TRAINING AI AND COPYRIGHT INFRINGEMENT: WHERE DOES THE LAW STAND?

Ravindra Kumar & Professor Pankaj Kumar, Manipal University Jaipur

ABSTRACT

Artificial Intelligence (AI) refers to computing systems that can make autonomous decisions based on stimuli, in a manner similar to human beings. It has various real world uses in practically every field from finance to healthcare. However, the growth of artificial intelligence comes with new legal challenges, particularly relating to intellectual property rights. AI systems are trained using massive amounts of data to recognize patterns and make intelligent decisions or predictions. This data often consists of copyrighted works, raising the question whether use of such works without authorization for machine learning constitutes copyright infringement. The copyright laws in diJerent jurisdictions lack clarity and uniformity as to the existence of a specific exception for the use of such works to train AI applications. The paper thus analyses the rules of access to copyrighted data in the United States copyright law, focusing on the fair use doctrine and its judicial interpretation, including the recent Google Books case. The subsequent part of the paper compares the US approach with the copyright laws of the European Union, looking particularly at the impact of the text and data mining exception introduced by the 2019 EU Directive on Copyright in a Digital Single Market. In light of the inferences drawn, the final part of the paper attempts to assess the Indian copyright law and explores the future course of law and policy on the issue, highlighting the role of the World Intellectual Policy Organisation.

Keywords: Artificial Intelligence, Copyright Infringement, Text and Data Mining, Fair Use Doctrine, Deep fake, Expressive Use, Training.

INTRODUCTION

Artificial Intelligence as a technology appeared for the very first timein 1956 and the term AI was coined by John McCarthy.[1] Artificial intelligence is concerned with the demonstration of intelligence in the machines and enables machines to imitate the learning and decision making capabilities of the human mind.[2] Machine learning is a concept developed by Arthur Samuel[3] in 1959.[4] It is a subset of AI which enables the machines to learn from enormous amount of data and make predictions without being expressively programmed. The 'Logic theorist' was the first AI program which could imitate the problem solving skills. Some major AI-based breakthroughs such as 'IBM Deep Blue', 'Kismet', 'Dragon Systems' and 'Alpha Go' were developed after 1980s.[5] Artificial intelligence has grown exponentially in fields ranging from medical to space exploration.[6] There are various examples of AI works which use great amount of data, such as the 'Next Rembrandt', 3D printed painting is generatedby artificial intelligence by collecting and analyzing great amount of Rembrandt works, and Google's 'Deep Dream Generator' which helps generate paintings by merging paintings with an uploaded picture in minutes.[7] Artificial intelligence has shown tremendous achievements not just in non-expressive uses but also expressive uses.[8] AI is increasingly learning about the human expression and generating expressive works such as books, prose and poems at par with copyright protected works.[9]

Arend Hintze has categorized artificial intelligence into four categories. First category is 'reactive machines' which work withoutmemory and past experience. They can only react to existing situations. An example of a reactive machine is IBM's 'deep blue' which is designed to play chess against a human. Second is 'Limitedmemory' which collects data and retains some

[1] John McMarthy, "What is artifcial intelligence?" STANFORD UNIVERSITY, (Nov. 12, 2007) available at http://jmc.stanford.edu/articles/whatisai/whatisai. pdf (visited on Feb. 24, 2021).
[2] Andy Peart, "John McCarthy, The Father of Artificial Intelligence (AI)" *available at* https://www.artificial-solutions.com/blog/homage-to-john-mccarthy-the- father-of-artificial-intelligence (visited on Feb. 24, 2021).
[3] Ayees Myat, "Machine Learning: Intro to the Future of Computing" *available at* https://www.seamgen.com/blog/machine-learning-future-computing/ (visited on Feb. 24, 2021).
[4] Russ Pearlman, "Recognizing Artificial Intelligence (AI) as Authors and Investors under U.S. Intellectual Property Law", 24 Rich. J.L. & Tech. (2018).
[5] Bruce E. Boyden, "Emergent Works in AI", 39 COLUM. J.L. & ARTS 377, 378 (2015).
[6] Annemarie Bridy, "Coding Creativity: Copyright and the Artificially Intelligent Author", 5 STAN. TECH. L. REV. (2012).
[7] Shlomit Yanisky-Ravid & Samuel Moorhead, "Generating Rembrandt: Artificial Intelligence, Accountability and Copyright", MICH. ST. L. REV (2017).
[8] Gideon Lewis-Kraus, "The Great A.I Awakening", N.Y. TIMES *available at* https://perma.cc/SBQ9-K899 (visited on Jan. 21, 2021).
[9] Daryl Lim, "AI & IP: Innovation & Creativity in an Age of Accelerated Change", 52 AKRON L REV813 (2018).

