DOI 10.5817/MUJLT2019-2-9

ALGORITHMIC COPYRIGHT ENFORCEMENT AND AI: ISSUES AND POTENTIAL SOLUTIONS, THROUGH THE LENS OF TEXT AND DATA MINING

by

ANDREA KATALIN TÓTH*

Although digitalization and the emergence of the Internet has caused a long-term crisis for copyright law, technology itself also seems to offer a seemingly ideal solution to the challenges of digital age: copyright has been a major use case for algorithmic enforcement from the early days of digital rights management technologies to the more advanced content recognition algorithms. These technologies identify and filter possibly infringing content automatically, effectively and often in a preventive fashion. These methods have been criticized for their shortcomings, such as the lack of transparency, bias and the possible impairment of fundamental rights. Self-learning machines and semi-autonomous AI have the potential to offer even more sophisticated and expeditious enforcement by code, however, they could also aggravate the aforementioned issues. As the EU legislator envisions to make the use of such technologies essentially obligatory for certain online content sharing service providers (via the infamous Article 17 of the directive on copyright in the digital single market), the assessment of the situation in light of future technological development has become a current topic.

This paper aims to identify the main issues and potential long-term consequences of creating legislation that practically requires the employment of such filtering algorithms as well as their solutions. This paper focuses

* andreakatalin.toth@gmail.com, legal officer at the Department of International Copyright Law, Hungarian Intellectual Property Office; Ph.D. candidate at the Department of Civil Law of Eötvös Loránd University Faculty of Law and Political Science in Budapest, Hungary.

on the potential role a broad copyright exception for text and data mining could play in counterbalancing the issues associated with algorithmic enforcement.

KEY WORDS

AI, Copyright Law, EU Law, Machine Learning, Technology, Text and Data Mining

1. INTRODUCTION: COPYRIGHT, EXCEPTIONS AND TECHNOLOGY

The purpose and aim of copyright law has traditionally been described along two major theoretical views: according to the utilitarian approach, copyright's goal is to promote the advancement of learning and culture by providing certain exclusive rights to authors and creators in order to stimulate the production and dissemination of intellectual works, while the natural rights-based justification argues that the relevant rights need to be afforded to authors and creators as a reward for their intellectual labour, as well as a protection of their personality enshrined in their works.[1] Even though the two main copyright law regimes, the common law based "copyright" system and the *droit d'auteur* (authors' rights) approach prevalent in continental Europe formulate and emphasize these ideas differently,[2] the underlying concept is similar in each jurisdiction. From an economic aspect, these exclusive rights (such as: right of reproduction, right of distribution, public performance, creation of derivative works) incentivize and reward the intellectual labour of copyright holders (who are usually the authors of the work), by giving them the sole authority to license and authorize the use and exploitation of their copyright--protected works to third parties.

However, this power does not create an absolute monopoly for the right holder: for the sake of long-term development, and in order to make

[1] Fisher, W. W. (2001) Theories of Intellectual Property. In: Stephen Munzer (ed.). *New Essays in the Legal and Political Theory of Property.* 1st ed. Cambridge: Cambridge University Press, pp. 169–171.

[2] Although the most obvious example of the embodiment of this idea is Article I. Section 8. Clause 8. of the United States' Constitution, it also appears in Recitals (2), (4) and (10) of the most important European copyright directive, the InfoSoc Directive (Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal of the European Union* (2001/L-167/10) 22 June. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32001L0029&from=EN [Accessed 10 January 2019]) as well as in the recital of the Hungarian Copyright Act, thus this concept is also deeply embedded in the continental "authors' rights" regimes.

the knowledge incorporated in copyright-protected works more easily accessible, some limitations on these exclusive rights have been put in place. One way to limit copyright is by introducing different exceptions[3] by declaring that certain specific uses that do not conflict with the normal exploitation of works and do not unreasonably prejudice the legitimate interests of the right holder[4] do not necessitate prior authorization and/or payment of royalties. These uses are excepted for different reasons, for instance due to their *de minimis* impact on right holders' rights (e.g. temporary acts of reproduction) or their socially beneficial nature (e.g. teaching illustration, criticism).[5] At the same time, however, exceptions also serve as an important tool for balancing between the legitimate economic interests of copyright holders and the fundamental rights (most importantly the freedom of expression and information) of users.

Another important feature of copyright law for the purposes of this paper is its connection to technology and the way the development of this specific field of law and the advancement of technology have always been closely intertwined: the appearance of the movable type and printing press and their contribution to the technology of dissemination of information proved to be a disruptive technology and resulted in the need for

[3] Even though there is no opportunity to explore the topic in detail in this paper, the distinction between the Anglo-American style of *fair use* or *fair dealing* system and the exhaustive list of exceptions found in continental European *droit d'auteur* regimes should be mentioned in relation to the subject of copyright limitations and exceptions. The former, more flexible scheme relies on the judicial interpretation of certain standards. Judges evaluate the following four factors in relation to the allegedly infringing use: (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes; (2) the nature of the copyrighted work; (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 (Sec. 107, Copyright Act of 1976). In contrast, the continental European system accommodates clearly and narrowly defined exceptions implemented by way of legislation (see also InfoSoc Directive, Article 5). For more on the American style fair use see: Leval, P. N. (1990) Toward a Fair Use Standard. *Harvard Law Review*, 103, p. 1105; Fisher, W. W. (1988) Reconstructing the Fair Use Doctrine. *Harvard Law Review*, 101 (8), p. 1659; Thatcher, S. G. (2006) Fair Use in Theory and Practice: Reflections on its History and the Google Case. *Journal of Scholarly Publishing*, 37 (3), pp. 215–229; Richard, K. (2018) Fair Use in the Information Age. *Richmond Journal of Law & Technology*, 25 (1); or the U.S. Copyright Office's information. United States Copyright Office. (2019) *More information on fair use.* [online] Washington, D. C.: USCO. Available from: https://www.copyright.gov/fair-use/more-info.html [Accessed 23 May 2019].

[4] This set of requirements is known as the "three step test" and it ensures that exceptions would not truncate copyright protection to an unjustified extent. The test first appeared in Article 9 of the *Berne Convention for Protection of Literary and Artistic Works* and the concept later became also enshrined in Article 13 of the *Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS)*, as well as Article 5 paragraph (5) of the InfoSoc Directive.

[5] Stamatoudi, I. and Torremans, P. (2014) *EU Copyright Law, a Commentary.* 1st ed. Cheltenham: Edward Elgar Publishing Limited, p. 441.

an exclusive right for publishers in order to secure their business. This later developed into an exclusive right for the authors of works,[6] and led to the appearance of copyright as a distinct field of law.[7] Throughout its history, technology and new technological inventions have had the most relevant impact on copyright's evolution: new inventions, such as the *Xerox* machine, the audio cassette or the VCR not only accommodated new forms of uses, but also upset the above-mentioned balance between the interests of right holders and users.[8] The most dramatic change and challenge for copyright law so far has proved to be digitalization and the emergence of the Internet. In this new, digital environment the costs of copying and sharing information and copyright-protected content converge towards zero, which fosters unauthorized mass production and distribution, and thus mass infringement.[9] As digital uses of copyright-protected works usually occur in a cross-border manner (given the globalized nature of the Internet) and under anonymity ensured by the World Wide Web, the proper enforcement of exclusive rights became exponentially more difficult for right holders. Many scholars, commentators, policymakers and legislators sought to find a solution to this "crisis" situation, by legislative or extra-legislative means, however, these efforts did not always bring the desired results.[10] Concerning law making, as the legislative process is and will always be slower than technological development, the application and interpretation of existing laws to new technologies and solutions constitutes another problem in the context of technological neutrality. Though this overarching principle of lawmaking aims to ensure that legal provisions are constructed in a way that is independent from any particular

---

[6] The first copyright act, the Statute of Anne was adopted in 1710 in Great Britain and it deviated from the earlier legislation that gave publishing monopoly to the Stationer's Company (an exclusive group of printers and booksellers) and it vested the rights and protection in the authors themselves. See: Joyce, C. (ed.). (2013) *Copyright Law.* 9th ed. New Providence: LexisNexis, pp. 17–19.

