A DEEPER LOOK INTO THE EU TEXT AND DATA MINING EXCEPTIONS

HARMONISATION, DATA OWNERSHIP, AND THE FUTURE OF TECHNOLOGY

CREATe Working Paper 2021/7

THOMAS MARGONI
MARTIN KRETSCHMER

CREATe

A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology

Thomas Margoni & Martin Kretschmer*

Abstract

There is global attention on new data analytic methods. Data scraping (a typical first step for advanced data analytics), text and data mining (TDM, the extraction of knowledge from data) and machine learning (ML, often also simply referred to as Artificial Intelligence or AI) are seen as critical technologies. The legal issues involved in the regulation of data range from privacy and data protection (such as the GDPR) to proprietary approaches (such as copyright, database rights, or proposed new rights in data themselves).

This paper focusses on one specific intervention, the introduction of two exceptions for text and data mining in the Directive on Copyright in the Digital Single Market (CDSM). Art. 3 is a mandatory exception for text and data mining (TDM) for the purposes of scientific research; Art. 4 permits text and data mining by anyone but with rightsholders able to "contract-out" (Art. 4), for example preventing TDM use of publicly available online content by "machine-readable means".

We trace the context of using the lever of copyright law to enable emerging technologies and support innovation. Within the EU copyright intervention, elements that may underpin a transparent legal framework for AI are identified, such as the possibility of retention of (permanent) copies for further verification. On the other hand, we identify several pitfalls, including an excessively broad definition of TDM which makes the entire field of data-driven AI development dependent on an exception. We analyse the implications of limiting the scope of the exceptions to the right of reproduction (which leaves the communication of research results in a grey zone). We also argue that the limitation of Art. 3 to certain beneficiaries remains problematic; and that the requirement of lawful access is difficult to operationalize.

In conclusion, we argue that there should be no need for a TDM exception for the act of extracting informational value from protected works. The EU's CDSM provisions paradoxically may favour the development of biased AI systems due to price and accessibility conditions for accessing training data that offer the wrong incentives. We also identify some old and new areas of the EU acquis which will play a crucial role in the future relationship of EU copyright law with technological innovation.

* Thomas Margoni is Research Professor of Intellectual Property Law, Centre for IT&IP Law (CiTiP), Faculty of Law, University of Leuven (KU Leuven) and Fellow at CREATe. Martin Kretschmer is Professor of Intellectual Property Law, University of Glasgow and Director of CREATe (UK Copyright & Creative Economy Centre). The research has been supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626870626 ("reCreating Europe: Rethinking digital copyright law for a culturally diverse, accessible, creative Europe"). The paper draws on earlier work presented at the European Policy for IP Association, Berlin (07/09/2018) and the Global Congress on IP & Public Interest, Washington (27/09/2018) that became an early reference point in the debate about a distinct EU approach to new data analytic technologies. The material is archived here: https://www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/.

Table of contents

I. Introduction

The Directive on Copyright in the Digital Single Market (CDSM)[1] incorporates a number of provisions (32 Articles and 86 Recitals) intended to modernise EU copyright law and to make it "fit for the digital age".[2] The Directive's reach is impressive: it covers exceptions and limitations (Arts. 3-6), out-of-commerce-works and licensing practices (Arts. 8-12); the reproduction of works of visual art in the public domain (Art. 14), and a whole chapter dedicated to the fair remuneration of authors and performers (Title IV, Ch. 3).[3] Some of these provisions have attracted extraordinary scholarly and media attention and were object of a lively debate in the light of their controversial nature (e.g. the changes in platform liability for copyright purposes contained in Art. 17[4]) or because they introduced a new right within the already variegate EU neighbouring right landscape (e.g. the protection for press publications contained in Art. 15[5]).

Far less attention, at least during the drafting phase, have attracted the provisions contained in Arts. 3 and 4 of the Directive which are dedicated to "Text and data mining" (TDM)[6], although Art. 4 may be seen as a "last minute addition".[7] The goal of Art. 3 is to introduce a mandatory exception under EU copyright law which exempts acts of reproduction (for copyright subject matter) and

[1] 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 (Text with EEA relevance).
[2] https://ec.europa.eu/digital-single-market/en/modernisation-eu-copyright-rules
[3] For a thorough analysis of the Directive see Dusollier S. (2020), The 2019 Directive on Copyright in the Digital Single Market: Some progress, a few bad choices, and an overall failed ambition, Common Market Law Review, 57(4), pp. 979-1030; Quintais J.P. (2020), The New Copyright in the Digital Single Market Directive: A Critical Look, European Intellectual Property Review EIPR 2020(1), pp. 28-41.
[4] European Copyright Society (ECS), General Opinion on the EU Copyright Reform Package (24 January 2017), p. 7, available at: https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf.
[5] Höppner T., Kretschmer M., Xalabarder R. (2017) CREATe Public Lectures on the Proposed EU Right for Press Publishers, European Intellectual Property Review 39(10) pp. 607-622, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3050575.
[6] Although see Geiger et al. (2018) The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects, Policy Department for Citizens' Rights and Constitutional Affairs, Directorate General for Internal Policies of the Union PE 604.941- February 2018, available at: https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf; Christophe Geiger, et al., Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?, 49 IIC 814, 814-844 (2018); Rossana Ducato & Alain Strowel, Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility", 50 IIC 649 (2019); Gonzalez Otero B., Machine Learning Models under the copyright microscope: is EU Copyright fit for purpose?, Forthcoming in: Nordic Law Review (NLR), Max Planck Institute for Innovation & Competition Research Paper No. 21-02; Eleonora Rosati, An EU Text and Data Mining Exception for the Few: Would it Make Sense?, 13 J. INTELL. PROP. L. & PRAC. 429, 429-430 (2018); Andres Guadamuz & Diane Cabell, Data Mining in UK Higher Education Institutions: Law and Policy, 4 QUEEN MARY J. INTELL. PROP. 3, 3-29 (2014).
[7] See Hugenholtz, Kluwercopyrightblog; Geiger et al., Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU.

extraction (for the Sui Generis Database Right, SGDR) made by research organisations and cultural heritage institutions (hereinafter research and cultural organisations) in order to carry out text and data mining for the purposes of scientific research. Art. 4 mirrors Art. 3 with one major (and a few minor[8]) differences: it is available to any type of beneficiaries for any type of use, but can be expressly reserved by rightholders – in other words it may be object of "opt-out" or "contract-out".

This paper focuses on these two new additions to the list of EU copyright exceptions and argues that their formulation, although underpinned by a strategic innovation policy goal, is conceptually wrong, theoretically flawed and normatively unambitious. Even worse, by employing an overlay broad definition of text and data mining, the provisions under analysis regulate by way of a narrow exception not only TDM but all forms of modern data-driven digital analytics that rely on "training" on data. This is a vast field that includes most forms of modern Artificial Intelligence (AI) applications relying on machine learning in areas as varied as Natural Language Processing (NLP), image recognition and classification, content filtering and robotics (hereinafter generally referred to as AI).[9]

The paper further argues that the implications of accepting the principle that EU AI can be developed only thanks to an *exception* or after securing proper authorisation reach far beyond the rationale and the evidence considered during the drafting phase of the new Directive.[10] The general regulation of technology, especially of a technology as pervasive as AI, exceeds the goals of copyright law. This is commonly accepted in AI policy and legislative venues where

[8] For reasons not fully apparent in the Directive's Preamble, Art. 4 explicitly includes in its scope the reproduction and the adaptation rights in computer programmes, while Art. 3 only refers to the reproduction rights contained in Directives 2001/29 (InfoSoc) and 1996/9 (Databases). The reference in Art. 3 to Directive 2001/29 should be sufficient to cover also reproductions of computer programmes (but arguably not adaptations) in the light of the hermeneutic principle that special law derogates general law, which implies that when special law does not derogate then general law applies. In the EU *Acquis Communautaire* (the *Acquis*) The Software Directive is considered *lex specialis* with regards to the general InfoSoc Directive (e.g. ECLI:EU:C:2012:407, Case C-128/11 UsedSoft, 51, 56), therefore the reference of Art. 3 CDSM to the general right of reproduction *ex* Art. 2 InfoSoc also covers the right of reproduction contained in the (special) Software Directive. An *a contrario* argument based on the explicit inclusion of Software in Art. 4 would not comply with such general theory rule. Other differences relate to the wording employed in relation to the possibility to retain copies for verification (Art. 3) or for text and data mining (Art. 4). There does not seem to be an equivalent faculty for rightholders to apply integrity measures in Art. 4.

[9] Margoni, T. (2021) Computational Legal methods: Text and Data Mining in Intellectual Property Research. Handbook of Intellectual Property Research Lenses, Methods, and Perspectives Publisher: Oxford University Press; Drexl, et al., Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective, Version 1.0, October 2019.

