# THE AI-COPYRIGHT CHALLENGE: AUTHORIAL, INFRINGING AND CULPABLE ATTRIBUTES OF AI MODELS

Chinmay
*Research Scholar, PhD Programme, Department of Law, University of Delhi, Delhi.*
*Email: chinmay@law.du.ac.in, ORCID iD: 0009-0000-7580-3017*

**Abstract**

*This research article examines the convergence of Artificial Intelligence and copyright law, with particular emphasis on the ramifications thereof on generative AI and text-data mining (TDM). The fundamental objective is to deliver a thorough descriptive, critical, and normative analysis of the intersection between these two domains. The research provides a comprehensive evaluation of legal difficulties by identifying probable grey areas within the current legal framework. The Gen-AI model poses copyright challenges at input and output stages. At the input stage issues that crop up relate as to how the copyright law treats the situation wherein prompts contain reference to copyright protected material; another concern is when copyright protected data is used to train the AI Model; further at the stage of execution, post prompt commands the AI model scanning huge data including copyrighted material and making copies; at what stage infringement occurs and who can be held liable; at the output stage wherein the process culminates into an art, who can be regarded as the author. Should the act of training the model with copyrighted material be exempted? Can the Model be held liable for infringing the material when producing copies? Can it be regarded as an author or joint author for that matter? These are some of the complex and intriguing questions that need urgent attention, arising in the contemporary technological environment at the very centre of which is the fact of AI Model using copyright protected material as a part of training process, wherefrom all the aforementioned issues arise. This paper also examines the practical and theoretical ramifications of these concerns, intending to guide both contemporary and future governance in India. The target audience comprises regulators, policymakers and scholars involved in AI and copyright law, as well as offering significant insights for legal professionals and intellectual property right owners adapting to the changing legal environment. The study aims to propose interpretations that significantly enhance current deliberations and policy debates in this swiftly evolving legal domain.*

**Keywords:** Artificial Intelligence, Intellectual Property, Copyright, fair use, Liability, Copyright Infringement, GenerativeAI, Text and Data Mining.

*In a world where code gives birth to art,*
*Generative AI plays its part.*
*But who can claim the brush or pen,*
*When machines create again and again?*

*Copyright's laws, once clear and bright,*
*Now wrestle with this digital flight.*

*Is it the coder, or the AI's mind,*
*That holds the rights, or leaves them behind?*

*Authorship, once a human trace,*
*Now challenges in cyberspace.*
*Infringement lurks, and courts must try,*
*To answer who, and what, and why.*

*As circuits hum and pixels gleam,*
*Legal theories must chase the dream.*
*The age of AI, with creative might,*
*Demands new rules to set things right.*

*-ChatGPT*

## 1. Introduction

The author of this paper gave ChatGPT, the most sought after Large Language Model (LLM), a prompt to write a poem delineating copyright challenges posed by generative AI. In not more than ten seconds a string of rhyming sentences started to appear on the screen, to the astonishment of this author, who gave the same command three more times only to find a better and finer selection of words by the AI model working in secrecy to aid the author. In yet another instance this author gave prompts in another AI application called *DALL E* retrieving donald duck dancing in Michael Jackson style at illuminated Eiffel Tower (image 1). In yet another attempt in response to relevant prompts, a picture of mickey mouse eating pizza while climbing on the Eiffel Tower with skates on, was obtained (image 2). Does these images pose any copyright challenges?

Image 1

Image 2

The AI models as mentioned above have evolved to perform various tasks including text and image creations. This paper deals particularly with AI models that function as text-to-image creators. The AI Industry in the domain of text-to-image creation platforms is very active. Users have many options to choose from for creating images of their choice in seconds.

However no matter how easy, funny, intriguing and interesting it may seem, the challenges posed are very complex when norms of copyright regime are pitched against the working of these AI models. The challenges are mainly seen at two different stages in the process involving creation of images from texts; the input stage and the output stage (figure no. 1).

