Access to data for training algorithms in machine learning – copyright law and "right-stacking"

Inger B. Ørstavik, professor, Department of Private Law, University of Oslo

Abstract (revised)
News services and information gathering services are increasingly relying on automated processes. As information sources are unlimited and often unreliable, it is important from an information security perspective that services using artificial intelligence to process information are able to draw and present reliable conclusions. This Chapter examines if and how exclusive rights in copyright law restrict data access for training algorithms for the purpose of machine learning. Because access to huge amounts of data is necessary to train algorithms to develop high quality artificial intelligence services, the rights of authors to individual works as well as rights in collections of works – database rights and publisher's rights – must be cleared. The Chapter discuss the role of copyright law in striking a balance between right holders and AI developers and how public interests can be included in the balance. The discussion plays into the larger policy debate: whether the role of intellectual property rights is developing from incentive mechanisms to control mechanisms.

1. Introduction

One effect of the digital revolution in the media industry is the employment of automated digital tools in the gathering, production and presentation of journalistic news and information services. While representing challenges on its own, artificial intelligence (AI) and its tools are valuable in tackling the challenges that digitalization also represents for journalism. Digitalization entails an information overload, as all content from every source is available instantly in digital format. Further, news production and distribution is becoming more pluralistic, moving away from traditional media and into social media channels without a legally responsible editor. Many media houses use AI systems to assist journalists in handling the massive amount of information readily available, by tracking down breaking news or by assisting in research, for instance by highlighting information deviating from a norm for evaluation by a human reporter.[1] The first "robot reporters" can provide brief notices of companies' financial reports or minor league sporting events.[2] AI's ability to handle bulk reading also enables automated tools for fact checking and efficient and reliable corroboration of information.[3]

[1] See Corinna Underwood, 'Automated Journalism – AI Applications at New York Times, Reuters, and Other Media Giants', 2019, https://emerj.com/ai-sector-overviews/automated-journalism-applications/, accessed 03.12.2020.
[2] See Jaclyn Peiser, 'The Rise of the Robot Reporter', 5 February 2019, New York Times, https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html, accessed 2 November 2020; David Caswell and Konstantin Dörr, 'Automating Complex News Stories by Capturing News Events as Data', Journalism Practice, 2019, 951–955, DOI:10.1080/17512786.2019.1643251; Javier Díaz-Noci, 'Artificial Intelligence Systems-Aided News and Copyright: Assessing Legal Implications for Journalism Practices', Future Internet 2020, 85; doi:10.3390/fi12050085.
[3] Richard Fletcher et al, 'Building the "Truthmeter": Training algorithms to help journalists assess the credibility of social media sources', Converg. Int. J. Res. New Media Technol., 2020, 19–34, at 20–21, doi.org/10.1177/1354856517714955.

Finally, most large news publishers use algorithms to support differentiated homepages and advertising to their readers.

Automated techniques applying AI are developed using machine learning techniques. Machine learning (ML) is an overall description of a process where computer systems are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data in order to make predictions or decisions without being explicitly programmed to do so.[4] An algorithm is a set of mathematical instructions or rules to a computer that will help it calculate an answer to a problem.[5] In automated systems, algorithms are trained upon massive amounts of data to optimize their problem-solving when applied to new data in the "real" world. To develop services related to the news industry and journalism, training data can consist of journalistic articles, notices, reports, social media posts, tweets and similar.

This Chapter discuss the role of copyright law in balancing the interests of authors and of owners of collections of protected works against the interests of AI developers and service providers, also considering the public interest in development and use of AI. By framing the discussion in the news services industry, public interests on both sides can be highlighted, making the findings applicable across industries. On the one hand, there is a need to promote development of automated tools to ensure information integrity. [6] The problem is extrapolated in the buzzword "fake news":[7] Verification of information and translucency in presentation of information is threatened as news and information production and distribution becomes more pluralistic. Furthermore, social media has the potential to exponentially magnify the impact of false or misleading information. There is an overt public interest in having a public discourse that is informed, not misinformed. On the other hand, AI systems, including well-trained algorithms, can be extremely valuable and overtake traditional media consumption based on human reading. To sustain human journalistic and edited content production, the law should ensure that authors and publishers are awarded a fair share of such earnings.

While the above mentioned public interests concern the application of AI, that is, operation of the services employing AI, a more nuanced approach is necessary with regard to development of AI and copyright law. Focus in this Chapter is on the use of works and collections of works in connection with training of algorithms to develop AI. An information service, such as an automated service offering summaries of news stories, "reads" news articles when operating. Development of the service,

[4] See e.g. https://en.wikipedia.org/wiki/Machine_learning.

[5] Drexl, Hilty et al, 'Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective', Version 1.0, October 2019, , available at: https://ssrn.com/abstract=3465577, 4 (Drexl, Hilty et al).

[6] Fletcher et al, (n 3), 19–20.

[7] The term no longer carries a specific meaning, but its connotations are illustrative. See Edson Tandoc, Zheng Wei Lim & Richard Ling 'Defining "Fake News"', Digital Journalism, 2017, 1–17. Available at: https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1360143, accessed 2 December 2020. See also Margi Murphy, 'Government bans phrase "fake news"', The Telegraph 23 October 2018, available at https://www.telegraph.co.uk/technology/2018/10/22/government-bans-phrase-fake-news/, accessed 2 December 2020.

however, has likely involved use of other works or collections of works, upon which the algorithm involved has "trained" to enhance its functionality. The balancing of interests between right holders and developers is not necessarily the same at the development stage as at the operating stage, and nor are the public interests involved. There is also a discrepancy between the development stage and the operating stage, as a training corpus may be used to train different algorithms, and one algorithm may be used to develop different services. The analysis here focuses on the development stage, that is, the training of an algorithm, but will also take into account the application of AI in a service, to pave the way for a discussion on how the interests at these different stages can be aligned.

Legal problems pertaining to the use by AI of copyrighted works have been subject to scrutiny in literature, as regards whether copyright protection is available for AI-produced works,[8] and whether authors' rights over their works could also extend to output produced by AI.[9] Furthermore, text and data mining techniques (TDM) have been subject to analysis.[10] While clearly overlapping, there are differences between TDM and ML that should be discussed. The literature offers very little concrete analysis of the application of EU copyright law to training of algorithms as part of ML. Under US law, ML is considered under the fair use doctrine; thus US law does not provide direct guidance for European law. The contribution of this Chapter is a discussion of the room for training algorithms in ML in EU copyright law, including the new DSM Directive.