[10] Although some early warnings were raised; Geiger et al., The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects, Policy Department for Citizens' Rights and Constitutional Affairs PE 604.941- February 2018; Margoni, T., Kretschmer, M. (2018) Data mining: why the EU's proposed copyright measures get it wrong. Publisher: The Conversation.

the role of copyright is often seen as secondary. However, the creation of property rights in data, i.e., in the building blocks necessary for erecting the complex structure called AI, is equivalent to the implementation of a system of authorisations that AI developers need to secure before engaging in their product development. Allocating the right to authorise or forbid the use of traditionally unprotected mere facts and data essential for AI development to certain market actors creates not only a market structure but also a system of social and moral values within which this technology will be compelled to evolve. In other words, by devising the rules that regulate access to a certain technology and by allocating ownership in the elements necessary to develop it, we are shaping that technology and its impact on society for the years to come.[11] These rules, today, in the EU, clearly state that firms, governments, citizens, journalists and anyone else who is not a research and cultural organisation acting for research purposes have to obtain a specific authorisation from rightsholders to develop AI. Outside the EU they do not. What this means for cultural and innovation policies, regulatory competition and the future of democracy is a complex question that far exceeds the scope of this article. However, it can be reasonably argued that the EU AI sector is put at a considerable disadvantage, if for nothing else, the much higher costs that AI development has in the EU due to the need to negotiate licences over vast amounts of works needed as input data.[12] Another important aspect relates to the type and quality of data available for AI training, since it is at least arguable that, unable to compete with dominant AI players, smaller firms or new market entrants may find it economically attractive to train their algorithms on "cheaper", which often means older, less accurate or biased, data, leading to the possible development of second class AI applications for those who cannot afford the costs of first class AI, thereby favouring algorithmic discrimination and inequality.[13]

In summary, the paper claims that a narrowly framed EU copyright exception may have become the formal recognition that in the digital environment EU copyright has achieved such an unprecedented hegemonic role in regulating knowledge production and circulation that it covers not only original expressions, as commonly accepted in copyright law and theory, but also

[11] Samuelson P. (2020) Regulating Technology Through Copyright Law: A Comparative Perspective. 42 European Intellectual Property Review; Benkler Y., The Role of Technology in Political Economy, LPE Blog July 2018, available at: https://lpeblog.org/2018/07/25/the-role-of-technology-in-political-economy-part-1/.

[12] Senftleben M., et al., Ensuring the Visibility and Accessibility of European Creative Content on the World Market: The Need for Copyright Data Improvement in the Light of New Technologies, (February 12, 2021). Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3785272.

[13] In this sense and with reference to the US legal system see the detailed analysis of Levendowski A., How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem, 93 Wash. L. Rev. 579 (2018).

mere facts and data.[14] This is the likely effect of the insertion of Arts. 3 and 4 into the current *Acquis Communautaire* characterised among other things by a rather low originality standard (even 11 consecutive words[15] and foldable bicycles[16]); a broad right of reproduction (covering copies in the cache memory of computers and satellite decoders as well as the transfer of ink onto different supports[17]); a right protecting non original databases (Art. 7 Database Directive); a closed non-mandatory list of exceptions that must be interpreted narrowly (Art. 5 InfoSoc) and which, at the same time, represents all the limits that EU copyright law's exclusive rights can be subjected to, including those connected to fundamental rights (*Pelham* case[18]). This stratification of rules enacted in different stages of the process of EU copyright harmonisation has the combined effect of absorbing a great deal of previously unprotected knowledge, such as mere facts and data, into low-original (or non-original in the case of SGDR) works protected against most forms of indirect, incidental and transient reproductions. In other words, a decisive, albeit disguised, enclosure of existing mere facts and data.

II. Reclassifying Arts. 3&4 CDSM: a functional perspective

The Directive defines TDM as "any automated analytical technique aiming to analyse text and data in digital form to generate information such as patterns, trends and correlations" (Art. 2(2)) as well as "the automated computational analysis of information in digital form, such as text, sounds, images or data" enabled by new technologies (Recital 8). This is a very broad definition which aptly identifies the potential of a tool able to analyse autonomously or semi-autonomously vast amounts of data. As a matter of fact, this definition reaches far beyond the taxonomy employed and comfortably captures most activities where digital technologies are utilised to analyse information and extract meaning. Nowadays, the dominant approach to perform this

[14] The paper does not discuss the related but systematically distinct issue of property rights in generated data; for an insightful analysis see Hugenholtz B., Against data property; Kerber W., A New (Intellectual) Property Right for Non-Personal Data? An Economic Analysis, GRUR Int, 11/2016, 989-999; Benterle F., Data Ownership in the Data Economy: A European Dilemma, In: Synodinou TE., Jougleux P., Markou C., Prastitou T. (eds) EU Internet Law in the Digital Era. Springer. The paper is also not directly concerned with whether AI outputs may or should be protected by copyright, see Hilty et al., Intellectual Property Justifications for Artificial Intelligence, Max Planck Institute for Innovation and Competition Research Paper Series; IViR & JIIP, Trends and Developments in Artificial Intelligence, Luxembourg, 2020; Kop M., AI & Intellectual Property: Towards an Articulated Public Domain, 28 TEX. INTELL. PROP. L.J. 297 (2020). Finally, the paper does not cover the issue of personal data, see e.g. Guarda P., Free data? open science in the age of personal data protection, Rooks by J. (Ed), Research Handbook on Intellectual Property and Technology Transfer, Elgar, 2020.

[15] Case C-5/08, *Infopaq I*, ECLI:EU:C:2009:465 and Case C-302/10.

[16] Case Case C-833/18, *Brompton Bicycle*, ECLI:EU:C:2020:461.

[17] Hugenholtz B., Kretschmer M., Reconstructing Rights: Project Synthesis and Recommendations, in Hugenholtz (ed.) Copyright Reconstructed: Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change, Kluwer Law International, 2018.

[18] ECLI:EU:C:2019:624 Case C-476/17, of 29 July 2019 Pelham v Hütter.

task is called Machine Learning (ML) in its various manifestations and developments.[19] Therefore, it can be argued that the definition employed in the CDSM is future-proof in the sense that it covers – and thus regulates – most areas of ML now known or developed in the future.

However, when such a broad definition is inserted into a narrowly construed exception, as the one under analysis, the result may not be that of opening up new technological and cultural practices as arguably was the original intention of the drafters, but rather the opposite. Not only TDM *stricto sensu* has been limited to research uses by research and cultural organisations, but virtually any automated technique that analyses information in digital form is captured under the narrow boundaries of the current formulation of Arts. 3 and 4. This certainly includes most modern, data-driven forms of AI, such as traditional machine learning and more advanced forms of deep learning and neural network structures. The policy reasons justifying the allocation of the power to authorise these cutting-edge technologies to upstream players in the database and content markets are far from clear. The interventions have the potential for anti-competitive effects and most importantly have not been addressed from a EU policy perspective in the explanatory documents of the CDSM.[20] This may suggest that – whereas the interests of the publishing industry in licensing their databases for TDM purposes as well as the needs of the research community to access them were duly considered in the Impact Assessment[21] – the deeper technological, innovation and cultural policy implications of the proposed legislation were not fully unpacked, despite calls in this sense.[22]

This unsighted approach to law making may favour the development of opaque AI systems or AI "black boxes"[23], an expression referring to a type of automated decision-making tool (e.g. AI)

[19] For common usage, see https://en.wikipedia.org/wiki/Machine_learning (visited 1 July 2021): "Machine learning (ML) is the study of computer algorithms that improve automatically through experience and by the use of data. It is seen as a part of artificial intelligence".
[20] The Impact Assessment discusses how exceptions may affect researchers and rightholders as well as the social and fundamental rights impact of certain provisions (although the latter two elements appear underdeveloped in comparison to the former), but in general does not consider broader industrial, innovation and cultural policy issues, see Commission Staff Working Document on the Modernisation of EU Copyright Rules Brussels, 14.9.2016 SWD(2016) 301 final PART 1/3, Sec. 4.3
[21] Id, at p. 114.
[22] ECS, Opinion on European Commission Proposals for Reform of Copyright in the EU, 2017, p.5 available at https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf; Margoni, Thomas; Kretschmer, Martin (2018) Data mining: why the EU's proposed copyright measures get it wrong, The Conversation, May 24, 2018; Geiger et al. (2018) The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects, Policy Department for Citizens' Rights and Constitutional Affairs, PE 604.941- February 2018.
[23] Levendowski A., How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem, 93 Wash. L. Rev. 579 (2018).

which makes decisions in ways that are not intelligible or transparent to humans.[24] Modern data-driven AI systems could be seen as extremely complex statistical machines. The way in which they develop a certain understanding of reality cannot be understood based on the cognitive abilities of biological beings, such as humans. The classical example in the literature relates (unsurprisingly) to cats. AI systems can become extremely accurate at distinguishing images of cats from images of, say, dogs. However, data-driven AI comprehension of what "cat" is or means cannot be compared to that of humans. This may have been (and to some extent still is) the case with "traditional" knowledge-driven approaches to AI, where the AI is "taught" to classify a cat following human categories, i.e., it is a mammal, sub-category feline, it has four paws, whiskers, tail, etc. This is closer to how human learning operates and may well be applied to certain fields of AI where rules, attributes and conditions follow a formal linear logic, such as certain attempts to encode contractual conditions in automated decision-making languages.