amount of information from previous data. AI can create new knowledge using the previous events. For instance, autonomous cars use pre-programmed data. Also, such self-driving cars gather data from nearby vehicle's speed, the direction of other vehicles, data about lane parking, and other related data to observe their surroundings and adjust their driving as necessary. Third category of AI is 'theory of mind' which attempts to imitate the mental state of human beings. No such AI system is in function yet and even world's most popular AI based robot 'Sophia' isn't capable of understanding human emotions in their entirety.[10] Fourth category of AI is 'self-awareness' and such machines are complex and highly sophisticated systems as they have human level consciousness.[11]

TRAINING ARTIFICIAL INTELLIGENCE:

Machine learning algorithms derive huge amount of data which may also include copyrighted work while providing training to the AI.[12] Machine learning algorithm follows three types of learning i.e., supervised, unsupervised, reinforcement. Under supervised learning the data is labeled and structured to generate mapping function which provides expected output. Unsupervised learning is when the data is not classified and structured. Reinforcement learning is when the machine is being exposed to a new environment and it takes prompt decisions and learn from its past mistakes.[13] Generally speaking, the training of AI can be divided into 6 stages. The first stage is data ingestion whereby enormous amount of data is being collected.[14] After the data is collected, any inconsistent, incorrect and skewed information has to be removed.[15] The source data has to be filtered and any prejudices or biases shall be removed. Further, data has to be formatted to achieve uniformity and remove anomalies. The process of data formatting should best fit the machine learning model. Further, data is converted into patterns which provide with relevant information which can be fed into the

[10] Calvert Solen, "Uncanny humanoid robot 'Sophia' to enter mass production" *available at* https://eandt.theiet.org/content/articles/2021/01/uncanny- humanoid-robot-sophia-to-enter-mass-prodution/ (visited on Jan. 24, 2021).
[11] Elizabeth Rocha, "Sophia: Exploring the Ways Al May Change Intellectual Property, Protections", 28 Dr-PAUL J. ARt, TECH. & INTELL. PROP. 1. 126 (2018).
[12] Kevin P. Murphy, "Machine Learning: A Probabilistic Perspective", MIT Press (2007).
[13] Michael Copeland, "What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?", NVIDIA BLOG (July 29, 2016), *available at* https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial- intelligence-machine-learning-deep-learning-ai/ (visited on Feb. 21, 2021).
[14] Amanda Levendowski, "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem", WASH. L. REV. 15 (2017).
[15] *Ibid.*

learning algorithms.[16] Lastly, the data has to be divided into 'training set' and 'evaluation set'. The'training set' consists of the data which is used by the model to makepredictions.[17]

In few instances, the training of artificial intelligence is consuming unprecedented amount of copyrighted works,[18] which includes prose, novels. However, the generated data by the AI is oftentimes similar to the copyrighted work which raises a flurry of questions about the legality of work under copyright laws.[19]

POTENTIAL COPYRIGHT INFRINGEMENT IN TRAINING AI:

Training AI can lead to issues related to copyright infringement. Themachine learning model use unauthorized input datasets and make digital copies of such works which is a reproduction of previously copyrighted work.[20] Literal reproduction of the copyrighted work canalso happen during the training process. As same input datasets are copied multiple times during the learning process.[21] In such a process, infringing copies are being made multiple times. On the other hand, non-literal reproduction in models and datasets happens at the input stage, as an enormous amount of datasets, which are unauthorized copyrighted work, is being fed into the system. Lastly, the output willbe substantially similar to the input data, thus it has the potential to infringe the rights of the copyright holder in his work.[22]