The Chapter starts with a presentation of ML and algorithms in section 2. Here, it will also be made clear how and why a copyright law analysis must differentiate between development of AI and application of AI-based services. In section 3, the question is first whether the process of training an algorithm infringes the right of reproduction to individual works under Article 2 InfoSoc Directive.[11] Training an algorithm, however, requires access to a large amount of data. The discussion is restricted to training models where the algorithm trains on a centralized collection of data, a training corpus. As training does not involve transfer of any of the data, only the right of reproduction is discussed in relation to individual works. Compiling and preparing a training corpus, as well as using it to train an algorithm, may infringe rights in collections of data, that is, database rights under the Database

[8] See Pamela Samuelson, 'Allocating Ownership Rights in Computer-Generated Works', 47 U PITT L REV 1185, 1226–28 (1986); Bruce E Boyden, 'Emergent Works', 39 COLUM JL & ARTS 377 (2016); Pratap Devarapalli, 'Machine learning to machine owning: redefining the copyright ownership from the perspective of Australian, US, UK and EU law, 2018 EIPR 40(11), 722–728; Bob L T Sturm et al, 'Artificial Intelligence and Music: Open Questions of Copyright Law and Engineering Praxis', Arts 2019, 8, 115; doi:10.3390/arts8030115.

[9] See Mirko Degli Esposti et al, 'The Use of Copyrighted Works by AI Systems: Art Works in the Data Mill', European Journal of Risk Regulation, 11 (2020), 51–69, doi:10.1017/err.2019.56 (Esposti et al), and Ole-Andreas Rognstad in another contribution to this book.

[10] See section 4.1 below and references there.

[11] 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.

directive[12] and the new publisher's right in Article 15 DSM Directive.[13] This is discussed in section 3.2 and 3.3.

Section 4 turns to interpretation of the exceptions from exclusive rights, and their application to ML. Focus is on the new exception for commercial text and data mining (TDM) in Article 4 DSM Directive. Section 4.2 asks whether the exception for temporary digital copying of copyrighted works in Article 5(1) InfoSoc Directive could supplement the exception for TDM.

In section 5, the discussion returns to the balance struck between the interests of right holders and AI developers, bringing together the policy arguments from the earlier sections, in particular whether a case can be made for maintaining an approach to copyright as a control instrument to ensure freedom of information. The economic rationales behind database and press publisher's rights better support control with ML, but do not give sufficient basis for controlling the societal impacts of services employing AI.

2. Machine learning and algorithms

This section takes a peek into the "black box" of ML to concretize how AI systems are developed and what training an algorithm means in practice.[14] Most AI services in use in the news industry today rely on *supervised machine learning*, where an algorithm is trained based on a pre-defined set of training data and on given examples. By using regression and classification techniques, the algorithm learns to make consistent predictions when applied to new data. More sophisticated machine learning processes exist. *Unsupervised learning* relies on unlabelled -that is, unstructured – data. The model is trained to identify similarities, parallels or differences in data, mainly using clustering techniques. The training process requires less human participation, as the training data do not have to be labelled and structured, but interpretation of the output requires more human involvement.[15] *Reinforcement learning* relies on evaluation of the results provided by the system. The system relies on algorithms to classify and define results. The newest learning methods are *neural networks*, where the system gives its own feedback for further learning, based on a pre-determined algorithm, thereby "learning" without human assistance.[16] In this Chapter, focus is on supervised learning, but as the research question

[12] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.
[13] 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.
[14] This section relies on the extensive and qualified discussions in David Lehr and Paul Ohm, 'Playing with the Data: What Legal Scholars Should Learn About Machine Learning', 51 UC Davis L Rev 653, 655–717 (2017); Thomas Margoni, 'Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?' CREATe Working Paper 2018/12, DOI: 10.5281/zenodo.2001763 (Margoni); Mauritz Kop, 'Machine Learning & EU Data Sharing Practices', Stanford-Vienna Transatlantic Technology Law Forum, 1/2020, 7 (Kop), and Drexl, Hilty et al (n 5).
[15] Definitions from Drexl, Hilty et al (n 5), 8.
[16] See Andres Guadamuz, 'Do androids dream of electric copyright? Comparative analysis of originality in artificial intelligence generated works', 2017 Intellectual Property Quarterly, 169 at 171.

relates to the process of training and not the output of the algorithm after training, the conclusions can apply similarly to other learning methods.

Very briefly, machine learning consists of three stages: the development and programming of a model architecture; the training process where a model is developed based on a training algorithm and training data; and finally the model is applied to new data.[17] In the first phase, the programmers define the problem that the algorithm will predict or evaluate. The algorithm can be chosen among standard algorithms available in online libraries or developed individually for the project. The criteria that the algorithm will apply to evaluate results must be defined. The purpose of what is here termed the "AI system" or "service" might differ from what the algorithm measures. The service might be an automatic personalization of the display of articles in an online newspaper, but the algorithm will measure the particular reader's historical reading habits to make predictions for future reading interest.

For the training stage, a corpus of training data must be collected. There must be a sufficient amount of data, the data must be representative, that is, giving variables measurement validity, and it must be generalizable, that is, give a basis for the algorithm to make the "right" decisions when running on new data. Here, I assume that the training data consists of natural language texts, such as articles, comments, posts, tweets, and so on. I also assume that these texts are protected by copyright. However, other materials may be included, such as dictionaries, data, databases and other works such as music, pictures, and the like. [18]

In a *centralized learning method*, the training data is collected as a corpus, and either downloaded to a local server, or access is secured for the software to run on a corpus of data remotely stored. Pre-processing activities prepare the data for the statistical analysis that the training of the algorithm really amounts to. Texts are automatically converted to a format that the system can read: plain text or similar. The data is "cleaned", correcting missing values, sorting out incorrect data and outliers that would obstruct generalizability. To define examples and classifications for supervised learning, metadata is added to label the text. These annotations define the text using the types of classifications that the algorithm will use as predefinitions for right or wrong answers in its training.

In the training process, the algorithm "learns" abstract probabilistic characteristics from the training data, which it will use to predict learned labels on unseen data.[19] The labels can be names, part-of-language tags, sentiment or meaning tags, and so on. The "learned" information is stored in a separate file, which once accumulated becomes the "trained model". After training, the algorithm can run without access to the training data. The algorithm runs back and forth on the data, testing a large number of "hypothesis" rules to enhance its function in a non-linear process of tuning, assessment and

[17] Drexl, Hilty et al (n 5).
[18] The findings here will be relevant to other kinds of works, although supplemental analysis is needed to include the particularities of each type of work. See eg Sturm et al (n 8).
[19] Margoni (n 14), 2.

evaluation. Those rules that prove correct based on the training data will be included in the final algorithm. The point is to enhance the function of the algorithm by minimizing the number of wrong predictions.

In a *decentralized learning model*, the algorithm is distributed to multiple decentralized devices, connected over the internet. The algorithm runs on the materials stored on local devices, and uploads the "learned" information to a central server, where all the training information from all the devices collectively make up the trained algorithm. Decentralized learning methods do not require compilation of a training data corpus.[20] These models may involve transfer of data from the training materials and might raise questions with regard to the right to communication to the public in Article 3 InfoSoc Directive. This is not discussed further here, but could be the subject of further research.

