
COPYRIGHT CONCERNS OF OPEN AI'S PRODUCTS

T.E.Krishna, Vellore Institute of Technology, Chennai

ABSTRACT

Artificial intelligence is one of the millennium inventions which gained a large quantum of users, admirers and haters at the same time. The Open AI Company, which is widely known for its chat bot known as “ChatGPT” has been the talk of the world ever since its incorporation. The whole intention of AI is to imitate human cognitive functions, among which the ChatGPT is a Generative Pre Trained Transformer AI which gives out novel as well as existing outputs for the command given by the users. To produce such results, the AI needs training through the existing databases. So technically, there are inputs which train the AI and outputs which it provides. Both inputs and outputs of this AI are prone to Copyright infringement concerns of the original artistic and literary works. When AI tries to imitate humans and produce human like responses, it's vulnerable to IP infringement. This article primarily aims to analyse the concerns of authors with respect to the Copyright infringement of Open AI and to also analyse Open AI defences against the claims.

INTRODUCTION

In the last few years, the pressure of tasks, assignments, projects of every section of the society has been relieved significantly by a tool named Artificial intelligence. It is known among the public generally as the “doer of anything” right from doing projects for us, come up with a creative write up for our necessities, generating novel images as required by us to presenting a broader view on bail jurisprudence in the High court¹. Such amount of reliance is what AI has on its shoulders at present. Its output, even though said to be an AI generated output, is not autonomously generated, it requires training of the AI with large amount of databases for it to understand the statistical combinations of command input to come up with the required outputs of the users. For example, for it to answer your question of “Who killed Dumbledore in the books”; it requires input of the text of the book “Harry Potter and the Half-blood prince”. Similarly, for it to come up with an image of “J.K.Rowling having kid Harry in her hands” it needs information about Harry’s appearance from the primary sources which are books and the appearance details of J.K.Rowling. So technically, it comes up with self-generated outputs which are derivatives of the input it is trained from. Eventually, when software and other technological domains use creative Copyrighted works such as the ones mentioned above, the claims for Copyright arises. Such claims have arose in the USA with respect to the Open AI company’s products from the authors of USA for using their Copyrighted works as inputs and generation of derivatives of their Copyrighted works for commercial purposes. Such claims haven’t come majorly into the picture in India yet, but the days aren’t far away. So this article tries to understand how these AI tools work and their Copyright infringement vulnerability and how transformed will it be when it comes into the discussions under the premise of Indian Copyright laws as against the US laws.

OPEN AI AND CHAT GPT

Open AI was founded on 2015 as an Non-Profit Organisation with a goal to research and create AI products which would benefit humanity as a whole without any condition or need to yield financial return. It was seen to state that this invention and development aimed to focus on a

¹ Ani, *In a first, Punjab and Haryana high court uses Chat GPT to decide bail plea*, Times of India (Mar. 28, 2023), <https://timesofindia.indiatimes.com/india/in-a-first-punjab-and-haryana-high-court-uses-chat-gpt-for-deciding-upon-bail-plea/articleshow/99070238.cms?from=mdr> (last visited on 16/10/2023)

positive human impact free from financial obligations². In 2020, the company Open AI transformed its structure from a NPO to a capped profit organisation which invited several comments from the public as well as from one of its notable member till 2018, Elon Musk, who called it a maximum profit company effectively controlled by Microsoft. When its CEO, Sam Altman, was questioned about the transformation, he answered in manner which meant that there is no shift in goals from what was stated in its previous form but the company required more capital than what was estimated then and the transformation is also to convene their investors as well as employers with a fixed return³. As a result of its transformation, leading technology companies like Microsoft, started to pour in subsequent investments in Open AI of \$1 billion, \$2 billion and \$10 billion in 2019, 2021 and 2022 respectively. The last time its valuation was estimated, Open AI was valued at \$29 billion⁴.