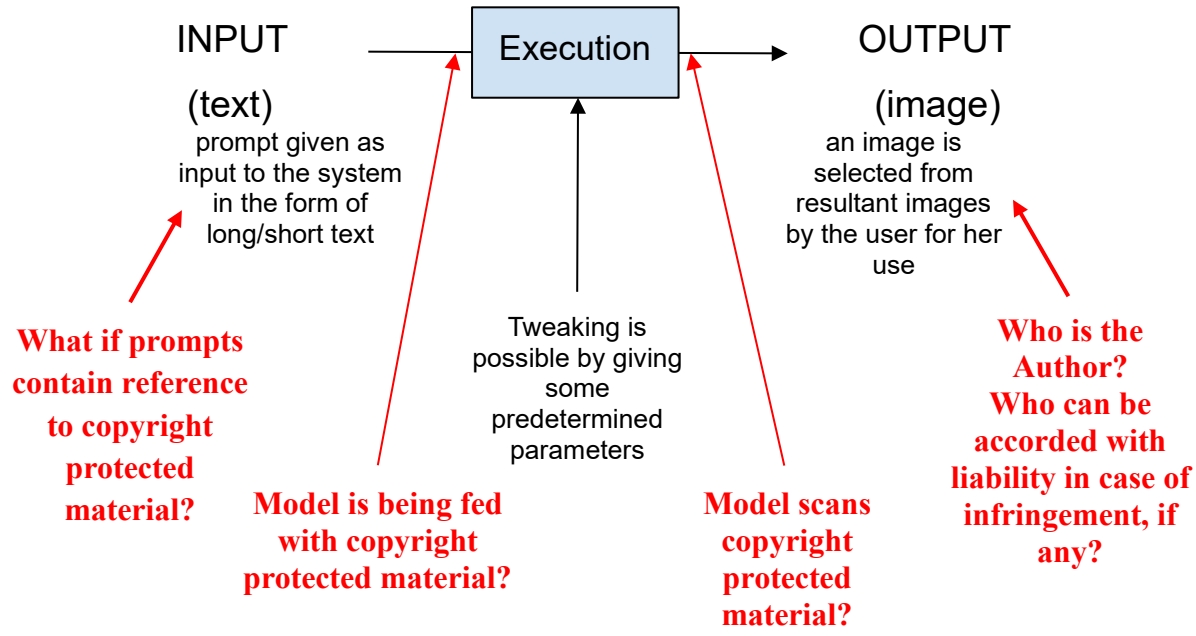


Figure No. 1. Stages in a text-to-image GenAI Model

Source: Author

At the input stage issues that crop up relate as to how the copyright law treats the situation wherein prompts contain reference to copyright protected material[1]; another concern is when copyright protected data is used to train the AI Model; further at the stage of execution, post prompt commands the AI model scanning huge data including copyrighted material and

[1] Andres Guadamuz, *A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs*, 73 GRUR Int. 111 (2024).

making copies; at what stage infringement occurs[2] and who can be held liable; at the output stage wherein the process culminates into an art, who can be regarded as the author[3]. Should the act of training the model with copyrighted material be exempted? Can the Model be held liable for infringing the material when producing copies? Can it be regarded as an author or joint author for that matter?

These are some of the complex and intriguing questions that need urgent attention, arising in the contemporary technological environment at the very centre of which is the fact of AI Model using copyright protected material as a part of training process, wherefrom all the aforementioned issues arise. It is crucial to note that these issues are pertinent not only for academic purposes but also for industry, policy makers, artists, musicians and other stakeholders[4]. The endeavour in this research article is to carefully analyse these issues from the lens of legal theory and also to critically analyse and evaluate the problems in light of the current copyright regime in India. The author also takes note of the developments occurring in the European Union (EU) relating to the aforementioned issues.

## 2. Structure of the Paper

The paper starts with introducing the concept of interface between copyright law and Generative AI technology. Thereafter the paper apprises the reader of the research objective, relevant research questions and the research methodology incorporated herein the paper. Thereafter the paper is divided into three sections. Section one provides a comprehensive understanding on the modus operandi of a Generative AI model. Section two discusses liability rules with respect to the Gen-AI Model. Section three provides a comprehensive analysis of copyright law vis-a-vis input questions that pose problems in the contemporary technological environment. Finally the paper concludes with some suggestions to comprehend the issues raised by Gen-Ai models in the copyright regime and also suggests areas of future research pertaining to the domain.

## 3. Research Objective and Methodology

The objective of the research is to critically analyse the input problems, as mentioned above, faced when emerging technology of generative AI is pitted against the current copyright regime. The research endeavors to suggest liability rules that may be taken into consideration at the time of

[2] *Id.*
[3] Mark A Lemley, *HOW GENERATIVE AI TURNS COPYRIGHT UPSIDE DOWN*, 25 Sci. Technol. Law Rev. 21 (2024).
[4] *Id.*

framing policies and regulatory frameworks. Generally the research will be conducted by incorporating a Qualitative Doctrinal approach. The background for the research is made by perusal of books, articles, research papers , news articles and official websites. This analysis seeks to determine the challenges and their possible solutions in the legal environment created at the intersection of copyright law and Generative-AI. The analysis tries to provide solutions conducive for implementation in our country and also ascertain steps that are needed in a direction towards bridging the vacuum created by absence of any regulation to govern Gen-AI issues.

## 4. Section I

### 4.1 The Gen-AI Model

Generally, a Gen-AI (Generative Artificial Intelligence) Model works on the principles of *prompt engineering* which thrives on prompt commands by the users. Other technical processes that the model undertakes include *execution* and *tweaking*[5]. The process of tweaking includes human efforts (figure no. 2) that give directions to and guide the process, though on the basis of predetermined standards. The following paragraphs explain the working of a Gen-AI model in detail.

Large Language Models (LLMs) have taken the world by storm. These LLMs are essentially a part of a bigger class of models called *foundational models*[6]. Initially libraries of AI models were fed specific data and the systems were so programmed so as to perform only specific tasks. Later it was predicted that all these task specific models can be combined resulting in a model that can perform various tasks including those specific tasks[7]. Such a model came to be known as a foundational model. Thus such a model with foundational capacity is a combination of specific models[8].

A foundational model is not task specific rather it can be *transferred* to perform any number of tasks. The question that arises is that how is such a model capable of executing multiple tasks and how can it be transferred to any number of tasks? The answer is that the model is being fed and trained on a humongous amount of data. The manner in which the data is fed is largely unsupervised and random, the data itself also being unstructured.