Machine learning, however, operates differently. It is based on highly complex statistical abstractions supported by enormous databases, e.g., millions or billions of pictures of cats and dogs which are used as training material by the learning algorithms.[25] Once the training is complete, a trained model, i.e. a file containing an abstraction of the data that the learning algorithm has found useful to accurately distinguish between cats and dogs will be retained. This file forms the "memory" used by the AI to analyse new and unknown data and to adapt its behaviour to this new reality, e.g., to establish whether a new, unseen picture is a cat or dog. The original dataset used as training material (the billions of pictures of cats and dogs) at this point is no longer necessary for the AI system to operate, only the trained model is.[26] However, humans are not capable of a proper understanding of this abstract statistical ML memory. A prospective user, a firm or a public body may know what data go in (a new picture, personal financial details, health records) and what data come out (it's a cat, credit or health related requests accepted or refused), but it is not possible for human observers to *understand* why.[27]

This reflects the characteristic of any ML based AI system to be a black box, but there are ways to mitigate this situation. One option to get closer to "understand" how the learning algorithm has reached certain decisions is access to the original training data. This would not necessarily

[24] Zittrain J., Intellectual Debt: With Great Power Comes Great Ignorance, 24 July 2019, https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c.
[25] Kaplan et al., Scaling Laws for Neural Language Models, 2020, arXiv:2001.08361 [cs.LG].
[26] Margoni T. (2018) Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?. AIDA; 2018(1), pp. 1-26.
[27] Zittrain J. Intellectual Debt: With Great Power Comes Great Ignorance, 24 July 2019, https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c.

explain the complex statistical process leading to those decisions but would make it possible to scrutinise the original training data for mistakes, omissions or bias and to replicate or reverse engineer those decisions and therefore to ensure a transparent, accountable and possibly unbiased decision-making processes.[28]

The fitness of a modern copyright system in this complex technological scenario needs to be assessed in the light of its ability to explicate a balancing function in this fast developing environment. Ensuring the undistorted availability of training data in order to produce more accurate results (efficiency), as well as securing their permanent accessibility in order to ensure that the produced results are in line with the system of fundamental rights and values embedded in our legal orders (fairness) will be key indicators of the fulfilment of copyright's role in emerging technology.

It may be argued that under the misleading label Text and data mining (TDM) what has been regulated at the EU level in Arts. 3 and 4 goes far beyond a mere copyright exception. In fact, it should be reclassified as the legal regulation of AI via the allocation of property rights in its building blocks, or in other words, as a *property-right approach to the regulation of AI*. This is possibly one of the most crucial legislative developments in the field of law and new technologies, one which will have a profound impact on the lives of EU citizens and on their effective enjoyment of rights and liberties, including those safeguarded in the EU Charter of Fundamental Rights.

II.1. Creating new knowledge from existing data: legal *versus* technological approaches

It has been shown that the global research community generates over 1.5 million new scholarly articles per annum[29] or approximatively one new paper every 30 seconds.[30] The same scientific community that has produced this knowledge is likely unable to maintain an adequate level of assimilation and understanding of it. This depicts a highly inefficient system where resources are spent to duplicate knowledge that probably already exists but remains undiscovered. Data seem to confirm this situation by showing that some 90% of all published scientific papers are never cited, whereas 50% of them are never red by anyone other than their authors, referees and

[28] Levendowski A., How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem, 93 Wash. L. Rev. 579 (2018).
[29] Ware & Mabe (2009) The STM report – An overview of scientific and scholarly publishing, STM, Oxford, 7, available at: https://www.stm-assoc.org/2009_10_13_MWC_STM_Report.pdf. See generally OpenMinTeD.eu for examples of how TDM techniques can be used.
[30] Spangler et al. (2014) Automated Hypothesis Generation based on Mining Scientific Literature, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, 1877, available at: scholar.harvard.edu/files/alacoste/files/p1877-spangler.pdf.

journal editors.[31] From a technical point of view, TDM could easily fix this problem by reading, processing analysing and classifying this wealth of knowledge in ways not yet even imagined. The new TDM exception will ensure that this will be permitted under certain conditions, chiefly when performed by research and cultural organisations for research purposes or when not reserved by rightsholders, something not completely clear under previous law.[32]

But there are numerous other examples that demonstrate how TDM could significantly improve the quantity, quality and speed of technological innovation, economic growth and social welfare which do not find proper recognition within the scope of the EU TDM exceptions. As a mere illustration, it has been attested that in the EU in fields such as linguistics and Natural Language Processing (NLP), the ability to develop automated translation tools is currently limited mostly to the official documents produced by the European Union,[33] which are openly available and reusable.[34] Augmenting the availability of original data sources beyond official texts of EU bodies (legal language cannot really be said to reflect how usually people talk) to include all information available on the Internet would open an entirely new set of opportunities. This would also put EU based firms, especially small and medium-sized enterprises (SMEs) and start-ups, on a level playing field with multinational companies, such as Google, Facebook, Amazon, Microsoft and Twitter which can benefit from copyright laws that permit *them* to engage in this type of activities without prior authorisation, therefore significantly reducing the cost of AI development. Another example that shows the problematic and likely unintended consequences ensuing from the formulation of Arts. 3 and 4 CDSM is the exclusion from their ambit of journalistic enquiry and the possibility to text-and-data mine online archives to verify the accuracy of certain facts and thus to combat fake news.[35] In this respect, it can be observed how the EC Impact Assessment focused its analysis on the needs of the publishing industry on the one hand and of academic and commercial research on the other. This is undoubtedly a very important aspect that needed to be addressed. However, this property-based approach to AI regulation failed to identify the deeper implications for society, the economy and democracy.

[31] Lokman (2007) The rise and rise of citation analysis, Physics World, 20(1), 32-36.
[32] Triaille et al., Study on the legal framework of text and data mining (TDM), European Union, 2014, p. 41, available at: https://op.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en.
[33] OpenMinTeD, TDM stories: How Zalando links languages with TDM, available at: http://openminted.eu/tdm-stories-zalando-links-languages-tdm/ (last accessed March 2018).
[34] Art. 4 of the Commission Decision of 12 December 2011 on the reuse of Commission documents (2011/833/EU).
[35] OpenMinTeD, TDM stories: A Text & Data Miner Talks About Analysing The Recent Past, available at: http://openminted.eu/tdm-stories-text-data-miner-talks-analysing-recent-past/.

There is an array of activities that from an economic and moral point of view seem at least as deserving as research conducted by research and cultural organisations, which are nevertheless excluded from the ambit of the EU TDM exception (or which remain in a sort of undefined status which depends on whether rightholders will reserve their use). Due to the specific characteristics of the *Acquis,* and particularly the right of reproduction, these activities are captured under the exclusive prerogatives of rightholders and thus cannot be lawfully performed.

II.2. The "exceptionalism" of EU copyright law and the right of reproduction

EU law defines reproductions as any "direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part" by Art. 2 of Directive 2001/29/EC (InfoSoc Directive)[36]. As for most acts performed digitally, to "text-and-data-mine" information it is usually necessary to make (at least temporary or indirect or transient) copies, that is to say, to reproduce the original material in a way that triggers Art. 2.

This paper agrees with propositions already formulated in the literature that in a properly designed copyright framework there should be no need for a TDM exception, as the extraction of factual information from protected content is external to copyright's remit.[37] Support for this thesis can be found in internationally recognised principles such as the idea-expression and fact-expression dichotomy, that is to say in the postulate that copyright protects original expressions, whereas ideas, principles, procedures, facts and data as such are not protected.

At the EU level there is no explicit general statutory recognition of the idea-expression doctrine, however, it can be found in the Software Directive (Recital 11 and Arts. 1 and 5.3) with a wording that gives away a certain universal ambition. The fact-expression doctrine may be found in Recital 45 of the Database Directive and in Recital 9 of the CDSM Directive. Additionally, the case law of the Court of Justice of the European Union (CJEU) has restated these doctrines under EU law, both by direct confirmation of its operativity,[38] as well as by identifying as a major canon of interpretation and integration of EU copyright law the international legislative framework which includes the TRIPs Agreements and the WCT both containing an explicit recognition of

[36] Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society.

[37] E.g. Flynn et al. (2020) Implementing user rights for research in the field of artificial intelligence: a call for international action. European Intellectual Property Review, 42(7), pp. 393-398; Matthew Sag (2019) The New Legal Landscape for Text Mining and Machine Learning, 66 J. of the Copyright Soc'y of the USA, 3, 9-19; Carys J. Craig (2017) Globalizing User Rights-Talk: On Copyright Limits and Rhetorical Risks, Am. U. Int'l L. Rev. Vol 33(1).

[38] E.g. Case C-833/18, of 11 June 2020, *Brompton Bicycle*, ECLI:EU:C:2020:461, at 27; Case C-683/17, of 12 September 2019, *Cofemel*, ECLI:EU:C:2019:721, at 29; Case C-393/09, BSA, of 22 December 2010, ECLI:EU:C:2010:816, at 49.

the doctrines.[39] Therefore, there should be no doubt about the general validity of an idea-fact-expression doctrine under EU copyright law.

Nevertheless, the effect of the dispositions contained in Arts. 3&4 CDSM is to formalise an interpretation that significantly reduces the ambit of application of the idea-fact-expression doctrine. This is achieved through the affirmation that non protected mere facts and data when contained in protected woks receive some sort of derivative or reflected form of protection since their (non-protected) reuse requires the making of some sort of transient or temporary copy of the (protected) containing work. In other words, the content is not protected in its own right, the container is. But because there is no viable form of using the content without also using the container the protection of the latter extends to the former. Technically, this is achieved via a broadly defined right of reproduction only partially compensated by corresponding exceptions. Whereas this might sound compelling from a certain point of view, it is a sort of improper syllogism that does not stand the test of a principled analysis of the law. In fact, by drawing a line between protected expressions and non-protected ideas and facts, both copyright law and theory establish a balance between the protection of certain interests on the one hand (investments of rightholders, personality of the authors, etc) and certain competing interests on the other hand (access to knowledge and information by the public). In this way, copyright can foster creativity, innovation and socio-economic welfare. Tilting this balance, while not impossible, should be done with great care and in full consideration of the implications for the fundamental rights at stake.