A machine learning model can be either static – that is, training is completed before the model is actually applied – or it can be dynamic – that is, training never ends, as the output of the system when applied continues to modify and optimize the system. A dynamic model requires constant feedback on the correctness of the output of the algorithm when applied. [21] An algorithm that displays articles on a web page based on readers' preferences is typically learning constantly. In dynamic algorithms, the learning process cannot be clearly discerned from the AI service.

ML uses very similar techniques as in text and data mining (TDM). TDM can be defined as computerized processes aimed at extracting and elaborating information from large digital data sets, extracting new knowledge, in particular by identifying correlations and trends.[22] TDM also involves a preparation phase where data are collected and prepared for automated analysis, and information is extracted by running an algorithm on the data. In a final recombination phase, the extracted data are arranged to visualise results.[23] In ML, the information extracted when running the algorithm on the data is incorporated into the algorithm, and the trained algorithm can be used for various purposes. Section 4.1 discusses whether these differences have legal implications.

The next section analyses exclusive rights to materials and collections that might need clearance to facilitate ML, starting with copyright in the individual materials (3.1) and then rights in collections of works (3.2 and 3.3).

[20] See Kop (n 14) with further references.
[21] Drexl, Hilty et al (n 5), 6.
[22] Maurizio Borghi and Stavroula Karapapa. *Copyright and Mass Digitization*, OUP 2013, 47. DOI:10.1093/acprof:oso/9780199664559.003.0003; Max Planck Position Statement, 1, available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2900110, accessed 3 December 2020.
[23] Romain Meys, 'Data Mining Under the Directive on Copyright and Related Rights in the Digital Single Market: Are European Database Protection Rules Still Threatening the Development of Artificial Intelligence?' GRUR Int., 2020, 457, doi: 10.1093/grurint/ikaa046 (Meys).

3. Exclusive rights in training materials

3.1. The right of reproduction, Article 2 Infosoc Directive

The wording of Article 2 InfoSoc Directive reserves for the copyright holder 'temporary or permanent reproduction by any means and in any form, in whole or in part'. The acts covered by this right of reproduction are construed broadly.[24] The CJEU applies a formal and technical approach, extending the right to every act of reproduction, 'however transient or irrelevant it may be from an economic perspective'.[25] This includes intermediate digital copies in the RAM memory of the computer, as well as other digital copies regardless of whether they are accessible by a human reader, such as cache memory, even if such copies may be intrinsic to (lawfully) accessing a work by computer, such as online browsing.[26]

The question is whether training an algorithm entails copying of training materials that infringe the right of reproduction. For centralized machine learning models, compiling the training data either by uploading it to a platform or by downloading it to a local server, results in copies of the materials. These copies are likely to have a permanent character in the sense of the Infosoc Directive, even if the corpus is deleted after completion of the training process.[27] Furthermore, pre-processing the training corpus could entail acts of reproduction.[28] Materials created for human reading, such as pdf-files, have to be converted to machine-readable formats, thereby creating a copy of the materials.[29] "Cleaning" the training data by sorting out outliers or irregular data could also include copying. Preliminary data processing may employ crawlers, but these will also "read" materials by creating temporary copies in the cache memory of the computer. Adding metadata and annotations to the materials – that is, supervising the algorithm – could also include such temporary copying.[30] Thus, it seems quite clear that compilation of a training corpus in a centralised learning model would entail several acts of reproduction infringing Article 2 InfoSoc Directive.

Whether the actual "learning" process – where the algorithm is running on the data – infringes the right of reproduction could be questioned. The algorithm can go back and forth between the data, as it is "testing" and modifying its "rules." The information extracted is factual information that when aggregated gives statistical information about correlations, trends, differences, and the like in the

[24] Preamble (21) InfoSoc Directive.
[25] Michel M Walter and Silke von Lewinsky, *European Copyright Law*, OUP, 2010, 968.
[26] *Ibid*.
[27] Art 5(1) InfoSoc Directive, discussed below, section 4.2. See Maarten Truyens and Patrick van Eecke, 'Legal aspects of text mining', Computer Law and Security Review, 2014. 153–170, at 162.
[28] Christophe Geiger et al., 'Text and data mining in the proposed copyright reform: making the EU ready for an age of big data? Legal analysis and policy recommendations', IIC 2018, 814-44, https://doi.org/10.1007/s40319-018-0722-2, at 818 (Geiger et al 2018).
[29] See Jean-Paul Trialle, 'Study on the legal framework of text and data mining' (TDM), EC Commission, 2014, 32, available at https://op.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en, accessed 3.12.2020 (Triaille).
[30] Esposti et al (n 9), 15.

training data. This information is stored in a separate file. When included in the final algorithm it takes the form of "rules" based on statistics, and the original works are not recognizable. The stored information is not likely to include sufficient parts of the training materials to infringe the right of reproduction. In the literature, some doubt has been expressed as to whether an algorithm, when running on the materials, makes relevant copies of the materials.[31] But, as even cache memory copies are considered reproductions under Article 2 InfoSoc Directive, most analytic techniques, including crawling, will entail copying, even if no human-readable copies are made.[32] Cache memory copying is intrinsic to the way a machine "reads". It therefore seems likely that training an algorithm infringes the right of reproduction.[33]

This very broad and formal construction of the right of reproduction has been criticised for going beyond the fundamentals of copyright, in particular the incentive and remuneration rationale for exclusive rights, which does not include use of a work to gather information as opposed to use of the work "as a work".[34] Seeing that copyright protects free speech to spread original ideas and enlightened communication between humans, acts of reproduction that do not impact this author-audience nexus should also fall outside the scope of copyright.[35] Following this argument, training algorithms with the objective of developing new services should not infringe copyright, as ML make use of the facts and information extracted from works, but not their creative expression.[36] This is not to say that access to copyright-protected works should be free and without restrictions for the purposes of ML. Clearly financial interests are involved in machine learning, and well-trained algorithms are extremely valuable, as is evidenced by large content service providers such as Netflix and Amazon. The value of the service can be traced back to access to the works in a training corpus. This value, however, does not relate to individual works, but to the fact that the training corpus consists of a large number of works. This is an economic value rather associated with collections of works (databases) than with individual creative works, and probably better managed through database rights or the new press