[7] Joyce, C. (ed.). (2013) Op. cit., p. 16.

[8] Latman, A. and Patry, W. F. (1986) *Latman's the Copyright Law.* 6th ed. Washington, D.C.: Bureau of National Affairs.

[9] Joyce, C. (ed.). (2013) Op. cit., pp. 45–47.

[10] For more on this, see: Mills, M. L. (1989) New Technology and the Limitations of Copyright Law: An Argument for Finding Alternatives to Copyright Legislation in an Era of Rapid Technological Change. *Chicago-Kent Law Review*, 65 (1); Geller, P. E. (2008) Beyond the Copyright Crisis: Principles for Change. *Journal of the Copyright Society of the USA*, 55, pp. 165–199; Litman, J. (2002) Revising Copyright Law for the Information Age. In: Adam Thierer and Wayne Crews (eds.). *Copy Fights: The Future of Intellectual Property in the Information Age.* 1st ed. Washington, D. C.: Cato Institute.

technology without any negative or positive discrimination,[11] the different approaches towards its conceptualization can lead to very different results.[12] Thus, even the more flexible and reactive jurisprudence and case law is unable to guarantee an adequate, appropriate and uniform answer to the questions of copyright law brought about by emerging new technologies.

2. ALGORITHMIC COPYRIGHT ENFORCEMENT AND ITS EVOLUTION

The so-called algorithmic enforcement of copyright appeared in light of the aforementioned problem triggered by digitalization and the spread of the Internet. As it became clear that the traditional ways of enforcement were inefficient and costly (individual users behind online infringements became extremely difficult to track down and identify and they are typically judgement-proof against large sums of damages), the idea of using technology itself to solve the issues brought about by technology appeared,[13] and the concept of controlling digital uses by digital means came to light.

In copyright, the first generation of algorithmic enforcement tools comprised of the so-called technological protection measures (TPM) [also known as digital rights management, or DRM technologies in the United States], which operated as digital locks: right holders could technically prevent unauthorized access to and control the subsequent use of the digital formats of their works, by way of encryption.[14] This provided a well--functioning technology for right holders, and ensured that users could only gain access to legally acquired works; the option to make digital copies was either completely disabled or limited to a small number of copies or even

[11] Greenberg, B. A. (2016) Rethinking Technology Neutrality. *Minnesota Law Review*, 100 (4), p. 1513.

[12] A more restrictive understanding of technological neutrality could result in the rigid application of old law to new technology regardless of its potential impact on the development of said technology, while the more laxed views also consider achieving equivalent outcomes and maintaining the purpose of copyright law itself. This can lead to opposing results when assessing whether an act is copyright-relevant or not. See: Craig, C. J. (2017) Technological Neutrality: Recalibrating Copyright in the Information Age. *Theoretical Issues in Law*, 17 (2), pp. 608–615.

[13] About the idea that "code is law" and the role of technology as a means for indirect regulation, see: Lessig, L. (2006) *Code v. 2.0.* [online] New York: Basic Books. Available from: http://codev2.cc/download+remix/Lessig-Codev2.pdf [Accessed 10 January 2019].

[14] Perel, M. and Elkin-Koren, N. (2016) Accountability in Algorithmic Copyright Enforcement. *Stanford Technology Law Review*, 19 (3), p. 484.

a restriction regarding the type and number of the devices used for the enjoyment of the works could be applied.[15] The most known applications of this technology were *CSS (Content Scrambling System)*, *Apple's Fair Play* or *Adobe's DRM*. These technologies suffered from a number of shortcomings: as they were easily hacked, an additional legal protection (in the form of the prohibition of the circumvention of TPM) was needed. In addition, even the introduction of such provisions could not help to remedy other problems, such as TPMs causing security risks and slowing down computers, limiting consumers' ability to enjoy their legally bought products by only letting them to access their content on a limited number of devices or generally overriding copyright exceptions by being overly preventive by design.[16] Although some technological tools to accommodate exceptions existed at the time (such as interoperability, the partitioning and authentication of users), they were not and could not be employed by the majority of TPM technologies.[17]

With the spread of social media and the emergence of platforms such as *Facebook*, *YouTube*, or *Instagram* as well as the proliferation of user--generated content that these new platforms enabled, the second generation of algorithmic enforcement technologies appeared. The main focus of these new tools became the online availability of copyright protected content.[18] *Facebook's Rights Manager*[19] or *YouTube's Content ID*[20] offer right holders a nuanced approach to digital copyright management. The best way to illustrate the functioning of such systems is through the example of *YouTube's ContentID* algorithm. Through this mechanism, right holders provide to *YouTube* information and data about their works that they do not wish to see unauthorized copies of on the video-sharing platform. Based on these data a digital fingerprint for that specific piece of content is generated. Each time a new video is uploaded to *YouTube*, the algorithm

[15] Kerr, I. (2010) Digital Locks and the Automation Virtue. In: Michael Geist (ed.). *From „Radical Extremism" to „Balanced Copyright": Canadian Copyright and the Digital Agenda.* 1st ed. Toronto: Irwin Law, p. 267.

[16] Myška, M. (2009) The True Story of DRM. *Masaryk University Journal of Law and Technology*, 3 (2), pp. 272–277.

[17] Akester, P. (2009) *Technological Accomodation of Conflicts between Freedom of Expression and DRM: The First Empirical Assessment.* Rochester, New York: Social Science Research Network, p. 103.

[18] Perel, M. and Elkin-Koren, N. (2016) Op. cit., pp. 478–481.

[19] Facebook. (2019) *Rights Manager.* [online] Available from: https://rightsmanager.fb.com/ [Accessed 10 January 2019].

[20] YouTube. (2019) *Copyright Management Tools – Content ID.* [online] Available from: https://support.google.com/youtube/answer/9245819 [Accessed 10 January 2019].

checks whether there are any matches between any of the fingerprints in the library and the video in question. In the event of a newly uploaded video matching a fingerprint, it becomes flagged as potentially infringing content. As a consequence, the right holder has a few options to choose from: they can follow the viewership statistics of the flagged video, block access to it, or they can also claim all advertising revenues in case the allegedly infringing video is monetized.[21] According to *YouTube's* statistics, *ContentID* is used by more than 9,000 partners, including television broadcast companies, movie studios as well as record companies, while the reference library contains more than 75 million digital fingerprints.[22] Nevertheless, it also means that the main beneficiaries of the *ContentID* mechanism are high-profile entertainment companies whose protected works are used in large numbers. As the employment of this technology necessitates the ownership of a significant amount of copyright-protected content, the submission of a high number of valid takedown requests and the resources to manage them, *ContentID* and its options mentioned above are mostly available for large and economically significant right holders.[23] Smaller companies owning copyright-protected content can benefit from the *Content Verification Tool*, which only makes it possible for the right holders to search for and request the removal of potentially infringing videos.[24] Creators of smaller scale (typically the authors of user-generated content) are offered the *Copyright Match Tool*, which scans the platform for unauthorized uploads of original videos. However, in case of matching content, the authors are only offered more limited options: they can email the uploader, request the immediate removal of the matched content, request a scheduled removal or archive the match without taking any action.[25] Thus, it is clear that the biggest actors in the industry dispose of the widest array of possibilities and most effective tools for enforcement, while smaller entities and creators of original content (who constitute the basis of *YouTube's* functioning and philosophy) have

[21] YouTube. (2019) *How Content ID works.* [online] Available from: https://support.google.com/youtube/answer/2797370?hl=en&ref_topic=2778544 [Accessed 10 January 2019].