This legislative technique, i.e., drafting a broadly defined right of reproduction corrected by specific carve-outs, is emblematic of a more general trend which reached its peak with the InfoSoc Directive of 2001. It is characterised by the full harmonisation of copyright's exclusive rights through broad and all-encompassing definitions (Arts. 2-4 InfoSoc), and by the systematic and semantic classification of any area not covered by copyright's exclusivity as an exhaustively listed "exception" (Art. 5 InfoSoc), a concept that in the theory of law possesses the very specific function to derogate from a general rule and therefore is subject to conditions such as that of strict interpretation.[40]

Consequently, the introduction of an exception establishing that in very specific cases TDM can be freely performed, leads to the exactly opposite effect: all uses that cannot be subsumed within the narrow construction of Arts. 3 and 4 are reserved. Had the legislative technique been

[39] See e.g., Case C-306/05, of 7 December 2006, SGAE, ECLI:EU:C:2006:764, at 35.
[40] As an example, "quotations" are classified as "free uses" under Art. 10 Berne Convention, but as "exceptions and limitations" under Art. 5(3)(d) Directive 2001/29/EC.

different, rejecting the rhetoric of "exceptionalism" and moving towards an approach where concurring rights are clearly delineated, the result would have been more in line with the identified international norms and theoretical frameworks. As a mere illustration, one could look at the path taken in Art. 14 CDSM. That article plainly clarifies that the digitisation of works of visual art does not create new rights in the copyright or related rights field. Similarly, the legislator could have simply clarified that the extraction of non-protected facts and data from protected works does not infringe copyright. Extra EU legal systems have embraced a variety of approaches where the different ingredients of exclusivity, access and technological development were combined to adjust to domestic priorities and legal traditions. However, in most of these systems, which can be counted as "competitors" of the EU in the technological, creative and cultural fields, the adopted solutions have all struck balances that on comparison are more favourable to technological development. Illustratively, the following main approaches can be identified: open and flexible standards,[41] the judicial construction of users' rights,[42] or a dedicated TDM limitation for *any purpose.*[43]

In relation to the effects of the broad definition of the right of reproduction in Art. 2 InfoSoc, it is insightful to note that already during the phase that led to its adoption in 2001 this approach was met with criticism. As Prof. Hugenholtz pointed out in his seminal article on copyright and freedom of information written in the wake of the InfoSoc approval:

> In commenting upon the Green Paper that preceded the [InfoSoc Directive], the Legal Advisory Board (the "LAB"), the body that advises the European Commission on questions of information law, observed: "[…] In the opinion of the LAB, the extent and scope of these rights are clearly at stake, if as the Commission suggests (Green Paper, p. 51-52), the economic rights of rightholders are to be extended or interpreted to include acts of intermediate transmission and reproduction, as well as acts of private viewing and use of information. […]" According to the LAB, the broad interpretation of the reproduction

[41] This is the US approach, but it has been adopted by other countries among which Singapore, South Korea, Malaysia, Israel, Taiwan. See Elkin-Koren N., Netanel N., Transplanting Fair Use across the Globe: A Case Study Testing the Credibility of U.S. Opposition, 72 Hastings Law Journal 1121 (2021).

[42] The interpretation of the fair dealing provision by Supreme Court of Canada led many authors to consider Canada's fair dealing as a type of fair use; see e.g., Geist, M. (2013) Fairness Found: How Canada Quietly Shifted from Fair Dealing to Fair Use. In Geist, M. (Ed.), The Copyright Pentalogy: How the Supreme Court of Canada Shook the Foundations of Canadian Copyright Law, University of Ottawa Press, available at: https://books.openedition.org/uop/969. A perhaps similar development could be seen – albeit still in an embryonic from – in some CJEU decisions, see M. Senftleben (2019) "Bermuda Triangle – Licensing, Filtering and Privileging User-Generated Content Under the New Directive on Copyright in the Digital Single Market" EIPR 41(8), 480, 481; Dusollier S. (2020) The 2019 Directive on Copyright in the Digital Single Market: Some progress, a few bad choices, and an overall failed ambition, Common Market Law Review, Vol. 57(4), pp. 979-1030; Sganga C. (2020) A new era for EU copyright exceptions and limitations?, ERA Forum, vol. 21, pp. 311-339.

[43] This is the course taken more than 10 years ago by Japan, see Ueno T. (2021) The Flexible Copyright Exception for 'Non-Enjoyment' Purposes – Recent Amendment in Japan and Its Implication, GRUR International 70(2), pp. 145-152.

> right, as advanced by the Commission, would mean carrying the copyright monopoly one step too far. [...].[44]

The advice of the LAB seems to have been largely ignored in the adopted text. However, its message should not be completely lost. To rebalance the amplitude currently enjoyed by the right of reproduction, the most direct intervention would be to redefine it, i.e. a modification of Art. 2 InfoSoc. However, this seems a highly unlikely course of action at present time.[45] Looking for alternatives, whereas the "exceptionalist" rhetoric of EU copyright law has been criticised above for carrying not only semantic but also meaningful prescriptive implications, a broad and possibly flexible TDM exception, or perhaps even better a "computational uses exception", could be an acceptable compromise. This would need to be wider than the current EU TDM one(s) and wider than what was known as "option four".[46] However, also this door appears to have been firmly shut after the contentious approval of the CDSM.[47] Remaining within the field of exceptions, a useful contribution could be found in a technology-oriented interpretation of an existing provision which, while not specifically drafted for TDM, the CDSM has confirmed as capable of covering certain TDM activities: Art. 5(1) InfoSoc.[48] While not specific to computational uses, Art. 5(1) was implemented with the goal of enabling certain technological uses (mainly internet browsing[49]) and to rebalance the excessive scope afforded to the right of reproduction. It is also the only mandatory exception of the whole InfoSoc Directive which has the important advantage of favouring cross-border uses.

Before proceeding to an analysis of Art. 5(1), it should be noted that the CDSM Directive clarifies that "Member States may adopt or maintain in force broader provisions, compatible with the exceptions and limitations provided for in Directives 96/9/EC and 2001/29/EC, for uses or fields covered by the exceptions or limitations provided for in this Directive".[50] For present

---

[44] Hugenholtz, Copyright and freedom of expression in Europe, in Cooper, Dreyfuss et al. (eds.), Innovation Policy in an Information Age, Oxford, OUP 2000, at p. 9.

[45] Proposing a different interpretation of the relationship "right-infringement" for Art. 2 ISD which relies *inter alia* on the CJEU "recognizability" test expressed in the *Pelham* case in relation to Art. 2(c), see R. Ducato, A. Strowel (2021) Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out, 43 European Intellectual Property Review EIPR 5, pp. 322-337.

[46] In the Impact Assessment, the EC identified as "Option four" a TDM exception not limited to research organisations for research purposes.

[47] Senftleben M. (2017) The Perfect Match – Civil Law Judges and Open-Ended Fair Use Provisions, American University International Law Review, Vol. 33, Issue 1; Hugenholtz (2016) Flexible Copyright: Can EU Author's Rights Accommodate Fair Use?, in Stamatoudi, (Ed.), New Developments in EU and International Copyright Law, Kluwer Law International.

[48] Recital 9 CDSM. See also Triaille et al., Study on the legal framework of text and data mining (TDM), European Union, 2014, p. 41, available at: https://op.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en; Margoni, T. (2018) Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?, AIDA 2018(1), pp. 1-26.

[49] Recital 33 InfoSoc Directive.

[50] Art. 25 CDSMD.

purposes this means that MS may maintain or introduce a new TDM exception usually on the basis of Art. 5(3)(a) InfoSoc (i.e. illustration for non-commercial teaching and scientific research). Beside the non-commercial *versus* research-purposes-by-research-institutions discussion, Art. 5(3)(a) exception is not mandatory therefore it does not represent an EU wide solution to the problem addressed in this paper and as such will not be discussed any further. It should be pointed out, however, that this exception, like all the exceptions listed under Art. 5(3) InfoSoc, covers both reproductions and communications to the public thereby offering an opportunity to MS interested in implementing a wider exception.[51]

II.3. Art. 5(1): An enabler for technological development?

The CJEU in *Infopaq I* and *II* had the occasion to clarify that temporary acts of reproduction made during "data capture" processes can be covered by the exemption of Art. 5(1) if its five cumulative and strictly interpreted conditions are met.[52]

Art. 5(1) requires that the reproduction be (1) temporary; (2) transient or incidental; (3) an integral and essential part of a technological process; (4) the sole purpose of which is to enable … a lawful use of a work; and (5) the act has no independent economic significance.

Regarding conditions (1) and (2), the *Infopaq I* Court clarified that temporary and transient acts of reproduction are "intended to enable the completion of a technological process of which it forms an integral and essential part". In those circumstances those acts of reproduction "must not exceed what is necessary for the proper completion of that technological process", being understood that "that process must be automated so that it deletes that act automatically, without human intervention, once its function of enabling the completion of such a process has come to an end".[53]

In *Infopaq II* the CJEU offered some further insights on the proper interpretation of the remaining conditions:

> (3) The concept of integral and essential part of a technological process requires the temporary acts of reproduction to be carried out entirely in the context of the implementation of the technological process. This concept also assumes that the completion of the temporary act of reproduction is necessary, in that the technological process concerned could not function correctly and efficiently without

[51] Some Member States took full advantage of this opportunity (e.g. France, Estonia, Germany), whereas others did not (e.g. UK); see Geiger et al., The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects, Policy Department for Citizens' Rights and Constitutional Affairs, PE 604.941- February 2018.
[52] Case C-5/08, *Infopaq I*, ECLI:EU:C:2009:465 and Case C-302/10, *Infopaq II*, ECLI:EU:C:2012:16.
[53] See *Infopaq I*, 61 – 64; Chiou T., Copyright lessons on Machine Learning: what impact on algorithmic art?, 10 (2020) JIPITEC 398 para 1, at 14, available at: https://www.jipitec.eu/issues/jipitec-10-3-2019/5025.