ChatGPT, the infamous product of Open AI on the other hand, ever since the transformation of Open AI has seen improvements in its functioning through upgrades in version. It stood as GPT - 3 in 2020, which was later refined and launched as GPT – 3.5 in 2022 which was the version which shoot up the graph of ChatGPT as one of the largest used product as on January 2023. It later got upgraded as GPT – 4, which was the first time since its inception, ChatGPT became available only on \$20 annual subscriptions and not for free. It is important to note here that both Open AI and Chat GPT have transformed itself into a “for - profit” organisation and “not for free” product respectively. The parent company and the product have started to make immense money from these products and it was stated by Open AI that 80% of employees of the Fortune 500 companies are using ChatGPT.

THE WORKING OF CHATGPT

As already stated above, ChatGPT is a Generative AI which is a pre-trained transformer. Its working is based on an algorithm known as large language models (LLM), it processes the natural

² *Introducing OpenAI*, <https://openai.com/blog/introducing-openai>. (last visited on 16/10/2023)

³ Aman Gupta, *Why did OpenAI go from non-profit to a capped-profit company? Sam Altman explains*, Mint (Mar. 27, 2023), <https://www.livemint.com/news/india/why-did-openai-go-from-non-profit-to-a-capped-profit-company-sam-altman-explains-11679895623795.html>. ((last visited on 16/10/2023)

⁴ Sissi Cao, *OpenAI, the Company Behind ChatGPT, Is Valued at \$29 Billion*, Observer (Jan. 6, 2023), <https://observer.com/2023/01/chatgpt-openai-valued-29-billion/> (last visited on 16/10/2023)

language input and predicts the next word on the basis of what it has already seen⁵. It is largely based on statistics i.e. the AI uses statistical patterns of in text datasets which they are trained with already. It responds to command inputs by correlating the relationship between words in it and similar combination of words in the data it's trained with. To make it precise, "training" under this context refers to the ingestion of data into the AI, mostly primary sources of data to produce quality results. Primary sources includes Copyrighted works such as books, articles etc. which would naturally draw Copyright claims if used without licensing and that is what has come up in the USA.

THE SOURCE OF INPUTS

The training of bots like the products of Open AI require so much data, Open AI generally considers the source of these data as proprietary but it has come out to state as much as that the corpora Open AI used to train its ChatGPT3 with, are excessively publicly available data which include Copyrighted works⁶. With regard to the specific disclosure of the sources of datasets, Open AI lastly, had slightly spilled out the sources through which it trained ChatGPT 3 which included filtered versions of Common Crawl, the famously known data set with no clean chit. Open AI stated that it has used around 410 billion tokens from filtered Common Crawl data which constitutes around 60% of the training⁷. To train ChatGPT which has to function in a multilingual basis, mC4 should serve as the dataset source i.e. mC4 contains the dataset in common crawl in a multilingual basis and it is of no surprise to see that it forms the corpora of ChatGPT 3⁸. The Dutch sources of the corpora that form mC4 are the hubs of Copyright violations, private data and fake news⁹. One such example of such source is the Docplayer which is one of the most important sites for pirated data. Conclusively, pirated data being a part of the corpus of the training input to train

⁵ Lucas Mearian, *What are LLMs, and how are they used in generative AI?*, Computerworld <https://www.computerworld.com/article/3697649/what-are-large-language-models-and-how-are-they-used-in-generative-ai.html> (last visited on 16/10/2023)

⁶ (Mar. 11, 2020), https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf (last visited on 16/10/2023)

⁷ <https://arxiv.org/pdf/2005.14165.pdf> (last visited on 16/10/2023)

⁸ Laura Herijgers, *'ChatGPT based on illegal sites, private data and piracy'*, Techzine Europe (June 8, 2023), <https://www.techzine.eu/blogs/privacy-compliance/107181/chatgpt-based-on-illegal-sites-private-data-and-piracy/> (last visited on 16/10/2023)

⁹ *'Dat zijn toch gewoon ál onze artikelen?'*, De Groene Amsterdammer (June 7, 2023), <https://www.groene.nl/artikel/dat-zijn-toch-gewoon-al-onze-artikelen> (translated by Google translate) (last visited on 16/10/2023)

ChatGPT is established by the admission of Open AI themselves and through other sources. Open AI has eventually used pirated freely available data to commercialise from its products ever since it turned into a “for profit” organisation. The use of Copyrighted works in its previous versions puts ChatGPT4 in a suspicious position. It is a fact that the output generated by ChatGPT would be of good quality only if it is trained with primary sources of the information such as Copyrighted books, research works etc.