One such instance of potential copyright infringement duringthe training of AI is in 'Deep fake technology'.[23] Such a technology can manipulate images, videos, or music using neural networks to such a degree that it is indistinguishable from the original work. Deepfakes are developed using adversarial training. The training process consists of two neural networks i.e.,

[16] Temboo, "Smart Predictions with Amazon Machine Learning" *available at* https://medium.com/@temboo/make-smart-predictions-with-amazon-machine- learning-ad4fa464947 (visited on Jan. 26, 2021).
[17] Zohar Karnin, "Elastic Machine Learning Algorithms in Amazon SageMaker", WSDM (2016), available at 377--386. https://doi.org/10.1145/2835776.2835781 (visited on Jan. 26, 2021).
[18] Peter Jaszi, "On the Author Effect: Contemporary Copyright and Collective Creativity", 10 CARDOZO ARTS & ENT. L.J. 293, 294 (1992).
[19] James Grimmelmann, "Copyright for Literate Robots", 101 IOWA L. REV. 657, 665 (2015).
[20] Sujith Ravi, "On-Device Machine Intelligence", Google Research Blog (Feb. 9, 2017) available a*t* https://perma.cc/WQ8L-WS5D (visited on Jan. 23, 2021).
[21] Brendan McMahan & Daniel Ramage, "Federated Learning: Collaborative Machine Learning without Centralized Training Data", Google Research Blog *available at* https://perma.cc/XVA2- J96J(visited on Jan. 24, 2021).
[22] Andres Guadamuz, "Artificial Intelligence and Copyright", 23 WIPO MAG (2017) http:/www.wipo.inti stpo magazineren12017/05/article_0003.html [https:/iperma.cc/SD4Q-KE9E (visited on Feb. 23, 2021).
[23] R Chesney, DK Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" 107 California Law Review 17 (2019.)

generator neural network and discriminator neural network.[24] The generator neural network tries to generate fake data by analysing the patterns and the discriminatorneural network tries to detect the fake data.[25] This process is repeatedmultiple times until the discriminator neural network can no longer distinguish between the fake data and the original data.[26] Such deep fake technology raises questions about the legality of the workcreated by AI as it uses enormous amount of copyright protected data.[27] There is no regulatory mechanism in place to supervise the works created by deep fakes due to absence of laws on this subject matter.[28] Deep fakes can be tested on the four-factor test of fair use doctrine. Some works created by deep fake technology have the potential to affect the market of the copyright author and infringe upon their economic rights.

US FAIR USE DOCTRINE AND AI:

The United States doesn't have a specific exception with respect to Text and Data Mining (TDM) activities. TDM refers to a computationalprocess which includes collection and analysis of data to find patterns and discover new information using unstructured data. However, such TDM activities are broadly covered under the 'Fair Use Doctrine'. The fair use doctrine as stipulated under Section 107 of US copyright law which determines whether the work amounts to copyright infringement or not by considering the four-factor test. The fair use doctrine is flexible with regard to recent technological developments in the field of Artificial intelligence, especially with regard to the TDM. The training data can be categorized into expressive and non-expressive use wherein expressive use includes learning from creative works whereas non expressive use is merelya collection of information which lacks expression and creativity. Such data is collected and analyzed to find patterns. TDM is one such example of non-expressive use and is generally protected under the fair use doctrine. Four factor test is applied under fair use doctrineto determine whether the work amounts to copyright infringementor not.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

[24] Samuel Calret, "Deepfake Detection Challenge Datase*t"*, FACEBOOK AI (Jul. 22, 2020, 11:54 PM), available at https://ai.facebook.com/datasets/dfdc/ (visited on Jan. 22, 2021).
[25] *Ibid.*
[26] James Vincent, "Facebook contest reveals deepfake detection is still an unsolved problem" THE VERGE (Jul. 22, 2020, 11:56 PM), *available at* https://www. theverge.com/21289164/facebook-deepfake-detection-challenge-unsolved- problemai(visited on Jan. 22, 2021).
[27] *Ibid.*
[28] *Supra* note 23.