[5] Stefan Feuerriegel et al., *Generative AI*, 66 Bus. Inf. Syst. Eng. 111 (2024).
[6] A team from the University of Stanford coined this term on witnessing confluence of artificial intelligence into a new paradigm. Task specific models were being replaced by new models that could perform various tasks.
[7] Feuerriegel et al., *supra* note 5.
[8] Peter Cohan, *What Is Generative AI?*, 1 *in* Brain Rush: How to Invest and Compete in the Real World of Generative AI 9 (1 ed. 2024), https://link.springer.com/10.1007/979-8-8688-0318-5_2 (last visited Feb 11, 2025).

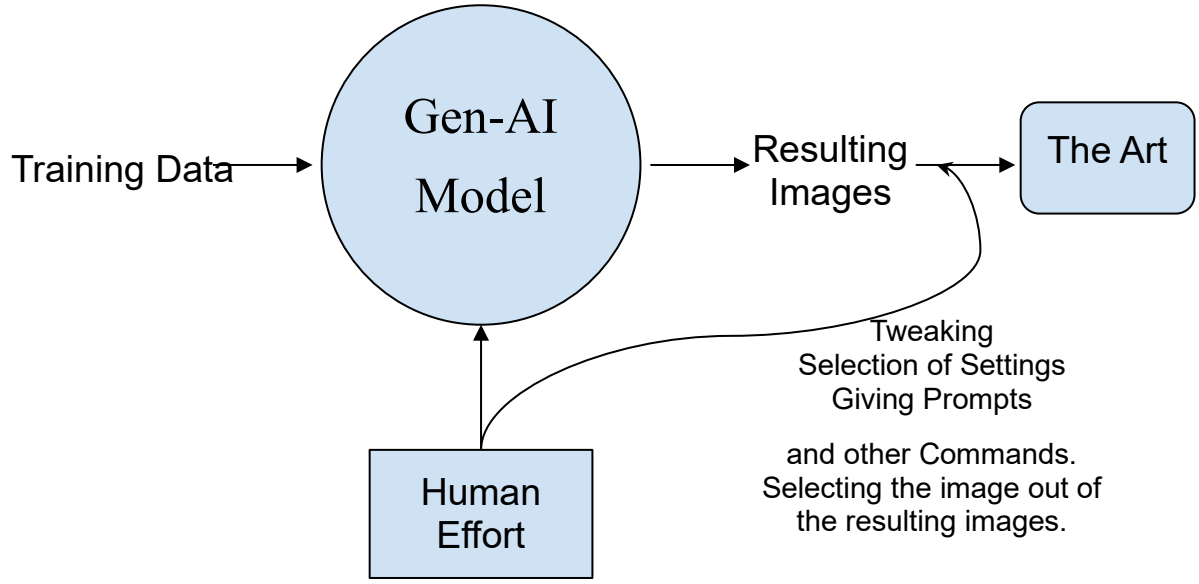


Figure No. 2 The Working of a Generative AI Model

Source: Author

Immense data containing sentences and images running into several terabytes is used to train the model. The trainer who feeds the data to train the system, programs the system and endeavours that, on his input, the model predicts the desired outcome based on all the data the system has seen before. For eg. if a user inputs the phrase

*"All the world's a stage and all the men and women are ……....."*

Provided the model has read his *Seven Ages of Man,* the system using its generative capacity predicts the remaining words and completes the famous quote from *Shakespeare* as -

*"All the world's a stage and all the men and women are merely players"*

This generative capacity of the model, predicting and generating on the basis of data the model has seen before, is at the core of the foundational models[9]. Foundational models, as they are generating something, are a part of the field called Generative AI (figure no. 3), which essentially perform generation tasks (eg. predicting the next word in the sentence).

Apart from generation they can be programmed, by the process called tuning, to perform other tasks like classification. Tuning can be done by introducing a small amount of data, resetting

[9] Feuerriegel et al., *supra* note 5.

certain parameters and performing specific language tasks[10]. Further the model can work pretty well even in cases where very few data points are available by the process referred to as prompt engineering.