OUTPUT OF CHATGPT

The output of ChatGPT, as everybody who has used it knows, is unrestricted as to the extent of content it spills out to answer to the commands put up. It never restricts from generating answers from any part of the inputs it is trained with. Until recently, it had the capacity to reproduce works in the garb of quotations but now it is cautious with such commands and responds that it cannot produce excerpts from Copyrighted works. However, it still continues to summarize Copyrighted works with so much accuracy and the company behind it, as noted earlier, directly jumped to commercialisation of this product by charging \$20 dollars for its latest version of GPT 4. There cannot be a dispute over the fact that the outputs ChatGPT produces regarding the Copyrighted works fall under the ambit of “derivatives of the original work and such production of derivatives falls under the exclusive right of the Copyright owner. These contentions have already been piled up in the US cases while these haven’t popped up in India. It is important to analyse Open AI’s defence to understand the issue further.

ANALYSIS

Open AI, ever since these contentions and allegations popped up established a position publicly that it believes that the use of Copyrighted information amounts to fair use under the US legislation. It is necessary to analyse the concept of fair use as against the functioning of ChatGPT

Section 107 of the US Copyright Act states the criteria under which fair use has to be judged

Notwithstanding the provisions of sections 106 and 106A, the fair use of a Copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting,

teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of Copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include—

- (1) the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes;
- (2) the nature of the Copyrighted works;
- (3) the amount and substantiality of the portion used in relation to the Copyrighted work as a whole; and
- (4) the effect of the use upon the potential market for or value of the Copyrighted work.

The application of this provision differs from case to case contingent to the facts and circumstances of this case and while judging a case with respect to these factors, one factor shouldn't be considered autonomously from the other but all of these factors have to be considered together¹⁰. To the question that which of the above criteria has more weightage over the others, it is to be understood that the effect of the usage upon the potential market or the value of the Copyright is the most important element out of the four¹¹.

To conveniently understand the Copyright infringement position of ChatGPT, a comparative analysis of Google Books position in the case of Authors Guild v. Google Inc. will help better. Google books, a product of Google, initiated a Google Library Project through which it will enter into agreements with worlds major research library to produce a copy of the books available with them to ingest the data into the Google Library web form which it would in turn display snippet views of the books in case a search appears with keywords related to the book. This mega project of Google books was contested on the grounds of Copyright violation for ingesting Copyrighted books in the Google library data without authorization and displaying snippets to the public unauthorized.

¹⁰ *Id.* at 577, 114 S.Ct. 1164

¹¹ *Harper & Row Publishers, Inc. v. Nation*

The production of snippets from Google works in way that it only gives out limited view of the snippets from the books. Pursuant to a keyword entered, Google will give the user 3 snippets which in size will be one eighth of a page and regardless of the number of servers or number of times or number of different keywords the user tries, the standard 3 snippets of a book will come in the results, any other snippet other than the standard snippet wouldn't come in the search results. Hence, Google library produces a restricted view of the books to enable views to decide whether the particular book is necessary for them or not.

