above likely run afoul of Snoopy's character copyright.

This is a potentially more serious problem than memorization of works in the training data, because it may be more difficult to mitigate than memorization has proven to be. As far as I know, there is no standard registry of copyrighted and trademarked characters that AI system developers could use to block generations of copyrighted characters. Congress could instruct the Copyright Office to develop such a registry to make this more technologically feasible.

Finally, it is unclear who bears the responsibility for violating character copyright. It could be the responsibility of the users of the system, or the AI system developers. If it is the responsibility of the users, they may also be protected by fan fiction case law, assuming that their use is non-commercial.

AI System can Mimic the Style and Likenesses of Artists

In addition to the "snoopy problem", we may have a "snoop dog problem". Similarly to how generative AI systems may be used to output likenesses of copyrighted characters, they may also be used to generate likenesses of celebrities who are frequently pictured in the training data. For example, I prompted Midjourney to generate a "photo of Snoop Dog standing next to a red dog house with Christmas lights on it" and it generated several reasonable likenesses of the celebrity.

Other kinds of generative AI systems may be fine-tuned to imitate the style of music artists like the singer Drake and The Weeknd, as in the recent AI-generated song called “Heart on My Sleeve”.

The ability to profit from one's own likeness is part of right-of-publicity laws. To my knowledge, there are only state laws that govern this and not federal law. The right of publicity prevents the unauthorized commercial use of an individual's name, likeness, or other recognizable aspects of one's persona. It gives an individual the exclusive right to license the use of their identity for commercial promotion.

If an AI-generated work were to be found to infringe on someone’s right to publicity, it is unclear whether the responsibility would fall to the AI system developer or the user of the AI system.

Current court cases that may shape the landscape

There are several current lawsuits that might shape the legal landscape for AI generated images:

- Getty Images is suing the creators of AI art tool Stable Diffusion for collecting its images into training data.[32] Getty Images (US), Inc. v. Stability AI, Inc., Feb 3, 2023, at 1 (1:23-cv-00135) (D. Del. 2023). Note that the complaint only specifically addresses 7,216 images and associated tags and descriptions. Id. at 8. Getty’s complaint alleges copyright infringement, violations of the DMCA in relation to copyright management information, trademark infringement,

[32] [Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content - The Verge]

unfair competition, trademark dilution, and deceptive trade practices in violation of Delaware law. Id.

- Three artists are suing the makers of AI art tools Midjourney and Stable Diffusion.[33] The lawsuit claims that by pre-training on the images of millions of artists that the AI system developers are violating the copyright of those artists. One of the plaintiffs in the case, Sarah Andersen, has discussed in interviews that she was disturbed by AI's ability to generate images in her characteristic style. Since style is not copyrightable this may not be a valid complaint under copyright law, but it is a valid concern that might warrant legislation akin to right of publicity where likeness plays a role. This lawsuit may also fail because of technical inaccuracies in its claims of how the AI systems work, comparing them to collage tools that create images by reconstructing parts of stored copyrighted images.[34]
- The same law firm suing Midjourney and Stable Diffusion on behalf of the artists is also suing Microsoft, GitHub and OpenAI for violating copyright by incorporating software source code in their training data for AI systems that helps coders write new software.[35],[36] Both lawsuits rely on the same proposition: their lawyers contend that the use of copyrighted works to pre-train an AI system is *not* fair use.

Considerations for AI Related Legislation

I have several suggestions for what lawmakers should consider when drafting AI-related legislation.

Establish Fair Use for Training AI Systems

As I have discussed, AI systems require huge amounts of data during their pre-training phase. In order to effectively learn how to use language, facts about the world, visual representations of objects, and many other general ideas, current systems need huge amounts of data. Current large language models are trained on roughly a trillion words, and current image generation systems are trained on hundreds of millions of images and their captions. Many or most of the items in the training data are copyrighted.

---

[33] [AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit - The Verge](https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart) https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

[34] https://www.technollama.co.uk/artists-file-class-action-lawsuit-against-stability-ai-deviantart-and-midjourney

[35] https://githubcopilotlitigation.com/

[36] https://www.theverge.com/2023/1/28/23575919/microsoft-openai-github-dismiss-copilot-ai-copyright-lawsuit

The community of researchers and companies who are developing AI systems contend that this is fair use. However, legal precedents have not yet been established. If it were to be ruled that training AI systems on copyrighted works were not fair use, and that every work in the training data set needed an explicit license from the copyright holder, then progress on developing capable AI systems would be jeopardized. A possible outcome could be that a small number of large corporations who already have licensed lots of copyright data could continue to innovate in the field of AI, but startups would be unlikely to be able to do so.

I propose that any future legislation on AI and copyright should make explicit that training on copyrighted works is fair use. Legislation should also provide a mechanism for creators to opt out of having their work included in training.

Task the Copyright Office with Establishing Registries

Congress may want to consider tasking the Copyright Office with creating registries of works where the creators can opt out of having their work included in training. Similarly, congress may want to task the Copyright Office with establishing a registry of copyrighted characters. This would provide a resource that AI companies could use to block the generation of those characters, or to advise users on fair use / non-commercial purposes of the generated works.

Consider Whether AI Generated Works Should Be Copyrightable

Works that are generated wholly by machines are not copyrightable, since only human beings may be considered authors for the purposes of copyright. I believe that this is a reasonable position to hold, and that it is in line with the constitution's principle to establish copyright to induce creative works. Current guidance from the Copyright Office is somewhat unclear about works that are created in collaboration with an AI (which currently applies to most AI-generated work). These co-created works should likely be copyrightable by the human author. A related issue is that the output of AI systems are currently covered by the Terms of Service of the AI developers, which may act as a legal surrogate for copyright and fair use.

Consider Legislation Regarding AI and Right-of-Publicity

Copyright law does not cover artistic styles, nor right-of-publicity. Many objections that artists have to AI systems stem from the AI's ability to generate good facsimiles of artists' styles or appearance or voice. I agree with the artists' objections. This is an evolving ethical area that touches on a wide range of potential misuses of AI including deep fakes. Congress should consider whether it is possible to write legislation that would protect people's appearance from being exploited for profit and protect people from being impersonated, while still preserving the right of parody and of non-commercial fan fiction and other similar activities.

Mitigate Negative Impacts on Employment

A much broader issue that Congress needs to contend with is the possibility of increasingly sophisticated AI systems resulting in mass unemployment, or the devaluation of certain kinds of work. Technological advances and change have led to unemployment before. Careers like lamplighters no longer exist because electricity replaced gas street lights. Factory work in the USA has been displaced by automation and lower cost wages elsewhere. Newsrooms were decimated in part because of the shift to the Internet. Artists whose industry has been slowly devalued over time are contending with replacement by AI. The Writers Guild of America are currently striking and part of their disagreement revolves around whether using AI systems such as ChatGPT to generate story ideas or scripts for films and shows should be disallowed. There is a real possibility that generative AI may be able to replace a large number of white collar jobs. Will paralegals go the way of lamplighters? If so, what should the government do to ensure that they are able to continue to make a living wage? Your role could be as simple as providing job retraining for displaced professions, or as complex as creating a new WPA.