[31] Meys (n 23), 460 and Geiger (n 38).
[32] CJEU case C-360/13, *Meltwater*, ECLI:EU:C:2014:1195. See also Rossana Ducato and Alain Strowel, 'Limitations to text and data mining and consumer empowerment: making the case for a right to "machine legibility"', IIC 2019, 50(6), 649–684, 658 (Ducato and Strowel).
[33] Christoph Geiger et al, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU', CEIPI Studies Research Paper No. 2019-08, https://ssrn.com/abstract=3470653 at 8 (Geiger et al 2020). The author's right is infringed if so much of the work is copied that it includes subject matter that is 'original in the sense that it is its author's own intellectual creation', case C-05/08, *Infosoc I*, para 37, ECLI:EU:C:2009:465.
[34] See in general P Bernt Hugenholtz (ed), *Copyright Reconstructed: Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change*, Alphen aan den Rijn, 2018. For TDM, see Ducati and Strowel (n 32), 667–668.
[35] Alain Strowel, 'Reconstructing the Reproduction and Communication to the Public Rights: How to Align Copyright with its Fundamentals' (Strowel), in Hugenholtz (n 34), 206–209.
[36] Lemley, Mark A and Casey, Bryan, Fair Learning (January 30, 2020). Available at SSRN: https://ssrn.com/abstract=3528447 or http://dx.doi.org/10.2139/ssrn.3528447, at 152, accessed 3 December 2020 (Lemley and Casey).

publisher's right.[37] Furthermore, requiring individual consent from right holders for all copying in the processes of ML is likely to be prohibitive to the development of AI-based services. Copyright risks being used to block the spread of information, thus working against the original aim of copyright.[38] On the other hand, the output of ML can be used for any purpose, and individual authors may feel strongly opposed to their works being used to develop services such as pre-emptive policing where individuals are sought out for surveillance if the algorithm finds them likely to commit crimes based on their social media postings.[39] As discussed below, a possible way to safeguard the interests of individual authors would be to require database owners to allow authors to opt out of including their works in ML licensing, and to require consent from database owners or press publishers for ML. A topic for further research would be to consider how individual authors interests could be protected by limited "digital ideal rights" when full exclusive economic rights are not warranted.

With the introduction of exceptions for text and data mining in the DSM Directive Articles 3 and 4, there is a presumption that an exception is also needed, and discussion of whether ML is copying relevant to copyright seems closed, at least for the time being.[40] For all practical purposes, the question is whether the process of training an algorithm can benefit from any of the exceptions to exclusive rights. The EU lawmaker has confirmed its approach to copyright as broad and all-encompassing exclusive rights. In addition, individual copyrights are supplemented with rights-protecting investment in collections of works, such as database rights and the new publisher's right in Article 15(1) DSM Directive. This layered system of rights poses additional challenges with regard to ML. The only practical approach to gaining access to the amount of works necessary for ML is through intermediaries, typically publishing houses or other entities holding libraries of relevant works. In sections 3.2 and 3.3, the discussion turns to database rights and the new publisher's right.

3.2. Database rights

A set of training data must contain a large amount of materials, and it is not practical to collect these individually. Collections of works may be subject to database rights.[41] A database may be protectable by copyright – Article 5 Database Directive if the choice of materials or their structuring in the database demonstrates originality.[42] Copyright to the database is only likely to be infringed if the whole database is downloaded for incorporation in a centralized training corpus, as only then will the

[37] Analysis of economic arguments in Joost Poort, 'Borderlines of Copyright Protection: An Economic Analysis', in Hugenholtz (n 34), (Poort).
[38] Strowel (n 35), 226–228.
[39] See eg the description of the AI system GPT2 in Alex Hern, 'New AI fake text generator may be too dangerous to release, say creators', The Guardian, 14 February 2019, available at https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction, accessed 3.12.2020.
[40] See preamble (8) DSM Directive and Geiger et al 2018 (n 28).
[41] A database is 'a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means', Art 1(2), Database Directive.
[42] Art 3(1) Database Directive.

structure of the database also be reproduced.[43] Hence, the question is whether running an algorithm on a database infringes the *sui generis* database right in Article 7(1) of the Database Directive, protecting databases for which "obtaining, verification or presentation of the contents" requires "a substantial investment". I discuss only the right of extraction, as the right to "re-utilization" in Article 7(2) (b) is not likely to be infringed in centralized ML models.

The right of extraction is related to the right of reproduction and defined in Article 7(2) (a) as 'the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form'.[44] It also covers the 'repeated and systematic' extraction of insubstantial parts of the database.[45] The CJEU has emphasized the economic interests of the *sui generis* right holder in protecting their investment, finding it less relevant whether copies are technically being made and transferred to another medium.[46] Whereas searching the database may entail digitally copying contents, this does not infringe the extraction right if the consultation is lawful.[47] The *sui generis* right is infringed if the investment in making the database is harmed, which it is if the right holder is deprived of revenue which should have enabled them to redeem the cost of investment.[48]

Applied to ML and training algorithms, it seems likely that downloading the whole or a substantial part of the database to compile a centralized training data corpus, would infringe the right of extraction. A labelled, annotated training data set can itself be awarded data base rights, as this type of investment is what the *sui generis* right is meant to protect.[49]

Under the *sui generis* right, the question whether actual training of the algorithm infringes the right, does not hinge on whether the algorithm makes temporary copies of materials in the cache memory of the computer or otherwise copies materials, but whether the right holder's investment is harmed. ML makes use of the economic value associated with works being part of a large collection of works. This is exactly the investment in the database that the *sui generis* right protects, and it therefore appears difficult to find arguments supporting a finding that ML should not require a licence.[50] On the other hand, ML is not an activity that could be described as "parasitical competing" activities.[51] ML rather reveals new knowledge and facilitates new and innovative services. However, when activities

[43] See Triaille (n 29), 34–35.
[44] Rec (44).
[45] Article 7(5).
[46] CJEU case C-304/07, *Directmedia*, ECLI:EU:C:2008:552, paras 33 and 35.
[47] CJEU case C-203/02, *William Hill*, ECLI:EU:C:2004:695 para 54; case C-304/07, *Directmedia*, ECLI:EU:C:2008:552, para 51.
[48] Recital (49) and CJEU case C-203/02, *William Hill*, ECLI:EU:C:2004:695, para 51; case C-202/12, *Innoweb*, ECLI:EU:C:2013:850, para 37.
[49] See Kop (n 14), 7.
[50] See CJEU case C-490/14, *Verlag Esterbauer*, ECLI:EU:C:2015:735, para 16; C-202/12, *Innoweb*, ECLI:EU:C:2013:850, para 46–48. See Lemley and Casey (n 36), 127; Geiger et al 2018 (n 28), 823–824.
[51] Recital (42) Database Directive; CJEU case C-203/02, *William Hill*, ECLI:EU:C:2004:695 para 47.

appropriate the value inherent in the database, the CJEU has emphasized that it is legitimate for the database holder to reserve a fee in consideration for use of the database.[52]

Even if the concrete ML model did not exploit the database by repeatedly extracting insubstantial parts contrary to Article 7(5), it is doubtful whether ML could be considered as "normal exploitation" of a database under Article 8(2) if not explicitly included in a licence.[53] The criterion is undetermined, but as discussed above, if the use harms the investment in the database, then the *sui generis* right is likely to be infringed.[54] TDM techniques have been considered in literature as going beyond "normal" exploitation of a database, although this might depend on a concrete assessment of what use the database was meant for, and what techniques are employed, that is, bulk reading, crawling, scraping or other machine-enabled analysis.[55] In conclusion, the exception for "normal" use in Article 8(2) does not provide sufficient legal certainty as a basis for ML.