[22] YouTube. (2019) *YouTube in Numbers.* [online] Available from: https://www.youtube.com/yt/about/press/ [Accessed 14 June 2019].

[23] YouTube. (2019) *Copyright Management Tools.* [online] Available from: https://support.google.com/youtube/answer/9245819?hl=en [Accessed 14 June 2019].

[24] YouTube. (2019) *Content Verification Program.* [online] Available from: https://support.google.com/youtube/answer/6005923 [Accessed 14 June 2019].

[25] YouTube. (2019) *Copyright Match Tool.* [online] Available from: https://support.google.com/youtube/answer/7648743 [Accessed 14 June 2019].

more constrained options to enforce their rights. The most striking difference is the lack of the option for monetization, the potential to claim the advertising revenues off the potentially infringing videos.

Even more so that this latter option is what provides the apparent benefit of the second generation systems: contrary to the first generation of enforcement technologies, they enable an *ex post facto* licensing mechanism through the possibility of claiming ad-revenues.[26] However, this solution is not completely in line with copyright law's concept: no prior authorization is granted as the collection of revenues takes place after the actual use has already happened; there is no direct agreement between the right holder and the user, thus there is no enforceable contract in place for the purpose of using the protected work. The punitive nature of the redirecting of revenues is also foreign in the licensing practice. At the same time, the content of the videos at least remain accessible to the public. This scheme accommodates freedom of expression and information better, as the default option is not to completely block the potentially infringing content, but to keep it accessible in order to generate revenue for the right holder. At first glance, this mechanism seems to offer a near to ideal solution to the digital copyright law crisis: videos can still be watched by the passive, consumer public, while right holders receive income after the use of their works. Nevertheless, the uncertainty about the type of content that can actually trigger the algorithm and would be flagged and qualified as infringing carries the potential to create a discouraging environment for active users (especially those producing user-generated content), resulting in self--censorship.

3. THE POTENTIAL ISSUES OF ALGORITHMIC COPYRIGHT ENFORCEMENT

Even though the technologies introduced in the previous chapter cater for an effective and seemingly well-functioning enforcement of digital copyright, the potential drawbacks of and issues caused by these algorithmic measures need to be considered and evaluated as well.

One of the main problems derives from the fact that codes and algorithms used as the basis of these technologies are mostly treated

[26] Perel, M. and Elkin-Koren, N. (2016) Op. cit., p. 512–513.

as trade secrets and as such are kept hidden from the public eye in order to secure competitive advantage as well as to prevent users from "playing the system" by exploiting loopholes in the functioning of the algorithms. The resulting non-transparency can lead to overprotection and abuse of power through a lack of accountability.[27] As a consequence, individuals with the intent to legitimately use these platforms are unable to adjust their behaviour to be compliant due to their unawareness of the boundaries of the rules implied by technology. The uncertainty about the type of content that can actually trigger the algorithm and would be flagged and qualified as infringing carries the potential to create a discouraging environment for active users, especially those producing user-generated content and resulting in self-censorship. Given that social media and content sharing platforms were specifically built on the idea of users creating and sharing their own original content, this issue goes to the core of the functioning of these service providers.

The second identified issue is that right holders can effectively disable copyright exceptions by exercising excessively strict control over their content. The problem with the current content identification technologies (including *YouTube's Content ID*) is that although they are capable of filtering out identical or matching content, they are not sophisticated enough to be able to distinguish infringing use from uses that fall under one of the categories of exceptions.[28] Thus, even excepted uses could be flagged and blocked from public availability. An illustrative example is of a review video about a newly released movie: in order to get the point across and to give a foundation to their arguments, the reviewer has the option to use some footage from the movie, which (also considering the extent of the use) could easily qualify as a copyright exception as comment or criticism.[29] Whether inside or outside of the realm of copyright exceptions, disproportionality may present another issue. The terms of the after-the-fact quasi licence contract (which essentially bears the characteristics of a "compulsory licence") embodied in the demonetization and ad-revenue claims could be highly unfair and disproportionate to the actual use of the protected content.[30] For instance, the use of a few seconds of a song as background music in a vlog or a gaming stream could essentially

[27] Op. cit., p. 483.

[28] Bartholomew, T. B. (2015) The Death of Fair Use in Cyberspace: YouTube and the Problem with Content ID. *Duke Law & Technology Review*, 13 (1), p. 70.

"hijack" the advertising revenue of videos of substantial length and views.[31] Regarding the incidental inclusion exception in EU copyright law and other jurisdictions where *de minimis* use falls outside of the scope of copyright protection, this issue relates back to the limitations of copyright.[32]

Finally, whenever legal provisions are translated into code, private and potentially biased actors analyse and interpret the law. As these entities determine the metes and bounds of specific rules, they have a substantial potential in building bias into the code that would favour their interests and discriminate against certain other individuals or groups.[33] The most possible form of bias in the context of enforcement algorithms is technical bias that originates from trying to make human constructs, such as a judgement on the substance of a legal provision, interpretable for computers.[34] Given that the interpretation of law is traditionally a public function

---

[29] In EU copyright law, Article 5, para. (3) d) of the InfoSoc Directive states that Member States may provide for exceptions or limitations to the rights of reproduction and communication to the public in the case of quotations for purposes of such as criticism or review, provided that they relate to a work or other subject matter which has already been lawfully made available to the public, the use is in accordance with fair practice, and to the extent required for the specific purpose. Similarly, Section 107 of the US Copyright Act (17 U.S. Code) states that criticism and comment are of the specific purposes that might warrant fair use in light of the evaluation of the four factors. For a specific example, a movie review about an infamously "bad" movie was given a copyright strike and blocked by the movie's director three days after its release. For the original review video see: I Hate Everything. (2015) *Cool Cat Saves The Kids – The Search For The Worst.* [online video] Available from: https://www.youtube.com/watch?v=HoTZZYm2HZI&t=42s [Accessed 10 January 2019]; for a comment on the video's removal and fair use, see e.g.: Channel Awesome. (2016) *Where's The Fair Use – Nostalgia Critic.* [online video] Available from: https://www.youtube.com/watch?v=zVqFAMOtwaI&t=53s [Accessed 10 January 2019].

[30] Bartholomew, T. B. (2015) Op. cit., p. 66.

[31] One of the most popular YouTubers with a significant number of subscribers, *Felix Kjellberg* (a.k.a. *PewDiePie*) often complains about record labels and production companies claiming the advertising revenue of his gameplay videos (the length of which can extend up to a few hours) for the use of a few seconds of a copyright protected song (that sometimes appear as part of the video-game itself). See for example: PewDiePie. (2017) *Life is cringe – life is strange – S2E01*. [online video] Available from: https://www.youtube.com/watch?v=PX4zk0G4IjM [Accessed 10 January 2019].

[32] Article 5 (3) (i) states that Member States may provide for exceptions or limitations to the rights of reproduction and communication to the public for the incidental inclusion of a work or other subject-matter in other material. This provision creates a legal basis for the introduction of *de minimis* limitations in EU countries' national laws. In the USA, trivial, or *de minimis* use is often allowed by courts. It means that the unauthorized use in question is so small and irrelevant that it would weigh against the finding of infringement both regarding the substantiality of the portion taken and the possible effect of the use on the potential market of the protected work (the third and fourth factors described above in footnote 2.). This doctrine has been developed by case law, mostly in relation to background objects appearing in movies. See: *Ringgold v. Black Entertainment Television, Inc.* (1997) 126 F.3d70, 16 September; *Sandoval v. New Line Cinema Corp.* (1998) 147 F.3d 215, 24 June; *Newton v. Diamond* (2004) 388 F.3d 1189, 7 April.