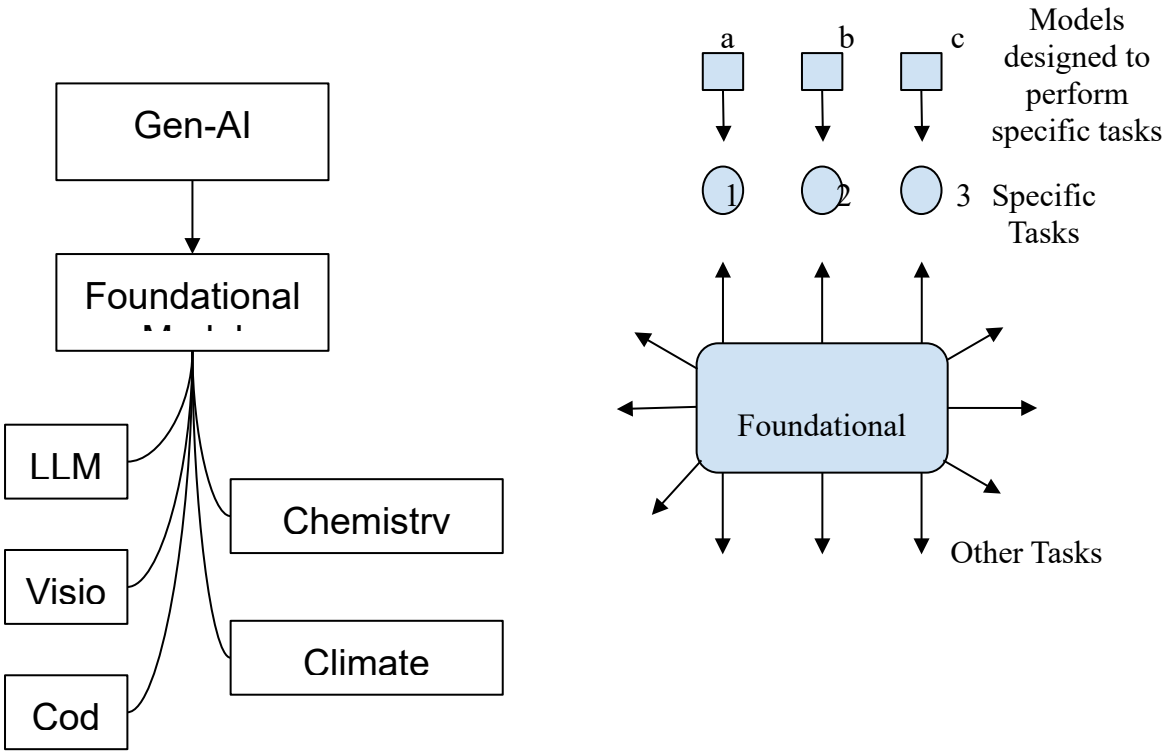


Figure No. 3 Type of Gen-AI Models- Foundational models and Large Language Models

Source: Author

These models come with its pros and cons. The advantages of these models are their efficiency, powerful performance and enhanced productivity. The disadvantages though play a crucial role in impending new companies to venture in the AI field. The *compute costs* of these models are immense. The models are very expensive to train which impedes a small enterprise to have its own model. The hosting, sustaining and functioning of the models is very expensive for big corporations too. The models also suffer from trust and biasness issues, depending upon the quality of data they are trained on.

Generally the data is retrieved from the internet and there is no fool proof mechanism to ensure that the data points are not influenced with biasness, hate speech, toxicity and material unsafe for children. Thus to conclude the Generative AI works on foundational models. These

[10] Cohan, *supra* note 8.

foundational models are not only on the language side (known as LLMs, eg. Chat-GPT) but are also on the visual side (eg. DALL-E), and can also be applied to other domains[11] like coding, chemistry and climate change. This much conceptual understanding about generative AI is sufficient and any further elaboration on the working of these models is outside the scope of this paper.

## 5. Section II

### 5.1 Liability

The process involved in image generation via a Gen-AI model involves many actors. Each actor such as programmer, data provider, trainer, operator and user have different roles to play in the said process. The question which emerges, in case of any conflict, is that on whom the liability can be imposed. Who can be held liable,out of these actors, in the wake of any wrong being committed by or through the process? The answer to this question is not at all easy to comprehend provided the complex nature of the technicalities involved along with other pertinent dependent factors such as risk control and risk attribution.

The traditional notion that the computers will do what one tells them to do is categorically challenged in the wake of emergence of disruptive technologies when the foreseeability is diminished and the output from computers is very different[12] from what the users had in mind. The era of diminished foreseeability calls for rethinking of the liability rules, importantly for three major areas of information technology, namely Robotics, Connectivity and Machine Learning (Artificial Intelligence) which have seen rapid advancements. Regulation of these technologies, which essentially means influencing the behaviour with respect to the technology vis-a-vis different stages along which the tech develops namely., the development stage, dissemination and application stage, results and application stage. Different areas of law, such as IP, competition, tort and security law, influence these stages differently.

All these areas function on humongous data, often coming from multiple sources, apart from the technology involved, regulation of which involves additional impediments such as inability to attribute the causation to one single actor. Further the programming coupled with training of the AI had brought forth a paradigm shift in the sense that the machine has evolved to learn on its own thereby interfering with the ability to predict and foresee the whole behaviour of the system. The impediments in foreseeability render the existing risks to undergo certain changes

[11] Feuerriegel et al., *supra* note 5.
[12] Guadamuz, *supra* note 1.

which shifts *risk control*[13] vis-a-vis the actors involved (figure no. 4). For Eg. in case of self driven cars the risk control of *producers* has increased and that of *operators* and *owners* have reduced.