3.3. Press publishers' right

The DSM directive Article 15 introduces a new press publisher's right that has been heavily criticised.[56] Under Article 15(1), press publishers are granted full exclusive rights as in Article 2 and 3(2) InfoSoc Directive, but only against "the online use of their press publications by information society service providers", and only for a period of two years from publication.[57] An information society service is 'any service normally provided for remuneration, at a distance, by electronic means and at the individual request of a recipient of services', that is, any service provided individually over the internet.[58] Due to the short duration of the right and its restricted scope to information society service providers, it will most likely only restrict dynamic ML, for example, use by news aggregation services of continuously improving and self-learning algorithms in their service. In these instances, the new right may provide an opportunity for centralized consent for AI services, but is not likely to have a separable impact on ML and training of algorithms.

---

[52] CJEU case C-203/02, *William Hill*, ECLI:EU:C:2004:695, para 57.
[53] Preamble (24). See also Geiger et al 2020 (n 33), 15–16.
[54] See also Estelle Derclaye, 'The Database Directive', in Irini Stamatoudi and Paul Torremans (eds), *EU Copyright Law*, Elgar, Cheltenham, 2014, 111.
[55] Irini Stamatoudi, 'Text and Data Mining', in Irini Stamatoudi (ed), "*New Developments in EU and International Copyright Law"*, Wolters Kluwer 2016, 278; Geiger et al 2018 (n 28), 823–824.
[56] The criticism concerns the paradox of ensuring a free press and free access to information by introducing yet another exclusive right, see preamble (54) DSM Directive. See eg Lionel Bentley et al, 'Strengthening the Position of Press Publishers and Authors and Performers in the Copyright Directive', study for DG IPOL, 2017, available at https://www.europarl.europa.eu/RegData/etudes/STUD/2017/596810/IPOL_STU%282017%29596810_EN.pdf, accessed 30 November 2020; Lionel Bentley et al 'Response to Article 11 of the Proposal for a Directive on Copyright in the Digital Single Market, entitled "Protection of press publications concerning digital uses" on behalf of thirty-seven professors and leading scholars of Intellectual Property, Information Law and Digital Economy', 2016, https://www.cipil.law.cam.ac.uk/press/news/2016/12/cambridge-academics-respond-call-views-european-commissions-draft-legislation, accessed 3 December 2020.
[57] Article 15(4).
[58] Directive (EU) 2015/1535 of the European Parliament and of the Council of 9 September 2015 laying down a procedure for the provision of information in the field of technical regulations and of rules on Information Society services Art. 1 (1) (b).

If protected materials are used in centralized ML within the two-year exclusivity period, then interpretation of the right raises some questions. First, the exclusive right does not extend to facts.[59] However, as the right is a full exclusive right of reproduction, it includes any temporary machine-generated copying. Thus, an exception might still be necessary to allow machine access to the data. Second, the protected subject matter is the publications contained, but in determining how much content can be reproduced without infringement the decisive factor is whether the investments by publishers in production of the content are undermined.[60] Only reproductions of individual words or short extracts can be lawful, for which purpose the exceptions in Article 5(1) InfoSoc Directive apply *mutatis mutandis.*[61] This combination of the different rationales from individual copyright and database rights that differ from the rationale behind the press publishers' right is somewhat of a paradox. The preamble uses the practical term "use" of publications to explain the scope of the right.[62] This could possibly signal a step away from the formalistic approach of the CJEU in interpreting the right of reproduction in Article 2(2) InfoSoc Directive, but the reference to the exceptions in Article 5(1) InfoSoc Directive probably leaves it with the CJEU to interpret whether ML infringes the press publishers' right. Finally, as individual or collective copyrights and database rights[63] are not affected by the new publishers' right,[64] the obstacles to obtaining authorization for ML are hardly overcome with the introduction of the press publishers' right.

4. Are the exceptions to exclusive rights sufficient to facilitate machine learning?

4.1. The exceptions for text and data mining in the DSM Directive

4.1.1.Definition

The DSM directive also introduced new exceptions to exclusive rights for text and data mining (TDM). The first question is whether the definition of TDM could also cover machine learning, that is, training algorithms. The DSM Directive Article 2(2) defines TDM as 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations'. The definition does not limit the type of digital data or the tools involved.[65] An algorithm can be an automated analytical technique, and it analyses text and data in digital form. When training an algorithm, the algorithm extracts trends and correlations from the training data patterns, in the same way as a TDM process.[66] The difference is that in TDM the purpose is to present the extracted information, whereas in ML the algorithm adds the new information to its set of "rules" in an automated process, enhancing its ability to make accurate

[59] Rec (57) DSM Directive.
[60] Rec (58) DSM Directive.
[61] Article 15(3).
[62] Rec (58) DSM Directive.
[63] Many press publishers will also have *sui generis* database rights in their online publications, see Lionel Bentley et al, "Strengthening the Position of Press Publishers and Authors and Performers in the Copyright Directive", study for DG IPOL, (n 55), 23.
[64] See Article 15(2).
[65] Meys (n 23), 464.
[66] See discussion in Triaille (n 29), 9.

predictions when presented with new data. Use of the information generated, however, is not part of the definition of TDM in the Directive. As the Directive points to how TDM techniques are used for developing new technologies, many ML models are likely to fall within the definition of TDM.[67]

The question in the following is whether the exceptions for TDM in the DSM Directive could apply to training of algorithms in ML. Under Article 4(1), the Member States must provide for 'an exception or limitation' to copyright, database rights and the press publishers' right, 'for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining'. Contrary to the limited exception for scientific research done by public interest research organizations,[68] the exception in Article 4 is available to commercial developers of AI. [69] Some interpretative questions are discussed in the following.

4.1.2.Application of the "three-step test"

Article 7(2) DSM Directive prescribes that the three-step test in Article 5(5) InfoSoc Directive shall apply for the exceptions for TDM. Accordingly, the exception in Article 4 DSM Directive may only apply to special cases which do not conflict with normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the right holder. The first question is whether the public interest in developing services to secure a free press and information integrity may be balanced against the interests of right holders through application of the three-step test.