[33] Friedman, B. and Nissenbaum, H. (1996) Bias in Computer Systems. *ACM Transactions on Information Systems*, 14 (3), pp. 332–333.

[34] Op. cit., p. 334.

of the judiciary or of the legislator, in instances when it is outsourced to private companies, the public scrutiny that courts, judges and parliaments are otherwise subject to can be easily evaded by these entities.[35]

4. A NEW GENERATION IN ALGORITHMIC ENFORCEMENT?

As artificial intelligence and machine learning[36] is gradually spreading across the world, algorithmic copyright enforcement seems to be an obvious field of application. One of the essential tools and technological manifestations of machine learning is text and data mining, which covers the process of gathering and analysing vast amounts of information in order to be able to forecast certain trends and patterns.[37] For autonomous and semi-autonomous systems, the supply of infinite amount of user-generated content[38] provides an invaluable pool of diverse and unfiltered training data, which ensures their effective and accurate functioning. Text and data mining is generally used to extract and classify data from large sets of information. Based on the *KDD-process*[39] *(Knowledge Discovery in Databases)*, it includes the selection, pre-processing, transformation, the actual mining and finally, the evaluation or interpretation of data. Machine learning algorithms, on the other hand, use these clean and targeted datasets and the trends and patterns drawn from them as training data to learn to predict future occurrences as well as to carry out certain tasks in a supervised or unsupervised fashion.[40] As these algorithms generally work better and produce the most accurate results if they have

[35] Citron, D. K. (2008) Technological Due Process. *Washington University Law Review*, 85 (6), p. 1298.

[36] Although these two terms are used interchangeably in the context of this article, machine learning and artificial intelligence are not exactly the same. Artificial intelligence is the broader concept, while machine learning is the manifestation of the study and learning processes that could be applied in artificial intelligence solutions. See: Ryszard S. Michalski, Jaime G. Carbonell and Tom M. Mitchell (eds.). (1983) *Machine Learning: An Artificial Intelligence Approach.* 1st ed. Berlin: Springer-Verlag, p. 3.

[37] Witten, I. H. and Frank, E. (2005) *Data Mining, Practical Machine Learning Tools and Techniques.* 2nd ed. San Francisco: Morgan Kaufmann Publishers, p. 23.

[38] According to some sources, there are 400 hours worth of videos uploaded to *YouTube* every minute and approximately 95 million pictures shared on *Instagram* daily. See at: DMR. (2019) *160 YouTube Statistics and Facts.* [online] Available from: https://expandedramblings.com/index.php/youtube-statistics/ [Accessed 11 January 2019] and Omnicore. (2019) *Instagram by the Numbers: Stats, Demographics & Fun Facts.* [online] Available from: https://www.omnicoreagency.com/instagram-statistics/ [Accessed 11 January 2019].

[39] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, 39 (11), pp. 30–31.

[40] Murphy, K. P. (2012) *Machine Learning: A Probabilistic Perspective.* 1st ed. Cambridge: Massachusetts Institute of Technology, p. 2.

as much and as diverse data as possible at their disposal,[41] the content managed by these platforms seems ideal for the implementation of machine learning technologies, especially in the field of enforcement.

Considering the issues of algorithmic enforcement discussed above, AI's and machine learning's main contribution towards algorithmic copyright enforcement could be their potential to spot and differentiate clearly infringing use from fair use with the help of their more sophisticated technology than those of TPM and the hashing and search algorithms that are currently employed.[42] Even more so considering that, based on *YouTube's* statement, their content recognition tools do not determine copyright exceptions or fair use.[43] However, in order to make these algorithmic systems more balanced in their functioning, the checks and limitations of the exclusive rights embodied in the exceptions and fair use should be part of their design.[44] Through an adequate flagging and training system, in which the initial enhanced human supervision embodied in marking and flagging infringing and non-infringing content could be later substituted by the algorithm's own assessment facilitated by high--quality and streamlined datasets,[45] the algorithm could be taught to identify cases of fair use or instances of copyright exceptions. Even though the different legal systems and jurisdictions regulate copyright exceptions differently,[46] the problem translated into code is rather uniform. For instance, there are several exceptions that necessitate the evaluation of the creator's intent and purpose as well as the context of the utterance: the relevant question is whether the work was used in relation to social commentary, a parody, teaching illustration or for quotation. AI is already getting better at understanding the intent of the writer or speaker and the context of the specific text through natural language processing.[47] Additionally, it is known that *YouTube* actually uses machine learning

[41] See e.g. Halevy, A., Norvig, P. and Pereira, F. (2009) The Unreasonable Effectiveness of Data. *Intelligent Systems, IEEE*, 24 (2); Banko, M. and Brill, E. (2001) Scaling to Very Very Large Corpora for Natural Language Disambiguation. In: Bonnie Lynn Webber (ed.). *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, 6–11 July. USA: Association for Computational Linguistics, pp. 26–33.

[42] Elkin-Koren, N. (2017) Fair Use by Design. *UCLA Law Review*, 64 (5), p. 1097.

[43] See: Google. (2019) *Frequently asked questions about fair use.* [online] Available from: https://support.google.com/youtube/answer/6396261?hl=en [Accessed 15 June 2019].

[44] Elkin-Koren, N. (2017) Op. cit., p. 1085.

[45] Lester, T. and Pachamanova, D. (2017) The Dilemma of False Positives: Making Content ID Algorithms more Conducive to Fostering Innovative Fair Use in Music Creation. *UCLA Entertainment Law Review*, 24 (1), p. 69.

[46] See footnote 3.

in order to distinguish and eliminate extremist content from its platform, and, according to the company, the algorithm seems to function quite well.[48, 49] Based on these assertions, it is not irrational to imagine that the different AI and machine learning applications could be combined together to deal with more complex expressions and issues, such as audio--visual content and copyright exceptions.

Nevertheless, even though the issue relating to fair use and exceptions could be potentially addressed by AI, the other problems already mentioned in relation to algorithmic copyright enforcement have the ability to be magnified through the employment of these novel technologies. Transparency of the decision-making process and the arguments behind its reasoning would essentially disappear: some forms of autonomous systems generate their own code, while deep learning applications and neural networks function effectively as "black boxes" due to their immense complexity, the lack of human intervention as well as the inability to reverse engineer the processes and the reasons behind the machine's actions.[50] As learning algorithms do not only implement the goals of the creator of the code but have the capacity to modify the meaning of the goals

[47] There has been recent developments both regarding sentiment analysis and sarcasm detection through deep learning. See: Sarikaya, R., Geoffrey E. and Deoras, A. (2014) Application of Deep Belief Networks for Natural Language Understanding. *IEEE Transactions on Audio, Speech and Language Processing*, 22 (4) and Zhang, M., Zhang, Y. and Fu, G. (2016) Tweet Sarcasm Detection Using Deep Neural Network. In: Eiichiro Sumita, Takenobu Tokunaga and Sadao Kurohashi (eds.). *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, 11–16 December. Japan: Japanese Association of Natural Language Processing, pp. 2457–2458.

[48] YouTube. (2017) *An update on our commitment to fight violent extremist content online.* [online] Available from: https://youtube.googleblog.com/2017/10/an-update-on-our-commitment-to-fight.html [Accessed 13 January 2019]. Based on *Google's* recent transparency report, almost 90,000 videos were removed between January and March 2019 due to being of violently extremist nature: Google. (2019) *Featured policies.* [online] Available from: https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism [Accessed 14 June 2019].