> that act. This condition is satisfied notwithstanding the fact that initiating and terminating that process involves human intervention.[54]
>
> (4) Temporary acts of reproduction must pursue a sole purpose, namely, to enable [...[55]] the lawful use of a protected work, which is in turn fulfilled when such use is authorised by the rightholder or where it is not restricted by the applicable legislation.[56]
>
> (5) Temporary acts of reproduction do not have an independent economic significance provided that the implementation of those acts does not enable the generation of an additional profit distinct or separable from the economic advantage derived from the lawful use of the work; and the acts of temporary reproduction do not lead to a modification of that work.[57]

The Court also importantly clarified that as long as the conditions of Art. 5(1) as interpreted above are met, the three-step test of Art. 5(5) is satisfied.

A very brief description of the facts of the *Infopaq* case may be helpful to properly situate these conditions within a data capture process which shares many logical steps with more modern TDM approaches.[58] In this case the Court was asked whether the compilation, extraction, indexing and printing of newspaper articles and keywords by a media monitoring service infringed the copyright in said articles. The Court identified five relevant phases in the process of data capture: (1) newspaper publications are identified and registered in an electronic database; (2) sections of the publications are selectively scanned, allowing the creation of a Tagged Image File Format (TIFF) file for each page of the publication and its transfer to an Optical Character Recognition (OCR) server; (3) the OCR server processes this TIFF file digitally and translates the image of each letter into a character code recognisable by computers and saves it as a text file, while the TIFF file is then deleted; (4) the text file is processed to find a user-defined search word, identifying possible matches and capturing five words before and after the search word (i.e. a snippet of 11 words) before the text file is deleted; (5) at the end of the data capture process, a cover sheet is printed out containing all the matching pages as well as the text snippets extracted from these pages.

The following is an example of the results produced by the Infopaq media monitoring service:

> 4 November 2005 — Dagbladet Arbejderen, page 3:

[54] *Infopaq International A/S v Danske Dagblades Forening*, Case C-302/10 of 17 January 2012, ECLI:EU:C:2012:16 [Infopaq II].
[55] "... either the transmission of a protected work or a protected subject-matter in a network between third parties by an intermediary or ..."
[56] *Infopaq International A/S v Danske Dagblades Forening*, Case C-302/10 of 17 January 2012, ECLI:EU:C:2012:16 [Infopaq II].
[57] Id.
[58] A detailed analysis can be found in Margoni, T. (2018) Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?, AIDA 2018(1), pp. 1-26.

TDC: 73 % "forthcoming sale of the telecommunications group TDC, which is expected to be bought".[59]

The Court found that the exception of Art. 5(1) only exempts the activities listed in points 1) to 4) above, whereas the activity of point 5), i.e., printing, constitutes a permanent act of reproduction which is therefore not covered by an exception for temporary copies. When this activity reproduces the original work in part as defined by Art. 2 InfoSoc, it has the potential to constitute a copyright infringement. In the same dispute, the Court of Justice clarified that it cannot be excluded that even 11 consecutive words, when representing the author's own intellectual creation, may qualify as an Art. 2 reproduction in part, i.e., as copyright infringement.

The conditions 1) to 4), which as the CJEU pointed out must be interpreted strictly as they derogate from the general rule,[60] are not always easy to meet in TDM processes nor is their interpretation always straightforward. That said, within a copyright framework that does not offer many other alternatives, Art. 5(1) represents an important ally as an enabler of technological development. This is an aspect acknowledged by the same CJEU, when it states that the function of 5(1) is to "*allow and ensure the development and operation of new technologies, and safeguard a fair balance between the rights and interests of rights holders and of users of protected works who wish to avail themselves of those technologies*".[61]

The statement's ethos seems to offer a comfortable safe harbour for modern TDM and data-driven AI processes. However, while the proposition seems directed towards a technology-enabling goal, it is not an equally comfortable exercise to imagine how the *rights and interests of users* of protected works to avail themselves of new technologies and the very same development of such new technologies can be safeguarded by a strict interpretation of the already narrowly defined five conditions of Art. 5(1).

II.3.a) Eroding lawful uses

It should be considered the eventuality that Art. 3&4 CDSM may have contributed to narrow even further the scope of Art. 5(1) InfoSoc. This is due to condition 4) and the concept of lawful use. A lawful use is a use authorised by the rightholder (e.g., via a licence) or not restricted by the applicable legislation.[62] In *Infopaq II* the Court states that "[…] the parties in the main proceedings do not dispute that in itself summary writing is lawful and does not require consent from the rightholders", that "such an activity is not restricted by European Union legislation" and

[59] Infopaq II.
[60] Case C-5/08 *Infopaq International*, EU:C:2009:465, paragraphs 56 and 57; Joined Cases C-403/08 and C-429/08 *Football Association Premier League and Others*, EU:C:2011:631, paragraph 162; Case C-360/13, NLA, ECLI:EU:C:2014:1195, paragraph 23.
[61] Case C-360/13, NLA, ECLI:EU:C:2014:1195, paragraph 24.
[62] Rec. 33 ISD; *Infopaq II*, 68; *FAPL*, 168.

finally that "it is apparent from the statements of both Infopaq and the DDF that the drafting of that summary is not an activity which is restricted by Danish legislation". These statements need closer scrutiny.

Regarding the first one, it seems that the Court is satisfied with the fact that parties in the main proceeding do not dispute the issue of summary preparation which allows the Court to avoid, on a procedural ground, a particularly tricky legal question. Regarding the second and third statements, it would be interesting (albeit beyond the scope of this paper) to verify whether it is domestic law which does not provide for a right of adaptation that covers the creation of summaries (or at least of summaries which reproduce in part the original work, such as 11 consecutive words); whether domestic law does it, but a specific exception excuses the activity; or finally whether this type of summaries are not covered by the right of adaptation due to the marked factual nature of the original articles. From a Union law point of view, the Court seems to implicitly reaffirm the absence of a horizontally harmonized right of adaptation which again allows the Court to avoid entering into an analysis of whether summaries are a form of "adaptation, arrangement and other alteration" (cf. Art. 12 Berne Convention). Regardless of the reasons that allow the Court to avoid and in-depth assessment of summary preparation under applicable copyright law, it is worth noting that the applicability of Art. 5(1) to the present case and therefore the more general permissibility of data capture processes under EU law entirely relies on the statement that the preparation of summaries is not a right reserved to rightsholders under applicable law. A statement that finds minimal support in the decision and which leaves open the possibility for domestic legal orders to deviate from this rule (as arguably many do).

This brief analysis intends to underscore the thin theoretical ground on which the entire concept of lawful use stands in Art. 5(1). If a lawful use is a use not reserved by law, but the law through a very wide right of reproduction reserves virtually any type of use save for when an exception applies, then the situation where Art. 5(1) finds application are logically limited to those situations where another exception already applies or when the use of a work does not trigger the right of reproduction (such as the preparation of summaries in the above case).

It follows, that if Art. 5(1) is only available when a certain use is not restricted by applicable legislation, the recognition that TDM is a reserved use of rightholders (excused when performed by research and cultural organizations for research purposes or when it is not contracted out), means that temporary acts of reproduction performed for TDM purposes outside the scope of Art. 3&4 CDSM are not permitted any longer as they do not meet the condition of lawful use. This is an odd and probably unforeseen effect of the provision, since the very same CDSM states that

Art. 5(1) should *continue* to apply to TDM (Rec. 9). It seems difficult to find a logical explanation for the described situation. Certainly, the crucial function of Art. 5(1), i.e., the right of users of protected works to avail themselves of new technologies seems incompatible with the described situation. If user rights and technological development are to be safeguarded under EU copyright law, the formalistic interpretation embraced by the CJEU in certain cases needs to be abandoned in favour of a teleological approach to EU copyright law which the same Court has adopted in other judgements.

II.3.b) The function of permanent reproductions in computational uses and in the development of trusted AI systems

Retaining permanent copies represent a crucial tool to mitigate the black box of AI (discussed at the beginning of section II). Greater transparency may enable trust in AI systems that make decisions affecting in ever more sophisticated ways the life and the rights of individuals. There are two types of fundamental reproductions in TDM and machine learning whose persistence needs to be ensured.

The first type is the one created by text and data analysis which corresponds to the "memory" of the AI application, also known as the "trained model". As it has been explained in more details elsewhere in relation to NLP,[63] in a typical ML workflow a learning algorithm trains a model, i.e., records in a permanent format (a file) the information that has been extracted from the original data. This model is the placeholder of what the machine has learned without which anything that has been inferred (patterns, correlations, links etc) would vanish as if it never existed. Sometimes this trained model only contains highly abstract statistical representations of the original data. This is especially the case with more sophisticated approaches to machine learning, such as so called "deep learning", where the expression "deep" indicates that the abstraction is structured in additional intermediate arbitrary categories, and thus the analysis reaches "deeper". At other times, in addition to the statistical information, the trained model also contains parts of the original data. When the original training data is protected (a literary work, a qualifying database) and when the information stored in the trained model qualifies as a reproduction in part (e.g., even 11 consecutive words, how many data points?) or when the trained model can be considered an adaptation of the original training data (e.g., a thumbnail representing the searched websites), Art. 5(1) is of no avail. In this case, an enabling provision should ensure that "functional" permanent copies (i.e. the trained model

[63] Eckart de Castilho et al. (2018) A Legal Perspective on Training Models for Natural Language Processing, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA).

containing the author's own intellectual creation or a substantial extraction of the database) can not only be stored but also shared (e.g. communicated to the public) for any purpose. As we will see, Arts. 3&4 CDSM have failed to fully address this first type of permanent copies.