The CJEU has repeatedly stated that the exceptions from copyright must be interpreted strictly.[70] However, interpretation of the conditions for exception must 'enable the effectiveness of the exception thereby established to be safeguarded and permit observance of the exception's purpose'.[71] Possibly, recent case law under Article 5(1) InfoSoc Directive may indicate a broader weighing of interests with the basis in Article 5(5).[72] The purpose of the exception for temporary digital copying in Article 5(1) Infosoc Directive, is to 'allow and ensure the development and operation of new technologies', and for this purpose to become effective, it must be interpreted in a way that 'safeguards a fair balance between the rights and interests of rights holders and of users of protected works who wish to avail

[67] Rec (18), see also Meys (n 23), 464–465.
[68] Article 3.
[69] Rec (10), (12) and (13). Some public broadcasting institutions can be "cultural heritage institutions" that are eligible for exception under Art 2(3). However, as they are not likely to offer large text work datasets online, they are not further discussed here. The original proposal only included the exception for research purposes in Art 3, and was heavily criticized for not facilitating commercial services protecting public interests, such as fact checking in journalism. See Geiger et al 2018 (n 28), 834, Meys (n 23), 466; rec (18) DSM Directive.
[70] CJEU case C-05/08, *Infopaq I*, para 56, ECLI:EU:C:2009:465; case C-302/10, *Infopaq II*, para 27, ECLI:EU:C:2012:16; case C-360/13, *Meltwater*, para 23, ECLI:EU:C:2014:1195.
[71] CJEU case C-403/08, *Premier League*, para 163, ECLI:EU:C:2011:631.
[72] Taina Pihlajarinne, 'Copyright exceptions and limitations – is the principle of narrow interpretation gradually fading away?', NIR – Nordiskt Immateriellt Rättsskydd, 2020, 117–122, at 121; P Bernt Hugenholtz, 'Flexible Copyright: Can the EU Author's Rights Accommodate Fair Use?', in Ruth L Okediji, *Copyright Law in an Age of Limitations and Exceptions*, CUP, 2017, 275–291, at 286.

themselves of those technologies'.[73] This points to a concrete balancing between the general interest of copyright holders and the narrow interests of users in availing themselves of new technologies in their use of works.

In *Infopaq I* and *Premier League*, the CJEU merely said that the conditions in Article 5(1) InfoSoc Directive must be construed with regard to the three-step test in Article 5(5), but it did not discuss the three-step test separately, perhaps not reading independent normative content into it.[74] In *Meltwater*, however, the CJEU balanced the interests of right holders against the interests of users both under Article 5(1) and under Article 5(5). The Court cited its statements in *Premier League*, pointing to the narrow balancing of interests under Article 5(1).[75] The Court then went on to consider the three-step test in Article 5(5) independently. This should theoretically further narrow the scope of the exception in Article 5(1), as further conditions must be fulfilled for the exception to apply. However, and especially when seen in connection with the more abstract statements about the interpretation of Article 5(5) in *Spiegel Online*, the Court appears to open up for a broader balancing of interests that also includes public interests. First, the *Meltwater* case indicates a more user-oriented approach where the interests of internet users, availing themselves of the right to information in the online media, must be safeguarded.[76] Specifically, they must be able to trust that the publishers of websites have fulfilled their obligation to obtain sufficient consent from the individual right holders, so that it is not necessary for internet users to obtain further authorization.[77] Second, in *Spiegel Online*, the CJEU characterizes the rights of online media users as fundamental rights, enshrined in the EU Charter of Fundamental Rights Article 11.[78] The fair balance to be struck between right holders and users of works is elevated to balancing two fundamental rights: Copyright as enshrined in Article 17(2) of the Charter, on the one hand, and freedom of information on the other.[79] As observed by Pihlajarinne, this development could possibly give more room for societal needs and a less stringent property-right approach, anchored in the three-step test in Article 5(5).[80] While development of AI can clearly serve the public interest in access to information and information integrity, a disconnect exists between developers and the public interest as AI can also be used for other purposes. Under the current copyright law regime, it is not

[73] CJEU case C-403/08, *Premier League*, para 164, ECLI:EU:C:2011:631; case C-360/13, *Meltwater*, para 24, ECLI:EU:C:2014:1195. See also preamble (31) InfoSoc Directive.
[74] CJEU case C-05/08, *Infopaq I*, para 56, ECLI:EU:C:2009:465; CJEU case C-403/08, *Premier League*, para 181, ECLI:EU:C:2011:631. See Annette Kur et al, *European Intellectual Property Law*, second edn, Elgar, Cheltenham, 2019, 384.
[75] CJEU case C-360/13, *Meltwater*, para 24, ECLI:EU:C:2014:1195.
[76] See also CJEU case C-403/08, *Premier League*, para 179, ECLI:EU:C:2011:631; Justine Pila and Paul Torremans, *European Intellectual Property Law*, (2 edn) OUP, 2019, 311, show how the national court in the Meltwater case came to the same conclusion, but based on the author-oriented argument that it would be better to seek a single, higher licence fee from Meltwater than to seek many small fees from end users.
[77] CJEU case C-360/13, *Meltwater*, paras 57–59, ECLI:EU:C:2014:1195.
[78] CJEU in case C-516/17, *Spiegel Online*, paras 54 and 57, ECLI:EU:C:2019:625.
[79] CJEU in case C-516/17, *Spiegel Online*, para 58, ECLI:EU:C:2019:625.
[80] See Pihlajarinne, (n 727), 122; rec (3) DSM Directive; Geiger et al 2018 (n 28), 282.

clear whether the need to control use of AI will be considered when applying exceptions to exclusive rights, or whether such control will be left to other law and regulation.

Secondly, it is not clear how the three-step test can be applied when the same subject matter is protected by several different exclusive rights. The exception for TDM in Article 4 DSM Directive applies to copyright to works and databases as well as the *sui generis* right and the press publishers' right. The three-step test as construed by the CJEU does not apply to the *sui generis* right to databases.[81] For TDM and ML, the exception in Article 4, might answer a need for a better balancing of interests in delineating the scope of the *sui generis* right.[82] However, it is very difficult to see how a balance of interests may be struck with any kind of legal certainty between AI developers also representing public interests on the one side and several holders of different rights based on different purposes on the other. Adding new rights and exceptions does not improve this.

4.1.3.Scope of the exception

The scope of the exception in Article 4 DSM Directive is wide, but its interpretation is not entirely clear. Firstly, the extractions and reproductions may be retained 'for as long as is necessary for the purposes of text and data mining'.[83] The wording indicates that more permanent copying might be included than under the exception for temporary reproductions in Article 5(1) InfoSoc Directive. However, it is not clear to what extent downloading of materials is included, and if so, how long the materials might be retained. While downloading is the practical way to compile a centralized training corpus, it also facilitates further use of the training materials. It may be costly to compile, prepare and annotate a training corpus, and it might be attractive to reuse it to train other algorithms or to license it. Licensing or selling a training corpus will likely go beyond normal exploitation of works and prejudice the interests of database or press publishers' right holders.[84] It is questionable whether a developer may reuse the corpus for training other algorithms.