Transformation

Open AI has time and again insisted on the defence that the outputs ChatGPT produce transformed form of the ingested information like Google Library. Transformation in this context refers to the change in form of the input to produce and output. Transformation in certain contexts is a defence of fair use under certain circumstances provided that such transformation is justified with a purpose. Google transformed books into snippets to enable public view the limited snippets to make choices i.e. they had a justification that such transformation will help their uses make choices. It is to be noted here that Google books snippets never reveal substantial part of the books neither summarizes it which enables it to claim a defence that it has reproduced very limited portion of the Copyrighted work under 107 (3) of the US Copyright Act. Whereas in the case of ChatGPT, it is different, Transformation which falls under the ambit of fair use has to be understood as distinct from "derivatives" which fall under the exclusive right of the Copyright owners. Acts like summarizing of texts to reveal its substantial portions without restrictions which ChatGPT does and artistic reproduction of the work like pictures which another Dall- E, another tool of Open AI, does cannot be classified into mere transformations but derivatives. The tools of Open AI pose no restrictions in quantity performing these functions unlike the snippets of Google which reveals only standard parts and not substantial parts. For instance, ChatGPT is going to reveal the whole story of Harry potter being sorted into Gryffindor if you command it to do so which is a substantial part of the Copyrighted work of J.K.Rowling unlike Google Snippets. Hence, the outputs of AI tools of Open AI are derivatives and are not transformative as claimed by Open AI.

To better understand what ChatGPT does, It is important to produce here the output it gave out to

my command “tell me the story of harry being sorted into Gryffindor by hat”

“Certainly! Here is the story of Harry Potter being sorted into Gryffindor by the Sorting Hat:

Harry Potter's journey to Hogwarts School of Witchcraft and Wizardry began when he received his acceptance letter to the school. On his first day at Hogwarts, he, along with the other new students, was led into the Great Hall for the Sorting Ceremony. The Great Hall was a vast and magnificent room with four long tables, one for each of the school's houses: Gryffindor, Hufflepuff, Ravenclaw, and Slytherin.

As the first years gathered in anticipation, Professor McGonagall, the Deputy Headmistress, placed a ragged and ancient-looking hat on a stool at the front of the room. This hat was the Sorting Hat, a magical hat with the ability to determine which of the four houses each student belonged to. The hat had a tradition of singing a song before it began sorting the students.

The Sorting Hat sang a song that outlined the qualities of each house, and then it was time for the Sorting Hat to make its decisions. One by one, the students were called up to the stool, and the hat was placed on their heads. It delved into the minds of the students, considering their strengths, weaknesses, and deepest desires.

When it was finally Harry's turn, he felt nervous and worried about where he would end up. The Sorting Hat had a brief internal debate before ultimately deciding that Harry belonged in Gryffindor. The hat shouted "Gryffindor!" to the applause of the students and the cheering of the Gryffindor table.

Harry was relieved and excited to be part of Gryffindor House, which was known for its bravery, courage, and a strong sense of justice. From that moment on, Harry's Hogwarts journey began as a Gryffindor, and he made lifelong friends in Ron Weasley and Hermione Granger, who were also sorted into the same house.

This marked the start of Harry's adventures and the beginning of his path to becoming a

legendary wizard in the wizarding world.”

As we can observe that ChatGPT summarized a substantial part of the book which is a derivative of the Copyrighted work.

Hence, there is no hesitation in concluding that Google’s Snippets fall under the ambit of Transformation¹² and with respect to the transformation defence of Open AI, there is no fear in concluding that ChatGPT infringes the exclusive right of the Copyright owner by producing derivatives and by the term derivatives, it is meant that it reveals substantial part of the book in the garb of summarising.

It is to be noted that this test of revealing the substantial parts thereof is not autonomous and should be adjudged along with the other factors.

Effect on potential market

In the case of Google Library snippets, the substantial parts are not revealed so that the potential markets of the Copyrighted works aren’t diminished, ironically, it improves the market of the Copyrighted works because if the user finds the book to be useful, he would initiate to buy the same. Contrarily in the ChatGPT, without any restrictions, the user can access the substantial portions of the Copyrighted work which fulfils the purpose with which he sought for the material. As it was clear from the case of Harry potter command where it narrated about the sorting ceremony of the first book of the series of Harry Potter, it served the user with the purpose they seek without even having to buy the book. This in turn would diminish the market of the Copyrighted work unlike the Google Library snippet view. Finding a fair use is more likely when very small portions, or less important passages of the subject work are reproduced¹³ and this is not the case with ChatGPT where it produces the derivatives of works serving the purpose of Copyrighted work majorly.