[49] Although *YouTube* claims that automatization is key in removing content before it could go viral, a *Counter Extremism Project's* report on ISIS content on *YouTube* found that 24 % of the examined 1,348 videos remained online for more than two hours, garnering close to 150,000 views, while 91 % of the extremist videos were later reuploaded. These data are however not completely indicative of the effectiveness of the machine learning algorithms, given that *YouTube* employs human review and hashing as well, while automatization is mainly used to locate extremist videos. Counter Extremism Project. (2018) *The eGlyph Web Crawler: ISIS Content on YouTube.* [online] Available from: https://www.counterextremism.com/sites/default/files/eGLYPH_web_crawler_white_paper_July_2018.pdf [Accessed 14 June 2019].

[50] For further information on this issue, see: Knight, W. (2017) The Dark Secret at the Heart of AI. *MIT Technology Review*, 11 April. [online] Available from: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [Accessed 13 January 2019].

themselves,[51] it would be close to impossible to tell if the machine made justified decisions and used the right criteria for assessing fair use. Similarly, accountability could present a new challenge, as the question of how AI could explain its decisions also touches on the issue of legal personality of artificial intelligence and how and to whom liability for damages and wrongdoings could and should be assessed.[52] Finally, the algorithm-driven pre-adjudication process could lead to biased decision making: even though the formal and public court proceedings would still be available for aggrieved parties, the trust put in algorithmic enforcement and automation bias[53] would discourage people from turning to the traditional judiciary when they feel that their rights as users have been violated by the application of automated enforcement measures, due to humans' tendency to ignore or not search for contradictory information, if a decision is generated by a sophisticated computer and believed to be correct.[54] This could affirm that any sort of bias embedded in the process would remain in the system, unchallenged.

5. THE DIRECTIVE ON COPYRIGHT IN THE DIGITAL SINGLE MARKET AND ITS ARTICLE 17

These concerns as well as the whole idea of automated algorithmic copyright enforcement have become even more relevant recently in Europe, in the context of the EU's recent copyright reform.

The most important part of the copyright reform package of 2016, the directive on copyright in the digital single market[55] (DSM Directive) envisions to modernize European copyright rules to meet the challenges

---

[51] Perel, M. and Elkin-Koren, N. (2017) Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement. *Florida Law Review*, 69 (1), p. 189.

[52] For the extensive literature on the issue of legal personality implications of artificial intelligence see: Solum, L. B. (1991) Legal Personhood for Artificial Intelligences. *North Carolina Law Review*, 70 (4); Čerka, P., Grigienė, J. and Sirbikytė, G. (2017) Is it possible to grant legal personality to artificial intelligence systems? *Computer Law & Security Review*, 33 (5); Allgrove, B. (2004) *Legal Personality for Artificial Intellects: Pragmatic Solution or Science Fiction?* [online] Available from: https://ssrn.com/abstract=926015 [Accessed 15 January 2019].

[53] Bamberger, K. A. (2010) Technologies of Compliance: Risk and Regulation in a Digital Age. *Texas Law Review*, 88 (4), p. 676.

[54] Cummings, M. L. (2006) Automation and Accountability in Decision Support System Interface Design. *The Journal of Technology Studies*, 32, p. 25.

[55] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. *Official Journal of the European Union* (2019/L-130/92) 17 May. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2019.130.01.0092.01.ENG&toc=OJ:L:2019:130:TOC [Accessed 15 June 2019].

of the digital world as well as to ensure the proper functioning of the internal market by stimulating innovation, creativity and investment in new content.[56] One of the most debated and controversial provisions, Article 17 aims to regulate the status and liability of certain online platforms. The provision's goal is to clarify and uniformize the *Court of Justice of the European Union (CJEU)* case law and declare online content sharing service providers that store and handle a significant amount of copyright protected work to be primary users of the content when they give the public access to these works or other protected subject matter uploaded by their actual end users.[57] This rule would mainly concern social media and content sharing sites, such as *YouTube*, *Facebook* or *Instagram*, while not-for-profit encyclopaedias, cloud services, educational and scientific repositories, open-source software developing platforms and online marketplaces fall outside the scope of the definition of *"online content sharing service provider"*. As primary users of copyright protected works, it will be necessary for these platforms to obtain licenses, pay licensing fees and to bear the burden of primary liability for copyright infringement. If no such license or authorization is granted, then platforms will be liable for the unauthorized acts of communication to the public, including making available to the public, of the copyright-protected works, unless they demonstrate that they made their best efforts to obtain an authorization and to ensure the unavailability of specific works (for which the right holder has provided the necessary information), and in any case, acted expeditiously upon the receipt of a notice to block or remove those specific works.[58] Nevertheless, the measures to comply with this obligations need to be proportionate to the type, audience, size of the service and the type of the works uploaded, as well as the availability of suitable and effective means.[59] If there is an authorization acquired, it will also have to cover the acts of the users, when they are not acting on a commercial basis or their activities do not generate a significant amount of revenues.[60] Regarding the tools to ensure the unavailability of unlicensed material, the earlier versions of the proposal even made an explicit reference to content

[56] Op. cit., Recital (2).
[57] Op. cit., Article 17 paragraph (1).
[58] Op. cit., Article 17 paragraph (4).
[59] Op. cit., Article 17 paragraph (5).
[60] Op. cit., Article 17 paragraph (2).

recognition technologies.[61] The *European Parliament's* approved report that constituted a basis for the informal trilogue negotiations was even more rigorous in this regard, as it did not even provide for an exemption as described above, thus placing the burden of strict liability for copyright infringement on the platforms concerned.[62]

Even though such measures are currently used by some online platforms voluntarily (as we have seen earlier through the example of *YouTube*), these sites could have still qualified as intermediaries in most cases based on the Ecommerce Directive. As such, they could also have benefited from the harmonized safe harbour provisions[63] shielding them from secondary liability.[64] However, if these platforms are to be considered primary users (meaning that they are going to be regarded as performing the copyright--relevant act of communication to the public themselves as well when their end-users upload a piece of content), the utilisation of content recognition technologies would essentially become obligatory for them to avoid liability for infringement. This creates a strong incentive for these platforms to over filter and block any suspicious and possibly infringing content, in the absence of a relevant authorization. In order to achieve the best results, platforms would also be interested in using the state of the art technology for the application of these preventive measures, which points in the direction of the employment of machine learning and artificial intelligence-based technologies.

Nevertheless, in case these technologies are going to be a ubiquitous part of online content creation and consumption, potential solutions and ways

---

[61] Article 13 paragraph (1), European Commission. (2016) *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market.* (COM(2016) 593 final) 14 September. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016PC0593 [Accessed 15 January 2019].

[62] The so-called "Voss-report" (named after the rapporteur) was supported by a significant majority of MEPs at the plenary session of the European Parliament on 12 September. See: European Parliament. (2018) *Report on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market.* (COM(2016)0593 – C8-0383(2016) – 2016/0280(COD)). Available from: http://www.europarl.europa.eu/sides/getDoc.dopubRef=%2F%2FEP%2F%2FTEXT%2BREPORT%2BA8-2018-0245%2B0%2BDOC%2BXM%2BV0%2F%2FEN&language=EN [Accessed 7 February 2019].

[63] Currently secondary liability and its requirements are not harmonized on the European level, however, Articles 12–14. of the Ecommerce Directive do provide for a harmonized liability exemption scheme. See: Nordemann, J. B. (2017) Liability of Online Service Providers for Copyrighted Content – Regulatory Action Needed? In-Depth Analysis for the IMCO Committee. *Directorate-General for Internal Policies, Policy Department A (Economic and Scientific Policy), European Parliament*, p. 19.

[64] The proposal even makes an explicit reference to the inapplicability of the Ecommerce Directive's safe harbour rules to online content sharing platforms that perform a communication to the public. European Commission. (2016) Op. cit., Article 13 para. (3).

to mitigate the drawbacks of the currently employed enforcement systems and the future issues of machine learning-based technology outlined in the previous sections need to be considered.