Secondly, right holders may expressly reserve against use of their works in ML.[85] Recognizing that materials to be used in TDM or in ML are likely to be obtained using automated tools, the reserve must be made by "machine-readable means" for content made available to the public online.[86] All right holders may reserve, including authors of individual works. With the absolute wording of Article 4(3), a reserve by one author will prevent a developer from running the algorithm on that work. As training data are likely to be collected or accessed through intermediaries who might have database or press publishers' rights, these individual reserves may pose challenges to developers in ensuring that

[81] The wording of the relevant Art 8(2) in the Database Directive differs from Art 5(5) InfoSoc Directive.
[82] In this direction Meys (n 23), 469; DG CONNECT, 'Study in support of the evaluation of Directive 96/9/EC on the legal protection of databases', 2018, 25, https://op.europa.eu/en/publication-detail/-/publication/5e9c7a51-597c-11e8-ab41-01aa75ed71a1, accessed 30.11.2020.
[83] Article 4(2).
[84] See Art 5(5) Infosoc Directive and the discussion in sections 3.2 and 3.3.
[85] Art 4(3) DSM Directive.
[86] Art 4(3) and rec (18).

training data are representative and generalizable. For the reasons discussed above in section 3.1, the better solution might have been to make the exception mandatory for copyright in individual works, but allowing database right holders and press publishers a right to reserve.[87]

4.1.4.The condition of lawful access

It follows from Article 4(1) that the user must have lawful access to materials to benefit from the exception. Lawful access can be based on a licence, and in that case the new exception gives publishers and database owners an incentive to distinguish prices for licensing for TDM or ML purposes and other uses.[88] The exception also applies to content that has been made available to the public online without reservation for TDM.[89]

A final question is whether "lawful access" should be construed to mean the same as "lawful use" in Article 5(1) InfoSoc Directive. Under that article, a use is lawful if it is either authorised by the right holder or falls outside the scope of the exclusive right of the author.[90] The CJEU has applied the exception in Article 5(1) to services that have as their output excerpts of the works so small that they do not reproduce the 'expression of the intellectual creation of the author'.[91] While this could be as little as eleven consecutive words, the problem with both TDM and ML and training algorithms is that while the techniques may make (temporary) copies of whole works, the output of the techniques does not reproduce any part of the works. The direct output of training an algorithm is an enhanced and trained algorithm. Using a trained algorithm in an AI-assisted service for fact checking might also not reproduce any part of any works. Hence, for the exception in Article 4(3) to have practical relevance for TDM and ML, the condition of lawful access should be construed narrowly, thus not including the output of the technique.

4.2. The exception for temporary reproductions, Article 5(1) Infosoc Directive

The exception for TDM in the DSM Directive entails a presumption not only that TDM (and ML) techniques infringe copyright, but also that the exception for temporary digital reproductions in Article 5(1) InfoSoc Directive is insufficient to cover TDM techniques and ML.[92] While discussion of the scope of the exception is now less potent, it remains open whether the exception in Article 5(1) could supplement the exception for TDM or benefit aspects of ML that are not covered by the TDM exception.[93]

---

[87] Similar Lemley and Casey (n 36) 130 with regard to US law.
[88] Geiger et al 2018 (n 28), 83.
[89] Preamble (18) DSM Directive.
[90] See CJEU case C-302/10, *Infopaq II*, para 42, ECLI:EU:C:2012:16, and case C-403/08, *Premier League*, para 168, ECLI:EU:C:2011:631 and preamble (33) InfoSoc Directive.
[91] CJEU case C-05/08, *Infopaq I*, para 39, ECLI:EU:C:2009:465; InfoSoc Directive Art 3(1).
[92] See rec (18) DSM Directive. See Meys (n 23), 465.
[93] Rec (9) DSM Directive.

A concrete analysis of the ML model is necessary to determine whether the exception in Article 5(1) could apply.[94] Only selected questions are discussed here. First, as the algorithm is running on training data, the computer will make temporary copies of the materials in its cache memory. Mostly, these copies are automatically deleted when the cache memory is full. [95] Only the information required to optimise the algorithm is stored in a separate file, later to become the trained algorithm. It seems likely that cache memory copies are considered transient, temporary, and to be an integral and essential part of a technological process, regardless of whether on-screen copies are made that would be readable by humans.[96] It will not prevent application of the exception that the process contains multiple steps where a copy is made,[97] as when the algorithm runs back and forth on the materials, or that human intervention is involved.[98]

Apart from compiling training data,[99] the exception in Article 5(1) could cover the individual steps of training an algorithm. The question of its application is conceptual, but can be tied to the condition that the act of reproduction must not have "independent economic significance." Training an algorithm that will present articles on a personalized front page, or perform a fact check will have as its output a trained algorithm which is not likely to contain a reproduction of the training materials. Hence, the overarching purpose of ML can be argued to fall outside the scope of the right of reproduction.[100] However, ML is based on the algorithm "reading", thereby intrinsically copying, huge amounts of works. In *Infopaq II*, the CJEU found that temporary copies can have "economic significance" if the copies can be exploited for gain, or if temporary copies transform the object of lawful use or if they facilitate other use.[101] The *Infopaq* cases concerned a service that included temporarily copying works, but where the output of the service was to present clients with short – and lawful – excerpts of the works. Applied to development of AI through ML, a differentiation should be drawn between training the algorithm and performing a service based on a trained algorithm. ML models based on static training, that is, where training is completed before the algorithm is used on new materials, may be seen as separate acts of exploitation of the works in the training corpus. As this exploitation is not

[94] See Margoni (n 14), 19.
[95] Automatic deletion would be necessary for Art 5(1) to apply, see CJEU case C-05/08, *Infopaq I*, paras 37 and 62, ECLI:EU:C:2009:465.
[96] In this direction CJEU case C-360/13, *Meltwater*, ECLI:EU:C:2014:1195.
[97] See CJEU case C-05/08, *Infopaq I*, para 65, ECLI:EU:C:2009:465.
[98] CJEU case C-302/10, *Infopaq II*, para 22, ECLI:EU:C:2012:16.
[99] Downloading or uploading a training data set will likely require permanent copying in the sense of Art 5(1) InfoSoc Directive. See also Triaille (n 29), 48.
[100] See Strowel (n 35), 213, 225 and Maurizio Borghi and Stavroula Karapapa, *Copyright and Mass Digitization*. OUP, 2013, 58–60 and ch 2, DOI:10.1093/acprof:oso/9780199664559.003.0003.
[101] CJEU case C-302/10, *Infopaq II*, paras 52 and 53, ECLI:EU:C:2012:16; case C-403/08, *Premier League*, para 175; Ole-Andreas Rognstad, Opphavsrett, 2. ed. Oslo, 2019, 192.