potential market.[29] The important factor to be considered is the 'transformativeness' and 'commercialization of the work. The more transformative the work, less will be the importance of the other factors.[30]

Under the fair use doctrine, it is pertinent for the copyright owner to prove that he owns a valid copyright. In *Cartoon LP* v. *CSC Holdings*[31], court held that there is a difference between voluntarily making a copy and issuing a command to the computer which obeys the command. Such non volitional uses of copyrighted work are 'intermediate operational use'. Such non volitional useof copyrighted works by the computer is not creative in nature and does not amount to infringement. In *Burrow-Giles LithographicCo.* v. *Sarony*[32], court observed that mere mechanical processes, encodings or transcoding are not expressive works and does not amount to copyright infringement. *Sega* v. *Accolade*[33], is the first case which recognized the non-expressive fair use. In this caseAccolade, video games developer company purchased some games to copy the functional code without the authoritisation of Sega and later they reverse engineered the games to get the functional code. Court held that the intermediate copying of functional code will be protected under the fair use doctrine as it is a functional element not amounting to copyright protection.[34] In *Kelly* v. *Arriba*[35], Arriba produced thumbnail images of Kelly. For that purpose, they used a web crawler which visited various websites to collect images and turned them into thumbnails. These images were featured on the Arriba search results. Court held it to be transformative fair use.In *Authors Guild* v. *Google*[36], google scanned millions of books including some copyrighted work. Google trained the AI by giving acorpus of books to make it machine readable and to perform keywordsearch for easy accessibility by the users. Second Circuit Court held that the scanning of books is transformative fair use as the purposeis distinctive than the original. Google books provides

[29] Pierre N. Leval, "Towards a Fair Use Standard", 103 HARV. L. REV. 1105, 1111 (1990).
[30] Benjamin L. W. Sobel, "Artificial Intelligence's Fair Use Crisis", 41 COLUM.
J.L. & Arts 45 (2017), Laura A. Heymann, "Everything Is Transformative: Fair Use and Reader Response", 31 COLUM. J.L. ARTS 445 (2008).
[31] *Cartoon Network LP, LLLP v. CSC Holdings, Inc*., 536 F.3d 121, 129-30 (2d Cir. 2008).
[32] *Burrow-Giles Lithographic Co. v. Sarony*, Ill U.S. 53 (1884).
[33] *Sega Enterprises Ltd. v. Accolade, Inc*., 977 F.2d 1510 (9th Cir. 1992.
[34] *Ibid.*
[35] *Kelly v. Arriba Soft Corp*., 336 F.3d 811 (9th Cir. 2003); *Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cit. 2007).*
[36] *Authors Guild v. Google Inc*., 804 F.3d 202 (2d Cir. 2015).

information about the books which is not the expressive part of the books. Hence,it doesn't amount to copyright infringement. [37]

Fair use doctrine excuses the use of massive amount of copyrightable work during training of the artificial intelligence. However, it should be used sparingly and judiciously. In *Harper & Row Publ* v. *Nation Enters*[38], court observed that fair use should not be interpreted broadly as to swallow the commercial value of a copyrightable work by failing to fully analyze the four-factor test.

Expressive Uses of Artificial Intelligence:

The use of copyrightable work in training of AI is not just restricted to non expressive use but has extended to expressive uses as well. Such expressive uses of AI poses some serious questions on copyrightinfringement and rights of copyright owner. Enormous data which is used for training purposes is further being exploited by third parties for targeted advertisements. There is a thriving market developed forthe input training data which fuels on expressive machine learning.[39]Unbridled and unrestricted consumption of copyright protected data for expressive use of machine learning can potentially deprive the authors from their rights and lead to economic losses. For instance, Jukedeck, an AI application through which users can compose songs,offers reasonable licensing options to the users.[40] The training data used by Jukedeck includes some copyright protected musical works.Applying the four factor test of fair use doctrine, Jukedeck has the potential to jeopardize the market of sound recordings and licensing,due to low rates of licensing and sound recordings.[41] There is a dearthof judicial pronouncements and legal mechanisms to regulate the training of AI for expressive use. Further, it is crucial to understand the legality of expressive uses of machine learning under the fairuse doctrine to protect the interest of the copyright owners and AI developers.[42]