Commercial Use

¹² Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) 202 (2015).

¹³ Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015) 202 (2015).

The first condition under section 107 is to test the nature of the infringement, that whether it is for a commercial purpose or not. This is not a conclusive criterion but has to be adjudged along with the other factors¹⁴. With the other factors established above and the transformation of Open AI from a NPO to a commercial organisation already discussed, there cannot be any dispute over the affirmation of this secondary factor to support the case of Copyright infringement.

With all the above information and contentions in hand, there is no hesitation in concluding that the functions of the AI tools of Open AI commit Copyright Infringement.

IMPLICATIONS TO INDIAN LAW AND SUGGESTIONS

The functions of ChatGPT would also constitute an infringement under section 51 of Indian Copyright Act which states;

(a) when any person, without a licence granted by the owner of the copyright or the Registrar of Copyrights under this Act or in contravention of the conditions of a licence so granted or of any condition imposed by a competent authority under this Act-

(i) does anything, the exclusive right to do which is by this Act conferred upon the owner of the copyright, or

¹[(ii) permits for profit any place to be used for the communication of the work to the public where such communication constitutes an infringement of the copyright in the work, unless he was not aware and had no reasonable ground for believing that such communication to the public would be an infringement of copyright; or]

(b) when any person--

(i) makes for sale or hire, or sells or lets for hire, or by way of trade displays or offers for sale or hire, or

(ii) distributes either for the purpose of trade or to such an extent as to affect prejudicially the

¹⁴ *Campbell*, 510 U.S. at 584, 114 S.Ct. 1164

owner of the copyright, or

(iii) by way of trade exhibits in public, or

(iv) imports ^{2***} into India, any infringing copies of the work:

³[Provided that nothing in sub-clause (iv) shall apply to the import of one copy of any work for the private and domestic use of the importer.]

While we are equipped enough to encounter the acts of infringement of AI with respect to Copyright infringement, the fixation of liability is unclear.

The fixation of liability in cases of AI infringing Copyright is more or less settled in the US, the liability is fixed in both the user and the AI owner but potentially the AI owner is likely to face the proceedings under vicarious infringement¹⁵. But under the Indian law, it is only clarified that AI cannot be recognized as an author to be granted copyright¹⁶ but there is no clarification about the fixation of liability for Copyright infringement under Copyright Act. It is suggested through this article that since AI cannot face legal proceedings, it would be suitable to follow the US doctrine to fix vicarious infringement liability on the AI owner. It is very likely that India is going to face such situations where AI's are sued for Copyright Infringement in the near future and it would be easy if India equips itself with necessary legislations to face the same.

CONCLUSION

Regardless of the assent of the Copyrights owners, AI is going to be the future of the world. Copyright owners and artists are already vulnerable about the growth of AI destroying their market and livelihood. It will not be apt to curtail technological development pertaining to the interests of artists and Copyright owner instead there is a need to equip the legislations with necessary remedies to counter the same. This article tried to bring out the effects and disadvantages of Copyright infringement which needs cognizance and recognition with the help of current

¹⁵ *Just a moment...*, <https://crsreports.congress.gov/product/pdf/LSB/LSB10922> (last visited on 16/10/2023)

¹⁶ Srishti Ojha, *Who owns AI-generated works? Here's what the laws say on copyright issue*, India Today (Sept. 22, 2023), <https://www.indiatoday.in/law/story/chatgpt-ai-generated-content-copyright-ownership-complexities-india-2439165-2023-09-22> . (last visited on 16/10/2023)

happenings around the world and the current products of AI in the market and it is believed that it served its purpose.