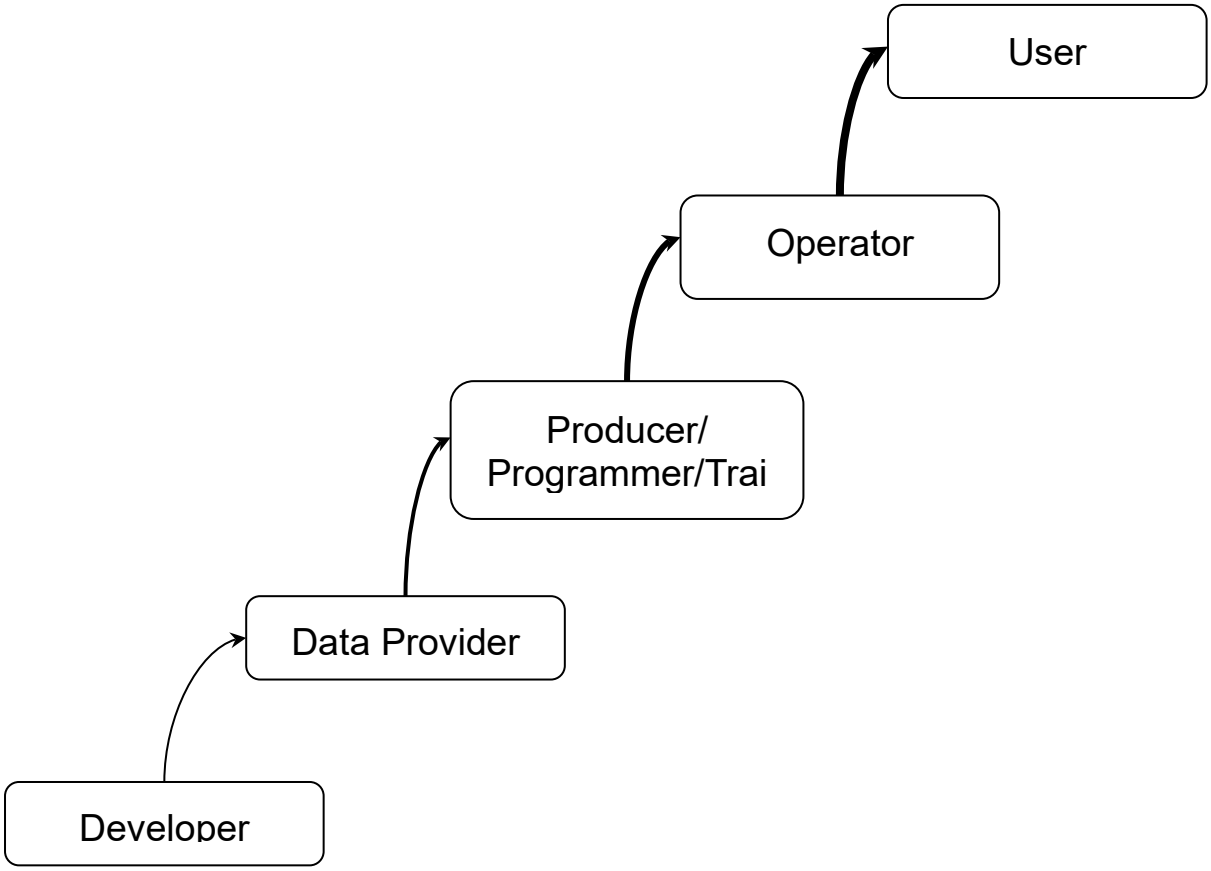


Figure no. 4 The Actors involved in an AI Model.

Regulation of technologies is crucial to mitigate the risks involved, create incentives and address stakes that lie in the development of technology. In order to regulate new technologies jurisdictions have resorted to direct and indirect regulation. Indirect regulation involves the technology to develop freely and be subjected to prevailing market conditions. Intellectual Property Law and Liability Rules[14] are regarded as indirect regulations which come handy in tech regulation and risk mitigation, owing to the insufficient knowledge of the state on how

[13] *Id.*
[14] Magdalena Tulibacka, Product Liability Law in Transition: A Central European Perspective (2016), https://www.taylorfrancis.com/books/mono/10.4324/9781315602318/product-liability-law-transition-magdalena-tulibacka (last visited Feb 11, 2025).

useful or risky the new technology might be, rendering the state unable to come up with a direct regulation.

The IP Law and the Liability rules are two sides of the same coin which function towards complete internalization of externalities. The IP Law particularly the patent law internalizes positive externalities while the liability rules internalize[15] the negative externalities. A harmonised effort to bring these two laws to function in a cohesive manner will help in building up an innovation friendly environment which incentivises people to bring up useful innovations and applications.

The EU AI Act Article 10 is an example of direct regulation which deals with data and data governance. Interestingly the Indian authorities are still on the fence about bringing a statute that exclusively deals with AI training data. However some pertinent laws that are currently in force in India are allegedly sufficient for regulation of the AI ecosystem.

The present liability rules deal with contractual liability, product liability[16], fault based liability, strict liability etc. The contractual liability stems from the contracts for supply of data. The contractual obligation binds the parties to provide data which is ethical, unbiased, clean, non-discriminatory and in accordance with the terms of the contract. In such contracts the transfer of the training data is the main obligation where data is supplied against consideration. Terms may also include providing access to data or data sources rather than providing data as such. These contractual obligations depend on *control* of the data supplier. But does the data supplier have complete control over the data?