[37] Maurizio Borghi & Stavroula Karapapa, "Non-Display Uses of Copyright Works: Google Books and Beyond", 1 QUEEN MARY J. INTELL. PROP. 21, 32-37 (2011).
[38] 471 U.S. 539, 560 (1985).
[39] Paul Ratner, "New Google AI Program Talk Like a Human and Write Music", BIG *available at* https://bigthink.com/paul-ratner/listen-to-new-google-ai- program-talk-likea-human-and-write-music https:/perma.cc/4J2C-75MH (visited on Feb. 22, 2021).
[40] Edward Rex, "Jukedeck: Building a Creative AI business", *available at* https:// entrepreneurship.blog.jbs.cam.ac.uk/author/ed-newton-rex/(visited on Feb. 24, 2021).
[41] Create unique, royalty-free soundtracks for your videos, JUKEDECK, *available at* https://perma.ccIL6EW9K4L (visited on Feb. 22, 2021).
[42] *Supra* note *40.*

EUROPEAN LEGAL PERSPECTIVE AND AI:

Before delving into the issues relating to copyright infringement in the course of developing AI systems, it is important to understand the European Union (EU)'s policy on artificial intelligence. The European Commission has recognized the economic implications of AI technology and has tried to direct its policies towards increasing Europe's competitiveness in the international AI market by way of greater support and funding for research and development in the continent. Simultaneously, there is a focus on putting a suitable legal framework in place that accounts for emerging concerns such as privacy, transparency, safety and accountability. Yet, little attention ispaid to intellectual property aspects in the EU's overall AI policy[43]. Itis most likely a consequence of this gap that Europe finds itself lagging behind Asia and North America both in terms of private investment in AI as well as patent applications for AI technologies[44]. In view of this situation, intellectual property policy assumes importance as a tool to foster technological innovation in AI. As explained in the first part of the paper, machine learning requires the developer to have access to a large dataset of good quality. This inarguably demands acopyright regime that can facilitate such access in an uncomplicatedand cost effective manner. In other words, to provide an impetus to the growth of AI, copyright laws need to be industry friendly. However,there is potential for conflict between the desire to stimulate the development of AI on one hand, and the necessity of protecting the interests of creators of the original works that constitute the training dataset. Therefore, a well-designed copyright regime should resolve the tensions between copyright protection and contemporary text and data mining practices used to train AI systems.

Text and data mining is an essential part of the machine learningprocess. It is a technique of collecting structured information from alarge amount of data, which helps to identify patterns and correlations.Works that are not copyright protected can be mined freely but manyTDM activities use copyright protected works and this is where copyright infringement may occur. TDM may involve making eithertemporary or permanent copies of works, thereby interfering with the reproduction rights of the copyright holder. In this respect, there are divergent views as to whether this should invite liability for copyright infringement[45]. In a situation where the AI developer is using copyright protected works to generate new expressive works, such use

[43] Joseph Straus, "Artificial Intelligence-Challenges and Chances for Europe", *29* European Review (2020).
[44] *Ibid*, at 149.
[45] Enrico Bonadio & Luke McDonagh, "Artificial Intelligence as producer and consumer of copyright works: evaluating the consequences of algorithmic creativity", I.P.Q. 112, 13(2020).

cannot be deemed fair, especially if the output is exploited commercially. The market of the right holder would inevitably be affected in an unfair way. Moreover, a scenario wherein the developer has the right to exploit the AI product unaccompanied by any liability for copyright infringement would give rise to inequitablecircumstances. There is also the risk of enabling a dual copyright system that disadvantages human creators, who cannot avail the same concessions as the AI algorithms. On the other hand, there is aplethora of arguments in favour of exempting TDM from copyright infringement. The temporary or permanent copies of protected worksmade for technical purposes in order to train AI may not necessarilyfall within the exclusive rights of the copyright holder. Not allsuch reproductions are intended to be substituted for the work ofthe author, meaning that they do not compromise on the essential interests of the author. Furthermore, the reproductions so made are not, in many cases, even communicated to the public. Not allowing for the exemption can also result in biased AI algorithms[46]. By virtueof the nature of copyright protection, current and contemporaryworks are not generally available in the public domain. The useof archaic material in training often leads to the perpetuation ofold fashioned beliefs and attitudes. Most importantly, the public's right to information, which includes the right to conduct and access research, is hampered if AI training is subjected to stringent liabilityfor copyright infringement.