6. TEXT AND DATA MINING AND ITS POTENTIAL IMPACT ON ALGORITHMIC ENFORCEMENT

Although there is a number of potential tools for such harm-reduction (such as the setting of certain standards of disclosure to ensure transparency and accountability[65] or an effective complaint and redress mechanism to tackle the problems of biased pre-adjudication and to provide human oversight), this chapter focuses on another provision within the DSM Directive, the exception on text and data mining and how it could alleviate the issues associated with algorithmic enforcement.

The essence of text and data mining can be captured through its definition, which denotes the extraction of implicit, previously unknown, and potentially useful information from data, for which machine learning provides the technical basis.[66] A study commissioned by the *European Commission* put text and data mining in the wider context of data analysis, which is the automated processing of digital materials, which may include texts, data, sounds, images or other elements, or a combination of these, in order to uncover new knowledge or insights.[67] Text and data mining is essential in realizing the full potential offered by the accumulation of huge amounts of data, and it is utilized in many different fields, such as commerce, finance, or marketing.[68] Additionally, text and data mining is becoming a useful tool in scientific and academic research[69] and based on the potential uses of machine learning outlined in the previous chapter, it could play an important role in the development of more sophisticated enforcement algorithms, which could differentiate between infringing and

---

[65] Lester, T. and Pachamanova, D. (2017) Op. cit., p. 70; Perel, M. and Elkin-Koren, N. (2016) Op. cit., pp. 529–530.

[66] Witten, I. H. and Frank, E. (2005) Op. cit., p. xxiii.

[67] Triaille, J. P., de Meeus d'Argenteuil, J. and de Francquen, A. (2014) *Study on the legal framework of text and data mining (TDM).* [online] Luxembourg: European Union. p. 17. Available from: https://publications.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en [Accessed 4 February 2019].

[68] Big Data Made Simple. (2014) *Top 14 useful applications of data mining.* [online] 20 August. Available from: https://bigdata-madesimple.com/14-useful-applications-of-data-mining/ [Accessed 4 February 2019].

[69] Filippov, S. (2014) Mapping Text and Data Mining in Academic and Research Communities in Europe. *The Lisbon Council Special Briefing Issue*, (16), p. 11.

non-infringing uses to a better extent. Text and data mining is closely related to machine learning, as, in general, knowledge extracted from examples of a task through data mining can allow for a better performance of the task, while learning process itself generates more knowledge in the form of data.[70]

As it has been stated earlier, one way to make these algorithms effective is to provide them with as much and as diverse information as possible. However, copyright law itself can constitute an obstacle in this process. The process of text and data mining includes the following stages: the business understanding of the problem, the data-specific understanding of the same problem and task, the preparation of data for analysis (the selection of relevant data and the creation of the final dataset), the modelling (the actual mining, which includes the choice of the proper method and its implementation), the evaluation of the prepared models, and finally, the application of the findings.[71] Text and data mining performed for machine learning purposes thus could potentially include copyright-relevant acts of copying, transforming, or communicating to the public while carrying out the steps above. This means that the analysis of data found within material that is protected by copyright or another right (such as the database right[72]) could necessitate the prior authorization of and additional payment to the right holders. This could be especially true in the case of platforms like *YouTube*, where the vast majority of videos are under copyright law's protection.

Large platforms, such as *YouTube* or *Facebook* operate with terms of service that already provide them with authorization to perform text and data mining on copyright-protected contents uploaded to their servers

---

[70] Calders, T. and Custers, B (2013) What Is Data Mining and How Does It Work? In: Bart Custers et al. (eds.). *Discrinimation and Privacy in the Information Society*. 1st ed. Berlin: Springer, p. 29.

[71] Based on the Cross-industry Standard Process for Data Mining (CRISP-DM). Vorhies, W. (2016) *CRISP-DM – a Standard Methodology to Ensure a Good Outcome.* [online] Data Science Central. Available from: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome [Accessed 4 February 2019].

[72] The database right is enshrined in Directive 96/9/EC of the European parliament and of the Council of 11 March 1996 on the legal protection of databases. It differentiates between databases protected by copyright law as the own intellectual creation of the author by reason of the selection or arrangement of their content and databases that merit protection due to the fact that the maker of the database has made a qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents of the database. This latter is called the *sui generis* database right. Makers of *sui generis* databases have the right to prevent the extraction and/or re-utilization of the whole or of a substantial part of the database.

through the all-encompassing global licenses.[73] Other actors, such as research or non-profit organizations, however, do not have this luxury of access and they lack one of the most important means to develop their own versions of ID algorithms: a minable database of considerable size.[74] This situation gives large tech-corporations a competitive advantage and essentially a monopoly on algorithmic copyright enforcement. We have seen earlier the problems that algorithmic pre-adjudication such as content filtering by private can pose, as these entities often have their own interests and agenda, which might be contrary to the interests of users as well as the freedom of expression.

A potential way to attenuate these consequences could be to remove the obstacle that copyright and other rights constitute. Although certain countries[75] already have provisions on a copyright exception for text and data mining already in force, there has been no such exception on the EU--level yet. However, a provision in the DSM Directive envisions to remedy this defect: among the rules on new, mandatory exceptions, Article 3 makes it compulsory for member states to introduce a copyright exception providing cultural heritage and research institutions the ability to freely use protected works for text and data mining for scientific research purposes. Another, originally optional provision that turned into a mandatory exception through the course of the negotiations (Article 4) additionally prescribes to member states to introduce a general and broad TDM--exception which would apply regardless of the nature of the beneficiary institutions or the purpose of the activity. This exception would provide an opportunity for other entities to more easily develop alternative methods and algorithms, as they would be free from the burden of authorization and remuneration-payment. In both cases, text and data mining could be carried out freely on works and databases to which they have lawful access to.

[73] See: YouTube. (2019) *Terms of Service, Section 8: Rights you license.* [online] Available from: https://www.youtube.com/static?template=terms [Accessed 7 February 2019] and Facebook. (2019) *Terms of Service, Section 3.3: The permissions you give us.* [online] Available from: https://www.facebook.com/terms.php [Accessed 7 February 2019].

[74] Although there are a number of public datasets that could be used freely for machine learning purposes, they usually do not contain information related to the consumption of copyright-protected content or copyright exceptions. For some lists of datasets see: Stanford, S. (2018) *The Best Public Datasets for Machine Learning and Data Science.* [online] Available from: https://medium.com/towards-artificial-intelligence/the-50-best-public-datasets-for-machine-learning-d80e9f030279 [Accessed 15 June 2019] or DeGroat, T. J. (2018) *19 Free Public Data Sets for Your Data Science Project.* [online] Available from: https://www.springboard.com/blog/free-public-data-sets-data-science-project/ [Accessed 15 June 2019].

[75] These countries include the UK, Ireland, Germany and Japan. See: Triaille, J. P., de Meeus d'Argenteuil, J. and de Francquen, A. (2014) Op. cit.

As a further limitation, right holders could expressly reserve the use of their works and protected subject-matter, thus retaining control over excluding TDM.

In any case, the exception on text and data mining could create competition that could cater for a more fair and transparent algorithmic enforcement. The possibility to be able to analyse and train semi--autonomous and autonomous systems is essential for the effective development of copyright enforcement algorithms. By ensuring that more and better data could be freely processed, the environment would be more adequate for the development of fair algorithms. If more non-profit and research organizations could create their own enforcement algorithms, it would not only ensure a more balanced competition through the possibility of choice for emerging platforms, but the aforementioned issues, such as transparency and bias could also be mitigated: if there are more actors, especially not-for-profit organizations, then trade secrecy becomes less of an issue and with a higher level of transparency the possibility of clandestine bias could be prevented as well. Nevertheless, the exception only concerns the actual acts of text and data mining, while the development of new algorithms is outside of its scope. However, the potential to license or sell the enforcement algorithms that have been based on the results of TDM carried out under the exception could either compel larger tech companies to take the development of their own content recognition tools seriously, or could create an alternative market and an incentive to outsource the creation of such algorithms to other entities.