On the other hand product liability can come into picture in case data sets are considered to be products. The product liability is enforced when the said products say consumer products turn out to be defective. In cases of infiltration technology if the information is *wrong* then the product may be termed to be defective. However in a leading case[17] it was adjudged that such scenarios attract the question of wrongful service and not defective product. Information products including softwares (AI) thus are out of the purview of the definition of *defective product.*

[15] *Id.*
[16] *Id.*
[17] See *VI v. KRONE - Verlag Gesellschaft mbH & Co KG CJEU C-65/20* wherein the court discussed liability in cases of defective products.

Contractual liability as seen above depends on control and product liability depends on the subject matter being a product. Training data lies midway on the spectrum wherein on one end is information and on the other is control (table no. 01). Thus training data does not does not come completely under the concept of contractual liability owing to lack of control and also not under product liability as training data sets are not products. It becomes crucial to appreciate at this stage the differences between information (simple data), training data and software.

| Simple Data | Training Data | Software |
|---|---|---|
| Analysed by humans | Used for training AI | Used for controlling computers |
| Does not control Machines | Indirectly controls machines | Directly controls machines |
| Causality via human knowledge | No human knowledge involved | No human in the loop |
| Can be treated as product | No conclusion so far on how to treat training data; as product or service | Comes in the category of service not product |

Table no. 01 Differentiating simple data, training data and software

The pertinent question emerges from the discussion above is how to treat the training data. The training data does not completely fit into the water tight compartment categories of product on one hand or software on the other. Therefore calls have been made to develop separate sui generis regimes different from product liability and contractual liability to address the emerging problems. Fault based liability is proposed to be an emerging and new liability rule in cases of AI. fault based liability is liability of negligence. It stems from damage and causation from the breach of a duty. Thus in cases where a duty of care[18] was breached this liability[19] can be imposed. The question to ask here is whether the data supplier had a duty to supply

[18] Donal Nolan & Ken Oliphant, Lunney & Oliphant's Tort Law: Text and Materials (2023), https://books.google.com/books?hl=en&lr=&id=bRSsEAAAQBAJ&oi=fnd&pg=PP1&dq=oliphant+ken+tort+law+2023&ots=qnW5hd5_Cn&sig=VxUNCZsAXJJRd-Ipxg5me8vgX20 (last visited Feb 11, 2025).
[19] *Id.*

immaculate data. If there was such a duty and it was breached, liability can be imposed. Strict Liability can also be imposed in cases of "high risks" AI models.

In conclusion the policy makers may endeavour to carve out sui generis rules to tackle the problem of liability in the cases of training data in AI models. Strict liability rules may be brought in for producers, programmers, trainers and also for suppliers of data particularly in cases of AI where high risk is involved such as in healthcare and education. Certain flexibilities and limitations shall also be incorporated in the rules in order to incentivise the suppliers to generate more data. Thus policies shall take care to balance out incentivisation and development of more immaculate data without stifling the growth of developing technology.

## 6. Section III

### 6.1 The Copyright Challenge

The copyright doctrine of fair use provides that there is no copyright infringement when protected works are put to certain uses regarded as fair. When copyright protected data is used for training an AI model, should there be any immunity? The stage of input as noted above brings up the issues that require scrutiny as to whether copyright infringement occurs when the model is being fed with the copyright protected material[20] as a part of its training process. It is also to be looked at whether such use of copyrighted material is exempted, meaning that whether such use will be deemed non-infringing.

At the stage of output the question is regarding the copyrightability of the works generated with the aid of AI. If copyright can be attributed to such work then who can be regarded as the author? In case copyright cannot be granted to such works, can these works be protected by neighbouring rights i.e. if there is no authorial copyright protection can there be a neighbouring right protection[21] to save the work from going into the public domain? Can the definition of "work" cover AI generated works? Can it be commercialised? Can you sell it as your creative output? These questions are interrelated but they have to be treated differently as the tests to check these problems are different. The questions at the stage of input are relatively more important than those at the stage of output and this article focuses on those questions only.

[20] Robert Brauneis, *COPYRIGHT AND THE TRAINING OF HUMAN AUTHORS AND GENERATIVE MACHINES*, 47 Columbia J. Law Arts 58 (2024).
[21] Lemley, *supra* note 3.

Allegedly it is an act of infringement when a copyright protected material goes into the model as a part of the training process[22]. Also, the AI models ardently require data to thrive. They are nothing without data input and require access to human source material. Therefore, on one hand, it becomes essential for the models to use existing copyrighted work for training and on the other hand there are allegations that such use will amount to infringement. Now how to harmonize these conflicting situations which require that technological advancement are not impinged upon and the rights of the copyright owner are also not infringed.

Every single process, from text to image creation, requires individualised scrutiny based on its own merits. The process involves several acts of copying and reproduction at several stages during the process, rendering occurrence of temporary and permanent copies. What, where and when is a copy an infringing copy in the system? The pertinent question to ask is whether all of these copies are infringing copies. Does the law exempt any of the copies? The copies that are fed to train the system are infringing? To answer this it is crucial to understand the principle of "text and data mining".