In view of all the above considerations, it is imperative thata proper balance be found between the rights of the copyright holder and the rights of the developer. It is in this context that EU's copyright regime must be assessed. As on the date of writing this article, the World Intellectual Property Organisation doesn't offer normative guidance as to whether the use of data from copyright works without permission amounts to infringement. As a result of this, different jurisdictions around the world have adopted differing stances on this issue. Recognizing the importance of the free flow ofdata with respect to AI applications, many countries have enacted exceptions in their copyright laws for TDM but the nature and scopeof these exceptions vary significantly. The previous section of the paper has discussed in detail how the United States' fair use doctrine regards the issue. A detailed analysis of TDM exceptions globally isbeyond the scope of the paper, but it shall suffice to state that generaltrend is to circumscribe the exception in narrow terms by allowing room for a multitude of restrictions in its application[47]. Restrictions on type of works and the range of rights to

[46] Michael W. Carroll, "Using Fair Use to Reduce Algorithmic Bias", 2019 JOTWELL: J. Things WE LIKE 1 (2019).
[47] Sean Flynn et al., "Implementing User Rights in the Field of Artificial Intelligence: A Call for International Action", EIPR 393,4-6 (2020).

which the exception extends, restrictions on commercial use and cross border transfers, the requirement of lawful access to the copyright protected content aswell as other technical and contractual restrictions serve to limit the scope of the exception. Whether exceptions drafted in this manner areconducive to the expansion of AI technology requires investigation.

Legal Framework on Copyright in the EU:

The European Union (EU) doesn't recognize the concept of fair use and the related concept of transformative use in the same way as the United States. The starting point of an examination of the EU copyright law pertaining to machine learning is the 2001 Information Society Directive that sought to harmonize copyright laws within Europe, particularly on copyright exceptions. The Directive exemptstransient copies of a mechanical nature from the scope of copyright infringement, provided they are necessary to the technological process in question and have no independent creative value, and the manner of accessing the copyright protected content is lawful. In machine learning, the system may make temporary copies of the input material for running it across the neural network, which do not need to be stored permanently. Thus, the exception under the Directive is applicable to the AI training process. However, the use of copyright works in this context must satisfy a three step test[48], which basicallyrequires that the aforementioned exception be applied 'in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the right holder". The first part of the three step test essentially means that the exception should have limited application for a specific purpose. The second part of the test i.e.non interference with the normal exploitation of the work causes some confusion unless it is assumed that the use must be such that itinterferes with the active commercial exploitation in relevant forms to a significant extent[49]. The third part of the test refers to absence ofunreasonable prejudice to the legitimate interests of the right holder implies that the use should not lead to unreasonable losses of incometo which he is rightfully entitled[50]. Further, adequate compensation may be necessary to meet this condition.

More recently, there is a specific TDM exception introduced bythe 2019 EU Directive on Copyright in the Digital Single Market available in respect of the right of reproduction. This

[48] Ted Shapiro & Sunniva Hannson, "The DSM Copyright Directive- EU Copyright Will Indeed Never Be the Same", EIPR 404, 6 (2019).
[49] Daniel Gervais, "Exploring the Interfaces between Big Data and Intellectual Property Law", 10 J. INTELL. PROP. INFO. TECH. & ELEC. COM. L. 3 (2019).
[50] *Ibid.*

