A BRIEF OVERVIEW OF FAIR LEARNING

ON THE TRAINING STAGES OF GPT AI SYSTEMS

Berk BORAZAN*

August 2023

*berk@legalbydefault.com

TABLE OF CONTENTS

1. INTRODUCTION

In this brief,[1] we have tried to explain the current discussions on the "Fair Learning" principle, with a focus on the Generative Pre-trained Transformer (GPT) Artificial Intelligence (“AI”) systems. Many AI companies (including Midjourney and Stable Diffusion) have been sued for developing GPT models through copyrighted photos without the prior permission of the copyright holders. In this brief, we will discuss whether copyright holders have a legitimate claim against the AI companies that train GPT models with their copyrighted material. Furthermore, we will elaborate on suggestions given by the researchers on this topic.

2. DEFINITION OF GPT MODELS

In the first stage of the GPT models, vast amounts of data are collected and later on, in the training stage, the model is trained. After that model is deployed and new outputs are generated. In this stage, without other safeguarding mechanisms, there is no guarantee of the factual correctness of the output produced. The AI field calls this factually incorrect, made-up information as "hallucinations".

Pre-training in GPT models takes place with carefully selected samples according to the intended use of the GPT model. Depending on the GPT model's purpose, this pre-training phase could have the GPT model predict the next sentence in the sample document, or it could ask the GPT model to fill in the blank. Nonetheless, the original sample document will provide the "ground truth" to the GPT model[2] and the GPT model will learn from that truth. It is important to understand that AI learns the patterns and connections between data points during training[3].

During the training stage, it is possible for GPT models to memorize certain data points and give outputs exactly as memorized data points. This is a well-known problem concerning intellectual property rights and there are approaches to overcome this problem both in the training and deployment stages of the AI.

[1] I would like to thank Abdulmecit İçelli and Hande Çağla Yılmaz for their help in writing this brief and their neverending comradery.

[2] Callison-Burch, Christopher. Understanding Generative Artificial Intelligence and Its Relationship to Copyright, 2023. https://www.cis.upenn.edu/~ccb/publications/understanding-generative-AI-and-its-relationship-tocopyright.pdf.

[3] Callison-Burch, *Understanding Generative Artificial Intelligence*,

Machines are very arduous scholars. The amount of raw data needed for training is beyond our capacity to imagine. For instance, it is estimated that OpenAI's GPT-4 was probably trained with over than 1 trillion tokens[4]. This will be used as an argument in the case of fair learning.

There is more than one approach for gathering the data used for training the AI systems and it is possible to implement many approaches together, such as "Data Filtering" and "Common Crawling". Common Crawl is one of the organizations which considers robots.txt while crawling the web[5]. However, the most common practice is to gather as much raw data as possible without any regard to copyright holders. This is primarily done via "Web Crawling".

3. WEB CRAWLING

Web crawling is one the oldest methods to gather data that is presented on the web. Google and other search engines have dedicated "crawlers" which are actually bots that search and index websites all over the internet. There is a method that is accepted by the community (and now by courts) to exclude certain websites (subdirectories to be exact) from the crawlers' indexing capability. It is called "robots.txt". Website developers, create a subdirectory within their website domains. This subdirectory is called "robots.txt" and it spells out the subdirectories from which crawlers are excluded from the index. Hence, crawlers decide not to search that subdirectory.

This is how robots.txt looks like for the mock-up web site "legalbydefault.com".

" URL: legalbydefault.com/robots.txt
User-Agent: *
Disallow: /natas.html "

This code snippet would result to crawlers disregarding the page "/natas.html".

However, this method is non-binding for the parties and more act like a courtesy rule. Thus, it is possible to disregard this method in its entirety. Nevertheless, it is not totally arbitrary

[4] Callison-Burch, *Understanding Generative Artificial Intelligence*,

[5] Callison-Burch, *Understanding Generative Artificial Intelligence*,

either. In the case of "Field v. Google Inc." consideration for robots.txt was mentioned. Court stated that Field could have used the robots.txt page to disallow Google from archiving the website which Field failed to do so. As a result, Court attributed legal consequences regarding the implementation of robots.txt. However, it is unclear what the future will bring considering the hasty development of the AI systems[6].

4. ONLY GATHERING NON-COPYRIGHTED MATERIAL

Data sets can have licenses allowing the permitted use of the data set. However, data sets contain a lot of data points and some of them could also have other licenses which would not permit the use of that data point[7].

Furthermore, copyright owners can change their licenses freely. So, there are no guarantees that once allowed management of a data point will always stay allowed[8]. Since almost everything on the internet is copyrighted, only using non-copyrighted material could also lead to unwanted behaviors. If that were to be the case, biases in the training data sets would likely occur since the overwhelming majority of the materials will be from very old times (such as materials in the public domain which are from the 1920s).

5. FAIR USE

Assessment of the fair use relies on four standards created by Courts. “(1) the purpose and character of the use; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used; (4) the effect of the use upon the potential market for or value of the copyrighted work.”