A second type of permanent copy is the one necessary for verification purposes. For something to be called scientific, it has to be based on replicable results, which in turn can only be achieved if the data, methods and analysis of the experiment are available for verification. This aspect is central to scientific enquiry where in the last decades the so-called "reproducibility crisis" of scientific results emerged. This phenomenon affects both social and hard sciences and has been extensively explored in the literature, which has identified both sector specific and more general issues at its basis.[64] A common cause of replicability failure is however the absence of sufficient disclosure of the data and the methods employed to reach a certain result. This situation has led to strong calls for more open and accountable disclosing and publishing practices, often under the name of Open Science.[65] Yet, it is not only scientific results which need to be obtained following a transparent and accountable methodology that allows an independent observer to understand and replicate them. Decisions affecting individual or collective rights should also follow similar principles and they usually do in the off-line world. Not only parliamentary statutes and acts, but also the preparatory materials that were used to draft them are normally available for public scrutiny, as are the parliamentary sessions where discussions are held. Similar patterns characterise many of the offices that make decisions affecting private and public interests, such as courts of justice, central and local governments, regulatory authorities and the like. The freedom to receive, impart and access information is a central tenet of modern democracies and is enshrined in EU's and MS's fundamental laws. Therefore, AI systems deciding whether a certain loan or credit card should be issued or whether access to a certain school, programme or job should be granted, or again decisions relating to macroeconomic, public health or epidemiologic aspects affecting the life of people should be open, accountable and verifiable. There seems to be little space, if any, in European's fundamental laws for public authorities to avail themselves of unaccountable AI applications. Private actors might decide that this is the right solution for them, and different legal systems

[64] Ioannidis J. P. A. (2005). Why Most Published Research Findings Are False. PLoS Medicine. 2 (8): e124, available at: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124.
[65] This is an explicit priority of the European Commission, see https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science_en (visited 1 July 2021).

may agree that market dynamics should regulate these decisions, either with or without public interventions to correct certain distortions in strategic sectors.[66]

In order to be able to "understand" what determinations are being made by AI systems even more important than the algorithm itself, it is the data used to train those algorithms. To fulfil this scope, such data must be available to the public to ensure that certain decisions were not based on information that may be old, inaccurate, incomplete or in any other way biased. A serious reflection on the kind of decisions that AI systems based on information (technological and scientific knowledge, values, morals, etc) that is out of copyright term (i.e. at least 70 years old) seems to be absent. Whereas it will not always be possible to understand why certain conclusions were reached by the AI, an open, accountable and verifiable approach will ensure that the same substantive and procedural guarantees of fairness, accountability and rule of law that have emerged in our societies over centuries of legal culture will not be obfuscated behind the unintelligible complexity of statistical inference.[67] While this type of permanent copy is not covered by an exception for temporary uses, some limited recognition of this aspect is present in Arts. 3&4 CDSM.

In conclusion, whereas Art. 5(1) retains a significant potential for TDM activities and computational uses, the cumulative, occasionally narrow and partially uncertain nature of its conditions and the fact that it only covers temporary reproductions, does not offer a clear and comprehensive solution within which not only science but virtually any human activity employing text and data analytics can move confidently.[68]

II.4. Was the EU TDM exception needed? Brief theoretical considerations

As stated in the Introduction, copyright protects the original expression of ideas, not ideas themselves, mere facts or data.[69] Accordingly, whereas TDM should not be considered a copyright infringement, it is not through a copyright *exception* that the issue is best addressed. The reason is that TDM mainly refers to the use of unprotected ideas, principles, facts and data,

[66] In the EU see the proposal for an AI Regulation: Proposal for A Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM/2021/206 Final (21.4.2021).
[67] Zittrain J., Intellectual Debt: With Great Power Comes Great Ignorance, 24 July 2019, https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c.
[68] Triaille et al. (2014) Study on the legal framework of text and data mining (TDM), European Union, 50, available at: http://ec.europa.eu/internal_market/copyright/docs/studies/1403_study2_en.pdf.
[69] E.g., Art. 2 WIPO Copyright Treaty and in Art. 9(2) WTO's TRIPs Agreements and Recital 8 CDSMD.

often contained in literary works or other types of texts (text mining) or in structured and/or unstructured datasets (data mining). TDM is simply external to copyright's scope.[70]

The same Berne Convention (BC) is based on the basic tenet that only *original expressions* are protected and not underlying ideas or facts.[71] This can be inferred not only from the general principles and the national traditions underpinning the BC,[72] but also from its literal meaning. Art. 2 ("Protected Works"), clarifies that every production in the literary, scientific and artistic domain is protected, whatever may be the *mode or form of its expression.*[73] The article further offers a list of non-exhaustive examples of these productions, all of which are specific expressions of human intellectual creations. The fact that protection offered by the BC to literary and artistic works "does not extend to the ideas embodied in those works, but only to the form in which those ideas are expressed" and that "[T]he same is true of factual information and subjects (in the case of artistic works): no writer or artist can have a monopoly over these things, which can be freely used in their works by other authors" are fundamental copyright axioms.[74] This is confirmed, among others, by the same WIPO guides to the Berne Convention which clearly states that "The scientific work is protected by copyright not because of the scientific character of its contents ... but because they are books and films" and that ideas are not protected but "it is the form of expression which is capable of protection and not the idea itself".[75] Similarly, that "only concrete original expressions of ideas are [protected] may be deduced from the basic meaning of the generic expression "production." A mere idea is obviously not yet a production; it is only transformed into a production when it is developed into a concrete form of expression".[76]

Furthermore Art. 2(8) bars protection for news of the day or to miscellaneous facts having the character of mere items of press information confirming that facts are explicitly excluded from copyright's ambit. They do not contain the minimum elements of intellectual creation and thus do not qualify as works.[77] The same principle underpins other articles in the international

[70] ECS, General Opinion on the EU Copyright Reform Package, 24 January, 2017, at 5, available at: https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf; Sag M. The New Legal Landscape for Text Mining and Machine Learning, 66 J. of the Copyright Soc'y of the USA, 3, 9-19 (2019); Carys J. Craig, Globalizing User Rights-Talk: On Copyright Limits and Rhetorical Risks, Am. U. Int'l L. Rev. Vol. 33(1) (2017).
[71] Ricketson & Ginsburg (2010) International Copyright and Neighbouring Rights, 407.
[72] Id., 406.
[73] Art. 2(1), BC.
[74] Ricketson & Ginsburg (2010) International Copyright and Neighbouring Rights, 407.
[75] Masouyé (1978) Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971), WIPO, Geneve.
[76] Ficsor (2003) Guide to the Copyright and Related Rights Treaties Administered by WIPO, 23 – 24.
[77] Art. 2(8), BC; See also Masouyé (1978), 22.

copyright framework such as Arts. 10(2) TRIPs and 5 WCT which clarify that copyright in compilations of data "do not extend to the data or material itself" or in Art. 1(2) EU Software Directive which clarifies that copyright protection for software applies "to the expression in any form of a computer program. Ideas and principles which underlie any element of a computer program, including those which underlie its interfaces, are not protected by copyright".[78] Similarly, the CJEU confirms that single words cannot be considered original expressions since words considered in isolation are not an intellectual creation of the author who employs them[79] and that "keywords, syntax, commands and combinations of commands, options, defaults and iterations consist of words, figures or mathematical concepts which, considered in isolation, are not, as such, an intellectual creation of the author of the computer program".[80]

It is not only creativity that is protected by excluding ideas, facts and principles from protection. Freedom of expression, i.e., the ability to freely express and receive one's ideas and opinions is a fundamental right recognised in all modern democratic constitutions and therefore any form of limitations to that right should be resisted and limited to specific cases identified by law. The law bears the crucial task of striking a balance between freedom of expression and other concurring rights, and in copyright theory this is done also by establishing that while ideas and facts cannot be limited by a concurring right, specific original expressions of those ideas and principles can be protected as property. This basic principle can be found in the case law of most modern legal systems. Illustratively, the U.S. Supreme Court in *Harper & Row, Publishers, Inc. v. Nation Enterprise*[81] recognised that the function of the idea-expression dichotomy is to define a "balance between the First Amendment and the Copyright Act by permitting free communication of facts while still protecting an author's expression". Similarly, the CJEU clarified in decisions such as *Promusicae*[82] and *Scarlet*[83] that a fair and proportionate balance must be struck among the different fundamental rights recognised by Union law, in particular the right to property, which includes intellectual property, the right to private life, the right to freedom of expression and the right to conduct a business.

If the above is plausible, then it should become clearer why addressing TDM as a copyright *exception* is conceptually wrong and theoretically flawed. Ideas, facts and data are not

[78] Goldstein & Hugenholtz (2010) International Copyright – Principles, Law and Practice, 2nd Ed., 5, 220; Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version), Art. 1(2).
[79] *Infopaq International v Danske Dagblades Forening*, Case C-5/08 of 16 July 2009, ECLI:EU:C:2009:465, 45; See also Bently & Sherman (2014) Intellectual Property Law, 4th Ed., 111 – 112.
[80] *SAS Institute v World Programming Ltd.,* Case C-406/10 of 2 May 2012, ECLI:EU:C:2012:259, 66 – 67.
[81] 471 U.S. 539, 556 (1984).
[82] *Promusicae v Telefónica de España SAU*, Case C-275/06 2008, ECLI:EU:C:2008:54.
[83] *Scarlet Extended v SABAM*, Case C-70/10 of 24 November 201, ECLI:EU:C:2011:771.

copyrightable elements, therefore there should be no need for a copyright exception in order to use those elements, even when they are contained in protected works. Nonetheless, in practice, in the current state of EU copyright law as it stands today a clarification of the legality of TDM was necessary and an exception, if properly formulated, could have been one of the ways to achieve this goal, albeit not the best one.