Machine learning, also known by the name of artificial intelligence, is the technology wherein, principally, the machine learns through data, algorithms and other user inputs and feedback and thereafter the ML mimics human intelligence. producing creative outputs. Analogy can be drawn with how a human learns using various similar methods. AI requires copies of works to learn and get trained to emulate human creativity. The training involves assessing the copies, reading the copies, preparing for reproducing the copies, analysing the copies and mining the copies. Does any of these constitute copyright infringement[23]? Or are there any exceptions to such infringement?

Everything is available on the internet which offers easy access to diverse media including audio visual works, literary dramatic works, cinematographic works etc. Further the models do not scrape everything off the Web. The nature of data sources and the manner in which data is collected are diverse. For extraction of data some methods involve making a permanent copy for use in generating works while some only make a temporary copy of the data. More the data, the better the models and much better the outputs. Training the AI requires accessing, reading, analysing, storing and copying the data to extract information[24]. The steps are collectively

[22] Guadamuz, *supra* note 1.
[23] Carys J. Craig, *The AI-Copyright Trap*, 100 Chic. Kent Law Rev. (2024), https://www.ssrn.com/abstract=4905118 (last visited Feb 11, 2025).
[24] Stefan Feuerriegel et al., *Generative AI*, 66 Bus. Inf. Syst. Eng. 111 (2024).

called data mining and are useful to train an untrained artificial intelligence model. The question here is whether these steps constitute copyright infringement[25] and if it does, do we have any exception or limitation for said text and data mining?

That data which is fed does not completely include only those works which are protected. Unprotected works such as ideas, numbers, stats and facts also form the bulk of the data. These datasets often comprise protected and unprotected works both. The unprotected works will be outside the purview of copyright infringement. Further no question of infringement arises if data is collected with authorisation of the owner or from open access databases or public domain or any other legitimate resources. Thus the trainers can avail data from legitimate and non infringing licensed sources. For an AI model only to rely on these sources makes it inefficient, oldschool and prone to bias towards data not fed to it. Mining and collection of data without permission will in principle infringe copyright and therefore trainers will have to avail the benefit of any exception, if any. In the case of mining and collection of data is the trainer performing acts that are covered under exclusive rights of the owner? Exclusive rights such as those related to exclusive reproduction (copying), distribution, adaptation, dissemination etc. What if the training process does not involve any coping, then obviously the right of reproduction is not infringed. In certain AI models only some attributes, encoded in the form of links, of protected work are scraped off and not the entire work is copied. Further these models in doing so often rely on temporary copies which are not kept or stored in the database. One can also say if such links are made it may be against the right to communicate to the public. However this communication is not regarded as to have been made to public in a strict legal sense. Therefore only links are extracted and copies are not made, hence no infringement of the rights of reproduction and dissemination as seen above. Thus the data set, as such, is not infringing.

Scholars argue that once the data set is used to train the AI it is no longer needed thereafter and the resulting trained model does not contain copies of the data set. Also the data set is essentially in a condensed form of information in a latent space . In a typical model images are taken in training process which are temporary copies and are encoded in a latent space. It is all a kind of lossy and compressed data and no copies of works are made in a legal sense. A question may arise, if in this process the right of adaptation can be said to have been infringed. In a training process what essentially happens is this, take all the things, break apart each thing

[25] Craig, *supra* note 23.

and categorise and situate similar components together. Now this is not a translation as the original work is completely gone. The content of data sets which are put to use for training AI and the content of the trained AI are very different from each other.

Till now the training data steers clear from copyright infringement. There is no copyright infringement in the process. However, still in some cases data collection or training the AI may infringe an exclusive right of the author. In that case the question to ask is whether there is an exception or limitation for such use?

The Infosoc directive of the UK states that copyright is not violated when creating a temporary copy that is transient or incidental[26], provided that this copy is essential to "a technological process, and its sole purpose is to facilitate the transmission of the work or for lawful use, and the temporary copy lacks *independent economic significance*[27]." This provision is a bit problematic. Firstly the term *lawful use* is subject to interpretations. Secondly the individual copy, though temporary, may not have individual economic significance whereas the value of combined accumulated information collected from such individual works is economically very significant. Though the copy is temporary and transient, it is very much essential for scraping of information without which the AI model does not exist. The question arises whether that copy for technological process is a legal use. Also the model is significant in terms of economy unlike single individual copies as aforesaid. It is not the exact work which is significant for the model but accumulated facts and information from the work. This brings up for discussion in future litigation to decide the individual value of work vis-a-vis collected value from accumulated works which essentially depends on hyper technicalities of training the model.

Another exception deals exclusively with collection of data itself, applicable for the training of a machine learning model. This is known as TDM (text and data mining) Exception. There are three models in vogue which provide TDM exemptions. A total exemption of TDM, including for commercial and scientific research, has been tried by jurisdictions[28], which brought with it, grave consequences. A Complete exemption can lead to huge losses in licensing revenues for copyright holders. Another model focuses on exemption for non-commercial research purposes. Here TDM is allowed for research purposes.

[26] Guadamuz, *supra* note 1.
[27] Section 28A CDPA (Infosoc Directive of the UK)
[28] The UK Intellectual Property Office opted for a total exemption for TDM. The notification ordering the said exemption was later called back due to grave repercussions in terms of huge revenue loss.