In the general use case for GPT models, GPT is trying to learn the patterns and connections between the data points[8]. When the training data is fed into the AI model, it is not copied verbatim by the model but learned through neural networks. For example, if an GPT is trained for recognizing certain materials (let's say apples), it actually disregards the copyrightable artistic choices such as lighting but rather focuses on the factual information about aimed material (apples); this should be considered fair use. However, if an AI system is

[6] Mark A. Lemley and Bryan Casey. 2021. "Fair Learning | Texas Law Review". *Texas Law Review*. https://texaslawreview.org/fair-learning/.

[7] Henderson, Peter, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. “Foundation Models and Fair Use.” arXiv.org, March 28, 2023. https://arxiv.org/abs/2303.15715.

[8] Lemley and Casey, *Fair Learning*,

[8] Lemley and Casey, *Fair Learning*,

trained to mimic certain artistic choices; the fair use claim is much weaker[9]. AI systems generally change the purpose of the data.[10] This can considered as a transformative use in the scope of copyright law.

6. NON-EXPRESSIVE USE

Copying the data and storing it without the purpose of using the original expression of the work[11], is considered non-expressive use. In GPT systems, the intention for gathering and consuming the data is quite distinct from the intended use of the data which is for human interaction[12].

A. Sega v. Accolade Case

One of the first cases discussing non-expressive use was Sega v. Accolade[13]. In this case, the video game-producing firm Accolade, reverse engineered Sega's gaming console Genesis and used code snippets to make their game compatible with Genesis. The Court ruled that since the code required for games to be Genesis compatible is used without the purpose of copying expression but studying the idea, non-expressive use of the Genesis' code snippet is fair use[14].

B. Kelly v. Arriba Case

Arriba Soft. Corp. which runs a search engine, used Kelly's copyrighted works for thumbnails in their search engine[15]. When clicked on the thumbnail, the user would be redirected to the original website containing Kelly's copyrighted photographs. Court ruled that smaller-sized thumbnails would not substitute for Kelly's photographs and Arriba's gathering of the photographs and making them smaller thumbnails that redirect to the original website would be considered a tool and protected under non-expressive use.

[9] Lemley and Casey, *Fair Learning,*

[10] Lemley and Casey, *Fair Learning,*

[11] Ginsburg, *Fair Use in the United States,*

[12] Comment of OpenAI. "Before the United States Patent and Trademark Office Department Of ..." Before the United States Patent and Trademark Office Department of Commerce Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation. Accessed August 1, 2023. https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf.

[13] REINHARDT, Circuit Judge: "Sega Enterprises Ltd. v. Accolade, Inc.." Legal research tools from Casetext, October 20, 1992. https://casetext.com/case/sega-enterprises-ltd-v-accolade-inc-2.

[14] Sobel, Benjamin. "Artificial Intelligence's Fair Use Crisis - Ben Sobel." www.bensobel.org/, 2017. https://www.bensobel.org/files/articles/41.1_Sobel-FINAL.pdf.

[15] T.G. NELSON, Circuit Judge. "Kelly v. Arriba Soft Corp.." Legal research tools from Casetext, February 6, 2002. https://casetext.com/case/kelly-v-arriba-soft-corp.

C. Authors Guild v. Google Inc. (Google Books) Case

Google created a platform where Google made digital copies of the books[16]. This digitalization process was made without the affirmative consent of the right holders. The platform made it possible for people to search certain keywords and find books containing them. In this case, Judge decided Google's copying of the books was non-expressive. Furthermore, Google Books did not substitute for the books it contained since Google Books only gave snippets of the books.

7. ARGUMENTS AGAINST NON-EXPRESSIVE USE IN AI SYSTEMS

It is possible to extract value from the expressive qualities of the work, rather than merely extracting information.[17] In the abovementioned precedents, it is stated in the Court's view; it is important to take into consideration the effects on the market of the tools used. Development of the AI could cheapen the labor that is produced by workers[18] and could overflow the market. AI systems could also affect the information shared for the public good. According to a study, "There is a 16% decrease in the weekly posts on Stack Overflow that could be attributed to Large Language Models."[19].

8. COPYRIGHT INFRINGEMENT VIA PROMPT ENGINEERING

It is possible through choosing correct prompts to reveal copyright-infringing training data sets in the deployment phase of the AI systems. In a study, by trying various prompts; researchers got the output containing the Dr. Seuss' story "Oh the Place You'll Go!" completely[20]. It is to be said when training data sets contain copyrighted material, sufficient filtering of the outputs is essential.

[16] Leval, Circuit Judge. "Authors Guild v. Google, Inc.." Legal research tools from Casetext, October 16, 2015. https://casetext.com/case/guild-v-google-inc-1.

[17] Sobel, *Artificial Intelligence's Fair Use Crisis*,

[18] Sobel, *Artificial Intelligence's Fair Use Crisis*,

[19] Rio-Chanona, Maria del, Nadzeya Laurentsyeva, and Johannes Wachs. "ArXiv:2307.07367v1 [Cs.SI] 14 Jul 2023." Arxiv.org, 2023. https://arxiv.org/pdf/2307.07367.