Directivehas brought sweeping changes to the existing copyright regime in Europe[51]. In a way, it can be seen as a setback for copyright protection in the continent as well as for right holders' groups. While the 2001 Directive was not an ideal system in terms of either copyright protection or technological innovation, it provided a semblance of balance by restricting the scope of the exceptions and providingfor contractual freedom to right holders to prevent the applicationof the exception to their works. This balanced approached has been undermined, to an extent, by the 2019 Directive. The first relevant substantive provision of the new Directive is Article 3. Article 3 contains the TDM exception for the purpose of scientific research. Research organisations and Cultural heritage institutions can make useof the exception. Research organisations can be universities, research institutions or any organisation that carries on research activities on anonprofit basis. Collaborative research efforts through public private partnerships can also avail themselves of the exception. However, theTDM exception is only applicable if the researchers have obtained access to the works being mined in a lawful manner i.e.by a contractto that end unless there is an open access policy. The exception grants the right of reproduction, and extraction and sui generis rights (in case of a database) to the researchers. They are, however, bound to keep the data secure and safe from unauthorized access by others. Further, Article 7 states that any contractual provision that seeks to curtail the operation of this exception will be invalid and unenforceable. The Directive also seeks to ensure that technologicalprotection measures do not narrow the exceptions provided therein. This is a marked divergence from the 2001 Directive. Article 4 of the 2019 Directive sets forth a broader exception than provided for in Article 3. Unless expressly reserved by the authors, the right of reproduction can be exercised by any category of beneficiaries, not necessarily researchers. The Directive also contains a private user exception which can possibly be claimed by individual researchers. Another key feature of the 2019 Digital Single Market Directive is the strengthening of the public domain. Article 14 of the Directive provides that once the copyright in a work of visual art has expired, any material derived from that work shall cease to be protected by copyright. This is a check on the attempt by authors to extend the period of copyright protection by creating a non-original derivative work.

However, the TDM exception available under the 2019 Directive suffers from several shortcomings which bring into question the utility of the exception for AI programmers[52]. Firstly,

[51] *Supra* note 29.
[52] Christophe Geiger et al., "Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data? Legal Analaysis and Policy Recommendations", IIC 814, 14-18(2020).

the TDM activities can only be carried out freely by research organisations or cultural heritage institutions only for the purpose of scientific research. Private companies that want to develop AI systems are at a disadvantage andthey can only engage in TDM activities for AI training under the exception given in Article 3 under a public private partnership and even then, a product developed through such a partnership may not be allowed to be commercially exploited. Therefore, the exception actually fails to create a climate of AI innovation that the industry demands. Even the more expansive exception under Article 4 that can seemingly be availed by businesses and not for profit entities is constrained by the fact that right holders can expressly deny the rightto mine. Most importantly, the Directive lacks clarity and certainty as to what sort of TDM acts are covered under the exception and how they are to be performed. There has not been enough progress towards implementation of the TDM exception across all EU member states so there is a lack of uniformity in the law in the continent, which adds to the challenges faced by AI researchers. The Directivein its present form leaves much to be desired from the point of view of researchers and technology companies. Many clarifications are required and ambiguous areas need to be addressed before it holds any real significance in terms of creating a harmonised and unified single digital market in Europe.

INDIAN LEGAL POSITION ON AI:

There is no comprehensive legal framework to regulate and supervise the rapidly growing artificial intelligence industry. Some initiatives has been taken by NITI Aayog by releasing a policypaper 'National Strategy for Artificial Intelligence' which covers five core areas for AI application such as education, healthcare, smart cities and infrastructure, agriculture and bandicoot robot.[53] A committee headed by V. Kamkoti is also formed to promote researchand development in the field of artificial intelligence and setting up a National Artificial Intelligence Mission.[54] The Policy paper also prescribes self-regulation and principles of transparency, privacy, equality, safety, inclusivity and accountability to be followed bythe members of AI industry. However, there still is no clarity on the legality of TDM and using data for training AI under the existing copyright laws. Section 52 of Copyright Act lays down an elaborate list of works which are covered under fair dealing and doesn'tamount to copyright

[53] National Strategy On Artificial Intelligence, *available at* http://niti.gov.in/ national-strategy-artificial-intelligence#:~:text=The%20Strategy%20is%20 termed%20%23AIForAll,of%20Sabka%20Saath% (visited on Feb. 23, 2021).