This exemption however may lead to the problem of *data-laundering,* also known as *academic washing*, discussed later. The third model also known as the hybrid model is a combination of the two wherein commercial purposes are exempted with certain restrictions. Renewed interests have emerged, owing to the emergence of technology, in the scholarly debates, to discern if data mining falls within *fair use.* Gathering the data, for non commercial research purposes, is covered by the exception, by analogy it can also be said to cover training the model as well. Gathering data and training the model are not conspicuously defined as different stages in law. The TDM exception for research has two fold problems, firstly, there is no clear cut legal or judicial opinion/interpretation of the term "research" and secondly, the problem of data laundering. Once the data is used for research, certain rules for the want of authenticity and transparency or for funding obligations may mandate disclosure of the data used in research, rendering the data to come within the public domain. Now the answer to the question whether an entity can use such data for commercial purposes, calming the data to be in the public domain, is not very clear. So far one thing is clear, that in order for training data to fall within fair dealing, there shall either be a temporary copy of the data or there be a TDM exception.

*Section 52* of the *Indian Copyright Act 1957* (the Act) deals with copyright exceptions. *Section 52 (1) (a) (i)* and *Section 52 (1) (b) and (c)* are crucial for understanding the scenario in light of the Indian jurisdiction. Copyrighted works can be used for research purposes in accordance with S. 52(1)(a)(i)[29]. Further, explanation[30] appended to this clause clarifies that even storage of a protected work's transient or temporary copy for the research purposes, will not constitute copyright infringement. Clause (b) of the section comes handy in the training AI process. Clause (c) is categorically important in the present case.

We saw above, how links to the data, used for training the AI, are stored, in condensed form. Providing such links, and storage of temporary copies to provide such links, does not constitute copyright infringement. The proviso attached to this clause may be used to solve the problem of data laundering. In the absence of any legislation to deal exclusively with the problems discussed so far leaves the opportunity for the curts to discuss and evolve the jurisprudence related to the intersection of Gen-AI models and Copyright laws. The author is of the view that

[29] *"Copyright Act 1957 Section 52. Certain acts not to be infringement of copyright. (1) The following acts shall not constitute an infringement of copyright, namely,-- (a) a fair dealing with any work, not being a computer programme, for the purpose of– (i) private or personal use, including research;"*

[30] *"Explanation.-- The storing of any work in any electronic medium for the purposes mentioned in this clause, including the incidental storage of any computer programme which is not itself in infringing copy for the said purposes, shall not constitute infringement of copyright."*

the domain calls for a renewed approach to rethink copyright law doctrines in the wake of these emerging technologies. This paper only discussed the input questions laying the foundation for future research pertaining to output questions and also other foundational legal doctrines of copyright regime such as idea expression dichotomy and substantial similarity doctrine to prove infringement.

## 7. Conclusion

The recent nobel prize recipient, Geoffrey Hinton, expressed grave concerns over his own work which led him winning the coveted prize. His work on Artificial Intelligence and Deep Learning has made him state that he regrets his life's work. Why? Even more so on similar lines Elon Musk, who has also regarded AI as being a risk to humanity, compels one to end up in a cassandra moment. The regulation of AI is an urgent global issue now and India is no exception. The EU AI Act is a groundbreaking piece of legislation that works on a risk based approach which may work as a point of reference and an inspiring model for building a framework here in India. Though, India has already made considerable strides in addressing these issues by its IndiaAI mission but much is yet to be achieved and what a law governing AI in India could look like remains to be seen.

This paper looked at serious policy concerns about copyright law as posed by the Generative AI, revolving around infringement, text data mining exception, bias, concentration, competitiveness and performance. These concerns if not addressed impact various stakeholders including members of art industry, cultural groups, photographers, filmmakers, artists, music industry members, libraries and other private and public entities. As is apparent in the litigation involving disputes related to Gen-AI, mushrooming around the world. The cases predominantly deal with issues related to licensing. This is not surprising as huge investments and other capital are at stake, thus issues pertain largely to money. India has not so far witnessed any significant litigation related to the area. Until that time, it is advised that endeavours are made to flourish the current IndiaAI mission and more and more public consultation and round table conferences are held to build an ethical and responsible AI environment.

The author suggests that a one-size-fits-all approach will be detrimental for regulation of Gen-AI Models. The thread of trust. transparency, inclusivity, collective dialogue and regular deliberations from all stakeholders shall continuously flow through the paradigm of policy and law making regarding the regulation of the Gen-AI models. Necessary provisions relating to flexibilities with respect to emerging technologies shall find place in the laws in order to

encourage the entrepreneurial ecosystem. On the other hand, international harmonization attempts shall be made with respect to the enterprises that are well established. Since the disruptive technologies of today's age are affecting all the sectors in the society, the stakeholders shall ensure adoption of inclusive approach while making legislations and regulations, taking into account the legitimate and required access of users and balancing it with the legitimate interests of the right holders. A proportionate legislation is the need of the hour that can function as a great leveler to balance the conflicting interests emanating from different stakeholders. As is often said with respect to the slow pace of the legislations as compared to that of technology, a regular and continuous assessment of the laws and regulations by calling upon working groups deliberations and round table conferences by diverse members and stakeholders to regularly check upon the developments in the concerned area, keeping in mind inclusivity and transparency.