[20] Henderson et al., *Foundation models and fair use*,

9. SUGGESTIONS

It is possible to create a registry for copyrighted works on the web, where it would be possible to compare training data points with those in the registry.[21] It should be possible to create an opt-out mechanism for the copyrighted works which could act like robots.txt.[22] A lot of researchers are also using AI systems for academic purposes. If the training data sets were to be only limited to non-copyrighted material, this could be detrimental to society[23].

There needs to be more tools to assess if an output given by AI is infringing on copyrighted material. The currently used text overlap method is inadequate. However, it is acknowledged by scholars that fair use can not be easily assessed by only using these tools[24].

In our opinion, we are only at the beginning stages of the AI development journey. Current methods of processing vast amounts of raw data will surely change. In the future, it should be possible for AI systems to learn from much smaller data sets. This would help at least the smaller stakeholders whose copyrighted materials are being processed without their consent.

Furthermore, AI models will be trained on copyrighted material even if we restrict it. So, instead of restricting it we should incentivize socially beneficial models and use cases. However, in doing so we still need to restrict certain type of use cases which hinders the growth of society and innovation. It is of utmost importance to make authors feel safe with their expression. If people believe their livelihood will be taken by an AI model, and even worse that AI model is trained with their own works, they will not contribute to society. Copying styles of authors is not something that should be protected when it is done by an AI model. This would result in authors not sharing their work publicly which is catastrophic for humanity.

Choosing fair use doctrine as the basis for these models is a deliberate decision. Fair use was applied for promoting innovation where the legislations were inadequate for solving the dispute. Fair use fits well on the dynamic environment of AI development.

[21] Callison-Burch, *Understanding Generative Artificial Intelligence,*
[22] Henderson et al., *Foundation models and fair use,*
[23] Callison-Burch, *Understanding Generative Artificial Intelligence,*
[24] Henderson et al., *Foundation models and fair use,*

BIBLIOGRAPHY

CALLISON-BURCH, CHRISTOPHER, Understanding Generative Artificial Intelligence and Its Relationship to Copyright (2023), https://www.cis.upenn.edu/~ccb/publications/understanding-generative-AI-and-itsrelationship-to-copyright.pdf, Last Access Date: 03.08.2023

COMMENT OF OPENAI, Before the United States Patent and Trademark Office Department of Commerce: Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation (2023), https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf, Last Access Date: 03.08.2023

CONGRESSIONAL RESEARCH SERVICE, Generative Artificial Intelligence and Copyright Law (2023), https://crsreports.congress.gov/product/pdf/LSB/LSB10922, Last Access Date: 03.08.2023

CREWS, KENNETH, Fair Use, https://copyright.columbia.edu/basics/fairuse.html#:~:text=Fair%20use%20is%20more%20likely,or%20criticism%20of%20the%20ori ginal, Last Access Date: 03.08.2023

GINSBURG, JANE C. “Fair Use in the United States: Transformed, Deformed, Reformed?” Columbia Law School Scholarship Archive, 2020. https://scholarship.law.columbia.edu/cgi/viewcontent.cgi?article=3680&context=faculty_scholarship.

LEVAL, Authors Guild v. Google, Inc. (2015), https://casetext.com/case/guild-v-google-inc1, Last Access Date: 03.08.2023

HENDERSON, PETER, XUECHEN, JURAFSKY, HASHIMOTO, MARK A. LEMLEY, AND PERCY LIANG. “Foundation Models and Fair Use.” arXiv.org, March 28, 2023. https://arxiv.org/abs/2303.15715.

MARK A. LEMLEY AND BRYAN CASEY. 2021. "Fair Learning | Texas Law Review". *Texas Law Review*. https://texaslawreview.org/fair-learning/.

REINHARDT, Circuit Judge: “Sega Enterprises Ltd. v. Accolade, Inc..” Legal research tools from Casetext, October 20, 1992. https://casetext.com/case/sega-enterprises-ltd-v-accoladeinc-2.

RIO-CHANONA, MARIA DEL, NADZEYA LAURENTSYEVA, AND JOHANNES WACHS. “ArXiv:2307.07367v1 [Cs.SI] 14 Jul 2023.” Arxiv.org, 2023. https://arxiv.org/pdf/2307.07367.

SOBEL, BENJAMIN. “Artificial Intelligence’s Fair Use Crisis - Ben Sobel.” www.bensobel.org/, 2017. https://www.bensobel.org/files/articles/41.1_Sobel-FINAL.pdf.

T.G. NELSON, Circuit Judge. “Kelly v. Arriba Soft Corp..” Legal research tools from Casetext, February 6, 2002. https://casetext.com/case/kelly-v-arriba-soft-corp.

U.S. COPYRIGHT OFFICE, LIBRARY OF CONGRESS. “Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence.” LIBRARY OF CONGRESS Copyright Office, 2023. https://public-inspection.federalregister.gov/2023-05321.pdf.

VYAS, NIKHIL, SHAM KAKADE, AND BOAZ BARAK. “On Provable Copyright Protection for Generative Models.” arXiv.org, July 21, 2023. https://arxiv.org/abs/2302.10870.