II.5. The enacted EU TDM exception(s): Practical considerations

The main criticisms against the current formulation of Arts. 3 and 4 (understood against the misguided theoretical approach) can be structured according to the following elements: 1) definition; 2) beneficiaries; 3) rights; 4) technological overridability; and 5) access to original sources. Two additional characteristics can be seen as functional to safeguarding the exception's scope: 1) contractual overridability (which will be addressed together with point 4 above); and 2) storage of copies for verifiability.

II.5.a) Definitions

As seen in the first part of this article, whereas a broad definition of TDM activities could be seen as functional to cover a broader set of free uses, its insertion into the current *Aquis* may have produced the opposite effect. We refer to the analysis developed above.

II.5.b) Beneficiaries

Art. 3 introduces a double limitation: it can only be performed by *research organisations and cultural heritage institutions* and only *for the purpose of scientific research.* Therefore, a commercial enterprise will not be able to benefit from the exception. Nor a university acting for any other purpose than scientific research. Other purposes commonly accepted as fundamental in democratic societies appear also excluded, such as journalism, criticisms or review.[84]

In the opinion of the drafters of the Directive, the current wording is thought to be less restrictive than the "non-commercial" limitation.[85] It seems however, that Art. 3's double limitation is very close to the non-commercial requirement and in certain respects even more restrictive in the sense that a "non-commercial" limitation would arguably allow a business acting for non-commercial scientific research purposes to benefit from the exception, something that is not possible under Art. 3 (although Public-Private Partnerships are explicitly allowed). This is a major

[84] Dusollier S. (2020) The 2019 Directive on Copyright in the Digital Single Market: Some progress, a few bad choices, and an overall failed ambition, Common Market Law Review, Vol. 57(4), pp. 979-1030.
[85] Commission staff working document – Impact assessment on the modernisation of EU copyright rules, SWD(2016) 301 final, Part 1/3, Brussels, 14.9.2016, 108 – 109; available at: http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=17211.

limitation to the efficacy of the exception that excludes important economic sectors and SMEs from benefiting from a critically important tool to compete on the global markets. This limitation appears in contrast with fundamental rights such as the freedom of expression and the freedom to conduct a business, even though the same preparatory material excludes such a contrast.[86]

Art. 4, which is not a direct emanation of "Option 4", but which may nevertheless have benefited from its assessment, is not limited to certain beneficiaries and thus potentially available to all. It is characterised however by the additional element of being capable of "opt-out" by right-holders, a provision that may very well frustrate its efficacy. It would be important, during the national implementation phase, to clearly identify how this opt out should be performed in the light of the general guidance offered by Art. 4. It would also be important to develop public awareness around the need to not unnecessarily restrict TDM.

II.5.c) Rights

Another significant limitation found in both Arts. 3 and 4 is that they only exempt potential infringements of the right of reproduction but not of the right of distribution or communication to the public, nor of the (unharmonized) right of adaptation.

This means that in all the cases when the results of an act of TDM include a protected part of the original "mined" work (and as seen above excerpts as short as 11 consecutive words could be protected) these results cannot be communicated to the public or redistributed. In certain areas this will not represent a cause of concern, however in other areas, e.g., Natural Language Processing (NLP), the fact that certain models trained on a number of copyright protected *corpora* (i.e. texts) could include reproductions in part, means that those models, the result of the research purpose conducted by the research organisation, cannot be redistributed or communicated publicly. Outside textual sources, e.g. in the case of audio-visual works it may be even more difficult to establish when this threshold is reached. The question of whether a trained model can be considered an adaptation of the original corpora is excluded *ratione materiae* from the EU assessment, but is an aspect that will need to be clarified.

II.5.d) Contractual and technological overridability

Arts. 3 and 4 state that contractual provisions intended to limit the TDM exception shall be unenforceable. This is an important provision, as many times access to databases is based on acceptance of Terms of Use that limit TDM. Nevertheless, if the same contractual provision contrary to the TDM exception is expressed through an effective technological measure, there

[86] Page 9 of the Proposal; See Geiger et al., Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?, IIC - International Review of Intellectual Property and Competition Law volume 49, pp. 814-844 (2018).

is no equivalent provision safeguarding the enjoyment of the exception. The approach taken by the CDSM is convoluted at best. Art. 6 second sentence reads "The first, third and fifth subparagraphs of Article 6(4) of Directive 2001/29/EC shall apply to Articles 3 to 6 of this Directive". In extreme synthesis this means that the TDM exceptions are inserted in the list of exceptions for which the InfoSoc Directive establishes that: 1) if a user with legal access to a work is entitled of an exception; and 2) that exception cannot be enjoyed due to the presence of an effective technological measure; and 3) rightholders have not voluntarily taken any measures to ensure that said user can enjoy the illegitimately restricted exception; then 4) Member States shall take appropriate measures to ensure that rightholders make available to said beneficiaries of the exception the means of enjoying it.

It is important to note that subparagraph 4 of Art. 6(4) does not find application in this case. Subparagraph 4 establishes that the reported mechanism (the obligation on MS to facilitate the enjoyment of an exception illegitimately restricted by rightholders via effective technological measures) is excluded when rightholders make available works to the public on agreed contractual terms in such a way that members of the public may access them from a place and at a time individually chosen by them, thereby rendering largely ineffective the entire provision.

Even though the CDSM recognises the importance of excluding subparagraph 4, it is the entire mechanism of Art. 6(4) InfoSoc that has proven highly ineffective due to its convoluted formulation and ultimately to the fact that it places the burden of reclaiming legitimate uses allowed by the law but illegitimately restricted by technological locks on the shoulders of end users. Illustratively, in the UK where the UK Intellectual Property Office (IPO) has set up a specific complain procedure,[87] a total of 11 applications have been filed since 2003, 9 of which failed as they related to computer programmes – an excluded category – 1 was rejected considering subparagraph 4 mechanism, and 1 lead to a voluntary solution.[88]

II.5.e) Lawful access to original sources

Art. 3 requires lawful access to the works that will form part of data analysis. Not much justification can be found in the preamble of the Directive. Some more details about the role of

[87] See https://www.gov.uk/government/publications/technological-protection-measures-tpms-complaints-process.

[88] See https://www.gov.uk/government/publications/complaints-to-secretary-of-state-under-s296ze-under-the-copyright-designs-and-patents-act-1988. These are data from 2015. A FOI request to the UK IPO revealed that since 2015, an additional two requests were filed, one rejected (due to paragraph 4 exemption) and one resolved on a voluntary basis. Ironically, this latter request, the only one that has somehow had a successful outcome in almost two decades, was based on the since repealed private copy exception.

the "lawful access" requirement can be found again in the Impact Assessment: "... the "lawful access" condition, i.e. [by the fact that] the exception would not affect publishers' ability to continue to authorise or prohibit access to their content and to generate revenues from selling subscriptions to universities and other research organisations".[89]

It has been argued that a TDM exception should be considered licit also when access to the training data does not fulfil the lawful access requirement.[90] The arguments to support such a position are multiple. As prof. Carroll puts it: "copies are made only for computational research and the durable outputs of any text and data mining analysis would be factual data and would not contain enough of the original expression in the analysed articles to be copies that count. Reference copies would be kept and shared only for reproducibility purposes or for further computational research and would not be otherwise made available".[91] Whereas such argument is developed within the US copyright framework which, as briefly discussed above, operates quite differently in relation to some of the elements of EU copyright law here scrutinised, it seems that the same rationale could find application also under EU law. Furthermore, it has been pointed out how the lawful access limitation could subject TDM research to private ordering[92] as well as severely impair other fundamental rights such as the freedom of information and to inform the public about specific undisclosed but publicly relevant issues, especially when these are "leaked" by whistle-blowers, and thus as such often failing the lawful access requirement.[93]

II.5.f) Storage of copies for verifiability

Art. 3 provides that "copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results". This is a very important element to ensure the verifiability of results. Regarding the fundamental importance of this condition, we refer to the analysis developed above. Regarding the present provision, while it is an important step to ensure the transparency and accountability of algorithmic decision-making tools, a degree of uncertainty connected with the specific formulation endures. In particular, it is not clear what the access dimension to such stored copies would be. In fact, if the research community needs access to the stored copies for verification purposes, the first researcher or institution who originally collected the material and who is storing it might engage in acts of communicaton or making those copies available to the public, whereas, as said, Art. 3 (and 4) are

[89] Impact Assessment, p. 114 Part I.
[90] See Michael W. Carroll, Copyright and the Progress of Science: Why Text and Data Mining Is Lawful, 53 U.C. Davis L. Rev. 893 (2019); see also Geiger et al., Text and Data Mining: Art. 3 and 4, cit., at 33.
[91] Id., at 954.
[92] See Geiger et al., Text and Data Mining Art. 3 and 4, cit., at 33.
[93] In this sense see Dusollier, cit., 987.

exceptions only to the right of reproduction. This appears an important area in need of clarification during the phase of national implementation. Additionally, Art. 3(4) establishes that "Member States shall encourage rightholders, research organisations and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligation and of the measures referred to in paragraphs 2 and 3 respectively". These obligations relate to safe storage provision and security and integrity measures. It would be important to ensure that Art. 3 will become effective as soon as it is transposed into domestic law, regardless of when the commonly agreed best practices are adopted.