7. CONCLUSION

Copyright law has gone through a number of significant changes in the past years, as it continuously struggled to keep abreast of technological development and to maintain its original goal as well as the level of protection to right holders. As enforcement of copyright has become more difficult with the proliferation of new technologies in the production and dissemination of copyright-protected works, the need for solutions employing technology appeared as well. Although cutting edge, new technology manifested in artificial intelligence and machine learning provide new possibilities for algorithmic copyright enforcement, they also present and potentially aggravate issues such as the lack of transparency and accountability, bias and the limitation of basic rights such

as the freedom of expression and information. These problems require specific attention with the DSM Directive entering into force: as Article 17 regards online content sharing service providers to be carrying out the copyright relevant act of communication to the public, these service providers could be exempt from infringement for copyright liability only if they demonstrate that they have made their best efforts to ensure the unavailability of unauthorized works on their platforms. This situation could easily prompt service providers to use the best and most effective tools.

The issues that could potentially emanate from the employment of AI and machine learning-based algorithmic enforcement mechanisms could be attenuated by two other provisions of the DSM Directive: the mandatory exceptions on text and data mining. Even though the original legislative intent behind the TDM-exception was to secure the development of data science and to close the gap that has appeared between the scientific community of Europe and other jurisdictions with more lenient copyright regimes (such as the United States, where the fair use doctrine offers a more flexible approach towards text and data mining, or China, where enforcement of intellectual property rights is still not in par with the European system), it seems to have a secondary, unintentional positive impact on algorithmic enforcement. It also serves as an example of how the different rules and the different sides of the same issue could be balanced out within the same legal instrument. Similarly, it is a reminder, that regulation and legislation concerning technology or other fields highly influenced by technology merit thorough preliminary analysis. Reactive law-making where only the existing problems are addressed with little to no consideration to the future direction of technological development and its possible implications should be avoided as it has the potential to result in an already obsolete and defunct regulation from the time of its entering into effect. This way, the potential benefits of AI and machine learning to copyright law could prospectively be overshadowed by the disadvantages and various issues brought about by these new technological phenomena.

LIST OF REFERENCES

[1] Akester, P. (2009) *Technological Accommodation of Conflicts between Freedom of Expression and DRM: The First Empirical Assessment.* Rochester, New York: Social Science Research Network.

[2] Allgrove, B. (2004) *Legal Personality for Artificial Intellects: Pragmatic Solution or Science Fiction?* [online] Available from: https://ssrn.com/abstract=926015 [Accessed 15 January 2019].

[3] Bamberger, K. A. (2010) Technologies of Compliance: Risk and Regulation in a Digital Age. *Texas Law Review*, 88 (4).

[4] Banko, M. and Brill, E. (2001) Scaling to Very Very Large Corpora for Natural Language Disambiguation. In: Bonnie Lynn Webber (ed.). *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, Toulouse, 6–11 July. USA: Association for Computational Linguistics.

[5] Bartholomew, T. B. (2015) The Death of Fair Use in Cyberspace: YouTube and the Problem with Content ID. *Duke Law & Technology Review*, 13 (1).

[6] Big Data Made Simple. (2014) *Top 14 useful applications of data mining.* [online] 20 August. Available from: https://bigdata-madesimple.com/14-useful-applications-of-data-mining/ [Accessed 4 February 2019].

[7] Calders, T. and Custers, B. (2013) What Is Data Mining and How Does It Work? In: Bart Custers et al (eds.) *Discrimination and Privacy in the Information Society.* 1st ed. Berlin: Springer.

[8] Čerka, P., Grigienė, J. and Sirbikytė, G. (2017) Is it possible to grant legal personality to artificial intelligence systems? *Computer Law & Security Review*, 33 (5).

[9] Channel Awesome. (2016) *Where's The Fair Use – Nostalgia Critic.* [online video] Available from: https://www.youtube.com/watch?v=zVqFAMOtwaI&t=53s [Accessed 10 January 2019].

[10] Citron, D. K. (2008) Technological Due Process. *Washington University Law Review*, 85 (6).

[11] Counter Extremism Project. (2018) *The eGlyph Web Crawler: ISIS Content on YouTube.* [online] Available from: https://www.counterextremism.com/sites/default/files/eGLYPH_web_crawler_white_paper_July_2018.pdf [Accessed 14 June 2019].

[12] Craig, C. J. (2017) Technological Neutrality: Recalibrating Copyright in the Information Age. *Theoretical Issues in Law*, 17 (2).

[13] Cummings, M. L. (2006) Automation and Accountability in Decision Support System Interface Design. *The Journal of Technology Studies*, 32.

[14] DeGroat, T. J. (2018) *19 Free Public Data Sets for Your Data Science Project.* [online] Available from: https://www.springboard.com/blog/free-public-data-sets-data-science-project/ [Accessed 15 June 2019].

[15] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal of the European Union* (2001/L-167/10) 22 June. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32001L0029&from=EN [Accessed 10 January 2019].

[16] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. *Official Journal of the European Union* (2019/L-130/92) 17 May. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2019.130.01.0092.01.ENG&toc=OJ:L:2019:130:TOC [Accessed 19 May 2019]

[17] DMR. (2019) *160 YouTube Statistics and Facts.* [online] Available from: https://expandedramblings.com/index.php/youtube-statistics/ [Accessed 11 January 2019].

[18] Elkin-Koren, N. (2017) Fair Use by Design. *UCLA Law Review*, 64 (5).

[19] Facebook. (2019) *Rights Manager.* [online] Available from: https://rightsmanager.fb.com/ [Accessed 10 January 2019].

[20] Facebook. (2019) *Terms of Service, Section 3.3: The permissions you give us.* [online] Available from: https://www.facebook.com/terms.php [Accessed 7 February 2019].

[21] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data. *Communications of the ACM*, 39 (11).

[22] Filippov, S. (2014) Mapping Text and Data Mining in Academic and Research Communities in Europe. *The Lisbon Council Special Briefing Issue*, (16).

[23] Fisher, W. W. (1988) Reconstructing the Fair Use Doctrine. *Harvard Law Review*, 101 (8).

[24] Fisher, W. W. (2001) Theories of Intellectual Property. In: Stephen Munzer (ed.). *New Essays in the Legal and Political Theory of Property.* 1st ed. Cambridge: Cambridge University Press.

[25] Friedman, B. and Nissenbaum, H. (1996) Bias in Computer Systems. *ACM Transactions on Information Systems*, 14 (3).

[26] Geller, P. E. (2008) Beyond the Copyright Crisis: Principles for Change. *Journal of the Copyright Society of the USA*, 55.

[27] Google. (2019) *Featured policies.* [online] Available from: https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism [Accessed 14 June 2019].

[28] Google. (2019) *Frequently asked questions about fair use.* [online] Available from: https://support.google.com/youtube/answer/6396261?hl=en [Accessed 15 June 2019].

[29] Greenberg, B. A. (2016) Rethinking Technology Neutrality. *Minnesota Law Review*, 100 (4).

[30] Halevy, A., Norvig, P. and Pereira, F. (2009) The Unreasonable Effectiveness of Data. *Intelligent Systems, IEEE*, 24 (2).

[31] I Hate Everything. (2015) *Cool Cat Saves The Kids – The Search For The Worst.* [online video] Available from: https://www.youtube.com/watch?v=HoTZZYm2HZI&t=42s [Accessed 10 January 2019].