accessory to otherwise lawful use, it would likely require the authorization of the authors.[102] This would also be the reason for excluding TDM from the exception under Article 5(1).[103]

For services based on trained algorithms, such as performing a fact check, or bulk reading to sort out information to consider reporting, the output of the service could be the primary act, to which the temporary copying involved in machine "reading" can be accessory. If the output of the service is lawful by consent or because it falls outside the scope of copyright, the conditions for exception in Article 5(1) can be considered in the same way as in the *Infopaq*, *Meltwater* and *Spiegel Online* cases. Following this reasoning, a service including an algorithm that continuously learns – a dynamic learning model – could benefit from the exception if the output of the service, such as display of articles on a personalized front page, would be considered lawful.

5. Conclusions: A missed opportunity for a balanced approach to machine learning and copyright law

Under the current EU legal regime, a differentiation is needed between operating online services and developing services. Services that rely on AI, and inherently trained algorithms, may operate on copyright-protected works as long as their output does not contain infringing reproductions of those works. Case law from the CJEU indicates that these services can be considered under Article 5(1) InfoSoc Directive. Looking only at the application of AI in online services, copyright seems fairly well adapted to striking a balance between the interests of right holders, both individual authors and owners of collections of works, and the interests of service providers. Although services employ AI, the situation does not significantly differ from other use of protected works in a digital market.

The development of new AI systems by off-line machine learning techniques, however, is dependent on consent from right holders in materials used to train algorithms, regardless of their output. In this regard, the DSM Directive represents a missed opportunity to balance the interests of right holders and users in relation to big data-related innovation. Although the exception for TDM in Article 4 DSM Directive likely covers ML and training of algorithms, the EU lawmaker has continued on its path of expanding exclusive rights towards any new uses of content. It is questionable whether the current regime will further the ambition for the EU to become a global leader in innovation in the data

[102] See also Ducati and Strowel, (n 32), 660–661. Same conclusion, but with some uncertainty, Geiger et al 2018 (n 28), 821–822.
[103] Irini Stamatoudi, "Text and Data Mining", in Irini Stamatoudi (ed.), (n 54), 251–282.

economy, for which development of AI is essential, [104] especially compared to the USA, where ML as a main rule is considered fair use of copyrighted works.[105]

First, problems with expanding copyright in individual works with regard to uses not related to the creative expression of a work have not been resolved for development of AI. To allow individual authors to oppose use of their works to train AI when those works are included in a collection or database, cannot be explained by the economic incentive system of copyright.[106] If the developer has lawful access to the works no market failure needs to be resolved by extending the exclusive rights of individual authors to the downstream market for AI systems.[107] It is also highly unlikely that individual authors' right to reserve would result in payment.[108] To train an algorithm, it is necessary to have access to so many works that individual payment is both practically impossible as well as prohibitively expensive. For an AI system in operation – that is, the service employing a trained algorithm – it is relevant to ask whether it reproduces such content from works it has "read" or been trained upon, so that the creative expression of the author is infringed. Using copyright-protected works to train algorithms is a use of those works as information, where the value of access to the works may be ascribed to the collection and whether the collection is representative, complete and generalizable, features that are unrelated to the individual works. This "added" value is related to the collection, and as such forms part of the incentive system for database rights or publishers' rights.[109] Economic arguments therefore support rules that strike a fair balance between database right holders and publishers and developers of AI when using the collection to train their algorithms, but with regard to the authors of individual works, the exception would have been more effective if mandatory.

Second, it is questionable whether the balance struck with the new exception for TDM – if applicable to ML – is fair. As automated processing of huge amounts of data becomes easier and has infinite future use cases, the income potential from big data-related services is likely to exponentially increase the value of data and content collections. Reluctance by publishers and database owners can be expected in terms of opening their repositories to be included in training corpora for commercial ML, and pricing may be prohibitive for start-ups and innovators.[110] It would have been a possibility to open

[104] See White Paper on Artificial Intelligence – A European approach to excellence and trust, COM(2020) 65 final, https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en, accessed 4 December 2020. See also Christian Handke et al, 'Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research', in B Schmidt and M Dobreva (eds), 'New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust', Proceedings of the 19th International Conference on Electronic Publishing, 2015, p. 120-130, doi:10.3233/978-1-61499-562-3-120.

[105] See, eg, Esposti et al (n 9), Lemley and Casey (n 36) and James Grimmelmann, 'Copyright for Literate Robots', Iowa Law Review, (101), 2016, 657–682.

[106] Poort, (n 37), 330.

[107] *Ibid.*

[108] Margoni (n 14),. 20.

[109] Poort, (n 37), 330.

[110] See also Geiger et al 2020 (n 33), discussing how the right to opt out could discriminate against small businesses and start-ups in access to data.

up for compulsory licensing for ML to develop services serving the public interest, such as journalistic tools for fact checking, or securing information integrity, to avoid so-called "fake news"-related issues. However, it is difficult to use one application of an AI service as justification for a mandatory exception or compulsory licence for the process of ML. As discussed above, an algorithm after training might be used to develop multiple services serving even opposing objectives, so the justification for ML will only cover ML for the purpose of developing that one service. Furthermore, a balancing of interests based on a public interest justification for the application of the service must cover both the interests of collective right holders in training materials for the algorithm in the development phase, as well as the interests of individual right holders in works used in running the (trained) service. To what extent these interests are interfered with will depend on the individual characteristics of the ML process and the service. Thus, it is difficult to make out a clear scope for a public interest justification that provides legal certainty both for right holders and for AI developers.

Finally, regulation of TDM and ML in EU copyright law, may be seen as a further step in developing copyright law as a system for balancing access to and control with uses of online available content in general, gradually moving away from the traditional functionality of copyright law, namely, as an incentive for creative efforts. Exclusive rights have been gradually expanded, but focus within copyright law remains narrow, limited to the economic interests of right holders. It is questionable whether the public interests concerned with development of AI, and training of algorithms can be sufficiently safeguarded within this system. Overlapping rights and many and complicated exceptions, create legal uncertainty.[111] The option to reserve against ML activities and the narrow focus on economic interests may pose risks to data transparency and integrity, as well as making it more difficult to ensure that datasets are complete, representative and generalizable and free of human bias that might pose a risk for secure use of AI. It would probably better serve the public interest – in particular to promote transparency – to base control of the use of AI on the use of service online or in the market, but not through privately enforced rights at the development stage.

[111] Geiger et al 2018 (n 28), 836.