[54] Central task force on AI recommends setting up of N-AIM *available at* https:// indianexpress.com/article/india/central-task-force-on-ai-recommends-setting- up-of-n-aim-5114130/ (visited on Feb. 21, 2021).

infringement. There is no specific mention of TDM activities or training of artificial intelligence. Although, section 52(1)(a)[55] can be exercised to protect TDM activities as it permits theuse of literary work for private or personal use, including research, criticism or review or news reporting. This applies to reproduction of a current economic, political topic or incidental storage of electronic links, provided that it is not prohibited by the copyright owner. The protection under fair dealing doesn't extend to commercial purposesas held in *Saregama India Ltd. & Ors v. Alkesh Gupta & Ors and Tips Industries Ltd v. Wynk Music Ltd. & Anr*[56]. TDM activities whendone for non-expressive use are protected under section 52 only to theextent of research and development not commercial purposes.

CONCLUSION:

The European copyright law needs reform in order to improve innovation in AI. The TDM exception, being limited to non- commercial use, is unnecessarily too narrow. Limiting beneficiaries of the exception to include only research organisations does notsuit the market realities, hindering meaningful work being done by startups and other unaffiliated researchers. In the long run, this can set the clock back on the progress of AI technology. To avoid this pitfall, any person who has access to a copyright work should be allowed to mine freely as long the requirements of the three step test are satisfied. Insistence on lawful access may be counterproductive as it leaves researchers or developers at the mercy of the right holder,who may very well engage in rent seeking behavior by demanding thepayment of exorbitant sums of money as license fees. This will cause immense hardship to institutions that do not have abundant funds at their disposal, leading to a situation where budgetary constraints mayimpact the quality of the research output. An additional concern in the EU is the delay in implementation of the Directive by member states, resulting in a fragmented copyright regime in the continent which is a bane to researchers who need some legal certainty in order to undertake TDM activities. Although the 2019 Directive provides a legal basis for TDM activities, the TDM exception must be expandedif it is to meet the EU's policy goals on artificial intelligence.

[55] Section 52(1)(a) a fair dealing with a literary, dramatic, musical or artistic work Private use including research.

[56] Divij Joshi, "Crawl Cautiously: Examining the Legal Landscape for Text and Data Mining in India" *available at* https://spicyip.com/2020/06/crawl- cautiously-examining-the-legal-landscape-for-text-and-data-mining-in-india- part-i.html (visited on Feb. 21, 2021).

There is some scholarship that suggests that the EU should lookto enact a more open exception along the lines of the US fair use doctrine[57], which offers more flexibility in today's dynamic economic and technological environment, but the fair use doctrine itself is also not completely up to the mark as it leaves room for ambiguity and necessitates judicial interpretation on a case to case basis, thus beingdevoid of a sufficient degree of predictability. Although the fair use doctrine seems to provide a competitive advantage to AI developers in the USA vis-a-vis those in Europe, it falters theoretically when it comes to the question of AI training which involves expressive use of copyright protected works. An excessively liberal construction ofthe fir use doctrine risks jeopardizing the pecuniary interests of the right holders.

The most appropriate solution may be the adoption of a clear and unambiguous provision that welds together the merits of the twosystems, particularly for India, where the AI regulatory framework isstill taking shape. In other words, a limited TDM exception must be accompanied by a wider clause that exempts any use that may not beexpressly stated in the law, but is in the interest of the general public,especially with respect to freedom of expression and the right to information. More discussion on the issue of copyright infringementin AI training is expected from the World Intellectual Property Organization in the coming days. The WIPO must take a balanced approach on the matter and ensure global coordination as far as possible. It must clarify the path ahead, keeping in mind the utility of AI technology in dealing with an unprecedented global health andfinancial crisis. It should provide guidance to states on the possible mechanisms that can be adopted to govern the use of copyright worksfor AI training in a progressive manner that will stand the test of time.

[57] Christophe Geiger & Elena Izyumenko, "Towards a European "Fair Use" Grounded in Freedom of Expression", 35 AM. U. INT'l L. REV. 1 (2019*).*