[32] Joyce, C. (ed.). (2013) *Copyright Law.* 9th ed. New Providence: LexisNexis.

[33] Kerr, I. (2010) Digital Locks and the Automation Virtue. In Geist, Michael (ed.). *From „Radical Extremism" to „Balanced Copyright": Canadian Copyright and the Digital Agenda.* 1st ed. Toronto: Irwin Law.

[34] Knight, W. (2017) The Dark Secret at the Heart of AI. *MIT Technology Review*, 11 April. [online] Available from: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [Accessed 13 January 2019].

[35] Latman, A. and Patry, W. F. (1986) *Latman's the Copyright Law.* 6th ed. Washington, D.C.: Bureau of National Affairs.

[36] Litman, J. (2002) Revising Copyright Law for the Information Age. In: Adam Thierer and Wayne Crews (eds.) *Copy Fights: The Future of Intellectual Property in the Information Age.* 1st ed. Washington, D.C: Cato Institute.

[37] Leval, P. N. (1990) Toward a Fair Use Standard. *Harvard Law Review*, 103.

[38] Lessig, L. (2006) *Code v. 2.0*, 2006, New York: Basic Books. Available from: http://codev2.cc/download+remix/Lessig-Codev2.pdf [Accessed 10 January 2019].

[39] Lester, T. and Pachamanova, D. (2017) The Dilemma of False Positives: Making Content ID Algorithms more Conducive to Fostering Innovative Fair Use in Music Creation. *UCLA Entertainment Law Review,* 24 (1).

[40] Ryszard S. Michalski, Jaime G. Carbonell and Tom M. Mitchell (eds.). (1983) *Machine Learning: An Artificial Intelligence Approach.* 1st ed. Berlin: Springer-Verlag.

[41] Mills, M. L. (1989) New Technology and the Limitations of Copyright Law: An Argument for Finding Alternatives to Copyright Legislation in an Era of Rapid Technological Change. *Chicago-Kent Law Review*, 65 (1).

[42] Murphy, K. P. (2012) *Machine Learning: A Probabilistic Perspective.* 1st ed. Cambridge: Massachusetts Institute of Technology.

[43] Myška, M. (2009) The True Story of DRM. *Masaryk University Journal of Law and Technology*, 3 (2).

[44] Nordemann, J. B. (2017) Liability of Online Service Providers for Copyrighted Content – Regulatory Action Needed? In-Depth Analysis for the IMCO Committee. *Directorate--General for Internal Policies, Policy Department A (Economic and Scientific Policy), European Parliament.*

[45] *Newton v. Diamond* (2004) 388 F.3d 1189, 7 April.

[46] Omnicore. (2019) *Instagram by the Numbers: Stats, Demographics & Fun Facts.* [online] Available from: https://www.omnicoreagency.com/instagram-statistics/ [Accessed 11 January 2019].

[47] Perel, M. and Elkin-Koren, N. (2016) Accountability in Algorithmic Copyright Enforcement. *Stanford Technology Law Review*, 19 (3).

[48] Perel, M and Elkin-Koren, N. (2017) Black Box Tinkering: Beyond Disclosure in Algorihtmic Enforcement. *Florida Law Review*, 69 (1).

[49] PewDiePie. (2017) *Life is cringe – life is strange – S2E01.* [online video] Available from: https://www.youtube.com/watch?v=PX4zk0G4IjM [Accessed 10 January 2019].

[50] European Commission. (2016) *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market.* (COM(2016) 593 final) 14 September. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52016PC0593 [Accessed 15 January 2019].

[51] European Parliament. (2018) *Report on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market.* (COM(2016)0593 – C8-0383 (2016) – 2016/0280(COD)). Available from: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=%2F%2FEP%2F%2FTEXT%2BREPORT%2BA8-2018-0245%2B0%2BDO%2BXML%2BV0%2F%2FEN&language=EN [Accessed 7 February 2019].

[52] Richard, K. (2018) Fair Use in the Information Age. *Richmond Journal of Law & Technology*, 25 (1).

[53] *Ringgold v. Black Entertainment Television, Inc.* (1997) 126 F.3d 70, 16 September.

[54] *Sandoval v. New Line Cinema Corp.* (1998) 147 F.3d 215, 24 June.

[55] Sarikaya, R., Geoffrey E. and Deoras, A. (2014) Application of Deep Belief Networks for Natural Language Understanding. *IEEE Transactions on Audio, Speech and Language Processing*, 22 (4).

[56] Solum, L. B. (1991) Legal Personhood for Artificial Intelligences. *North Carolina Law Review*, 70 (4).

[57] Stamatoudi, I. and Torremans, P. (2014) *EU Copyright Law, a Commentary.* 1st ed. Cheltenham: Edward Elgar Publishing Limited.

[58] Stanford, S. (2018) *The Best Public Datasets for Machine Learning and Data Science.* [online] Available from: https://medium.com/towards-artificial-intelligence/the-50-best-public-datasets-for-machine-learning-d80e9f030279 [Accessed 15 June 2019].

[59] Thatcher, S. G. (2006) Fair Use in Theory and Practice: Reflections on its History and the Google Case. *Journal of Scholarly Publishing*, 37 (3).

[60] Triaille, J.P., de Meeus d'Argenteuil, J. and de Francquen, A. (2014) *Study on the legal framework of text and data mining (TDM).* 1st ed. Luxembourg: European Union. Available from: https://publications.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en [Accessed 4 February 2019].

[61] United States Copyright Office. (2019) *More information on fair use.* [online] Washington, D. C.: USCO. Available from: https://www.copyright.gov/fair-use/more-info.html [Accessed 23 May 2019].

[62] Vorhies, W. (2016) *CRISP-DM – a Standard Methodology to Ensure a Good Outcome.* [online] Data Science Central. Available from: https://www.datasciencecentral.com/profiles/blogs/crisp-dm-a-standard-methodology-to-ensure-a-good-outcome [Accessed 4 February 2019].

[63] Witten, I. H. and Frank, E. (2005) *Data Mining, Practical Machine Learning Tools and Techniques.* 2nd ed. San Francisco: Morgan Kaufmann Publishers.

[64] YouTube. (2017) *An update on our commitment to fight violent extremist content online.* [online] Available from: https://youtube.googleblog.com/2017/10/an-update-on-our-commitment-to-fight.html [Accessed 13 January 2019].

[65] YouTube. (2019) *Content Verification Program.* [online] Available from: https://support.google.com/youtube/answer/6005923 [Accessed 14 June 2019].

[66] YouTube. (2019) *Copyright Management Tools.* [online] Available from: https://support.google.com/youtube/answer/9245819?hl=en [Accessed 14 June 2019].

[67] YouTube. (2019) *Copyright Management Tools – Content ID.* [online] Available from: https://support.google.com/youtube/answer/9245819 [Accessed 10 January 2019].

[68] YouTube. (2019) *Copyright Match Tool.* [online] Available from: https://support.google.com/youtube/answer/7648743 [Accessed 14 June 2019].

[69] YouTube. (2019) *How Content ID works.* [online] Available from: https://support.google.com/youtube/answer/2797370?hl=en&ref_topic=2778544 [Accessed 10 January 2019].

[70] YouTube. (2019) *Terms of Service, Section 8: Rights you license.* [online] Available from: https://www.youtube.com/static?template=terms [Accessed 7 February 2019].

[71] YouTube. (2019) *YouTube in Numbers.* [online] Available from: https://www.youtube.com/yt/about/press/ [Accessed 14 June 2019].

[72] Zhang, M., Zhang, Y. and Fu, G. (2016) Tweet Sarcasm Detection Using Deep Neural Network. In: Eiichiro Sumita, Takenobu Tokunaga and Sadao Kurohashi (eds.). *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan, 11–16 December. Japan: Japanese Association of Natural Language Processing.