III. EU Copyright law and data property

The position embraced in the CDSM about the proprietarisation of mere facts and data is ambivalent. Whereas when considered *as such* they seem excluded from protection, when they are contained in a protected work, they become object of exclusivity. The reason is to be found in the well-known ubiquity of copies in the digital environment. Whereas this reason is global, EU copyright law has developed an idiosyncratic approach characterised by a relatively low level of originality, the protection of qualifying non-original databases (and therefore of factual data) and by a broadly defined and broadly interpreted right or reproduction that is able to capture most types of digital uses. This formalistic approach to computational uses should be wholly rejected. It frustrates and renders ineffective some of the most important fundamental copyright principles, such as the idea-fact-expression dichotomy, the concept of intellectual creation, exceptions and limitations and ultimately the very same concept of work of authorship – all principles that embed fundamental rights such as freedom of expression, property and economic initiative. Copyright has never been about controlling the use of information contained in a work, unless we look at the censorial prototypes predating the early statutes of the eighteenth century. In the past *use* only referred to human use as no other type of uses were known. Today, as demonstrated above, this principle should extend to machines, i.e. to *computational uses*. Controlling the use of information and cultural productions is the domain of other fields of law, such as media and telecommunication law and areas of public and criminal law, which implement public ordering procedures and guarantees to avoid the dangers of censorial abuses. Copyright should not be employed to regulate aspects for which it was not designed and does not possess the tools or the procedures. However, if a conscious decision was to be made to move towards this function of copyright law, then this should be made explicit and be part of an open and transparent process. More importantly, this crucial development should not be part of a tacit, possibly surreptitious and probably unintentional effect.

III.1. Two futures separated by a common provision

The current EU copyright framework seems to be caught in between two possible futures. This unenviable situation may be connected to certain underlying and unresolved contradictions. Two seem particularly pressing. The first is common to many copyright systems worldwide and is caused by the well-known inadequacy of rules devised in the past, sometimes a remote and analogue past, to regulate modern digital practices. After all, the problems here under discussion are intimately related to the advent of digital technologies and the EU's reaction to this advent. A reaction that as evidenced by the roadmap proposed in the Green Paper of 1988[94] interpreted technology mainly as a challenge, which certainly it was, but failed to see it also as an opportunity. It is a known aspect of (not only EU) copyright history that throughout the 1980s and 90s, faced with the paradigm shifting changes brought by digital technologies and under pressure from a content industry that witnessed an unexpected dramatic shift in business models and a potentially steep decline in revenues, EU copyright law tightened its defences, made rights broader, demoted free uses to exceptions which had to be found in a closed but not mandatory list and shielded this new reality behind encryption, i.e. technological protection measures.[95] The striking erosion of free uses and of the public domain can be seen as a direct consequence of this tension. However, this also caused the disruption of the fine balance that copyright used to explicate. Consequently, economic, social and cultural initiatives often clashed against rules that lost the ability to channel innovation while maintaining incentives for investment and protecting the moral dimension of creativity.

The second contradiction is idiosyncratic of the EU legal order and is caused by the inadequacy of national copyright rules to regulate the circulation of information in a single market made up of 27 harmonised but still distinct and territorial copyright markets. This situation is exacerbated by the only partial power that the EU has (had) to regulate copyright, a power which largely relies(d?) on internal market attributions as a legal basis. As explained elsewhere,[96] this limited allocation of competences has led to a patchwork of at least 12 Directives (and 2 Regulations) which, with few exceptions, have harmonised EU copyright law "vertically", i.e. only in relation to

[94] Commission of the European Communities, Green paper on copyright and the of technology, COM(88) 172 final, Brussels 7 June 1988.

[95] Dusollier S. (2020), The 2019 Directive on Copyright in the Digital Single Market: Some progress, a few bad choices, and an overall failed ambition, Common Market Law Review, Vol. 57(4), pp. 979-1030; See also Green Paper on Copyright and the Challenge of Technology - Copyright Issues Requiring Immediate Action. COM (88) 172 final, 7 June 1988; Kretschmer M. (2001) Digital Copyright: The end of an era, European Intellectual Property Review EIPR 25(8), pp. 333-341.

[96] *Ex pluris* Margoni (2016) The Harmonisation of EU Copyright Law: The Originality Standard, in Perry (Ed.), Global Governance of Intellectual Property in the 21st Century, pp. 85-105, Springer; Ramalho (2016) Conceptualising the European Union's Competence in Copyright: What Can the EU Do?, International Review of Intellectual Property and Competition Law, 2, 178.

certain rights or certain subject matter.[97] One of the few directives that has taken a "horizontal" approach (Directive 2001/29/EC, InfoSoc Directive) has done that following an unambitious and to a certain extent contradictory legislative technique based, as already discussed, on the full harmonisation of only certain aspects of copyright (mostly rights) and leaving MS ample discretion with regards to other aspects (mostly exceptions).[98] This approach has resulted in further fragmentation and uncertainty since having diverging rules within a market that proclaims to be single produces tensions. Use of a copyright protected work may be considered lawful in one MS but not in another.[99]

It is also in the light of these considerations that the CDSM aimed to regulate in a mandatory manner and with rules of full or almost full harmonisation at least certain elements of EU copyright law such as the TDM exception. This is certainly laudable. However, whereas the 2019 Directive is timidly but clearly moving in the right direction regarding the second of the above identified tensions – thanks to the mandatory nature of a number of provisions such as Arts. 3 and 4 – it fails to properly address the problems connected with the first tension. In other words, the challenge of digital technologies, after more than three decades, remains a challenge for the EU copyright law.

III.2. Non original property

In order to offer an overview of the issue of property in mere facts and data, a brief mention should be made of other stances where EU copyright law has moved towards a process of propertization of non-personal data. This will offer additional support to the critique here developed concerning the inability (or unwillingness) to address technology as an opportunity. The *Sui Generis Database Right* (SGDR) naturally stands out as a unique EU creature that protects against substantial extractions of data in both original and non-original qualifying databases, thereby *de facto* protecting data under certain circumstances. This approach to data property was rejected in almost any other legal order due to its anti-competitive and anti-information effects. After a quarter of century of its existence, it is far from clear that the SGDR has contributed in any way to the development of the EU (at the time nascent) database market.[100]

[97] See Bechtold in Concise Copyright Law.
[98] Hugenholtz (2000) Why the Copyright Directive is Unimportant, and Possibly Invalid, European Intellectual Property Review EIPR, 11, pp. 499-505; Guibault (2010) Why Cherry-Picking Never Leads to Harmonisation: The Case of the Limitations on Copyright under Directive 2001/29/EC, 1, JIPITEC 55, para. 1.
[99] An illustrative case is *Criminal proceedings against Titus Donner*, Case C-5/11 of 21 June 2012, ECLI:EU:C:2012:370.
[100] The Commission own assessment is revealing: "Despite providing some benefits at the stakeholder level, the sui generis right continues to have no proven impact on the overall production of databases in Europe, nor on the competitiveness of the EU database industry." https://ec.europa.eu/digital-single-

Certainly, it has contributed a discrete amount of work for national and EU courts and has been used in ways that have negatively impacted on consumer's rights and access to knowledge.[101] Nonetheless, as it has been pointed out, it may be extremely difficult to repeal EU legislation, including when, in the words of its drafters, it failed to deliver.[102]

An important observation for an accurate outline of the SGDR is that only substantial investments in obtaining, verifying or presenting data count. *Created* data do not qualify for protection. After all, it is a database right, not a data right. To remedy this void of protection, a data producer right has been proposed and, at least for the moment, abandoned.[103]

III.3. Is the solution to the problem outside the problem?

A final element in the account of the EU approach to data property and its implications for (AI) technology is placed outside the realm of copyright law and allied rights. The new Public Sector Information (PSI) directive of 2019, also referred to as the Open Data directive regulates the reuse of information held by Public Sector Bodies (PSB).[104] Whereas it would be out of the scope of this article to explore such an important legislative tool in detail, a few specific elements are worth mentioning. First, within the broad principle of re-use by default which has gained more and more strength in the evolution of PSI legislation, the Open Data directive specifically includes research data resulting from public funding under its ambit (Art. 10). This is an important expansion of the scope of the Directive over its predecessors and has a direct impact on the issue of transparency, accountability, and replicability of EU science, contributing to make it a reference at the international level. A second important element of the new Directive relates to the adoption by the Commission (via a future implementing act) of a list of high-value datasets held by public sector bodies and public undertakings to be made available free of

market/en/news/staff-working-document-and-executive-summary-evaluation-directive-969ec-legal-protection.

[101] See ECLI:EU:C:2015:10 Case C-30/14 of 15 January 2015 *Ryanair v PR Aviation*; for a detailed discussion see M. Borghi and S. Karapapa (2013) "Contractual restrictions on lawful use of information: sole-source databases protected by the back door?" European Intellectual Property Review, 37(8), pp. 505-514.

[102] See also Husovec M., The Fundamental Right to Property and the Protection of Investment: How Difficult is it to Repeal New Intellectual Property Rights?, CREATe working paper 2020/02, available at: https://www.create.ac.uk/blog/2020/05/07/new-working-paper-the-fundamental-right-to-property-and-the-protection-of-investment-how-difficult-is-it-to-repeal-new-intellectual-property-rights/.

[103] Hugenholtz, against data right; Yu, Peter K., Data Producer's Right and the Protection of Machine-Generated Data (October 22, 2018). Tulane Law Review, Vol. 93, pp. 859-929, 2019; Drexl J., Designing Competitive Markets for Industrial Data – Between Propertisation and Access, 8 (2017) JIPITEC 257; Zech H., Data as a Tradeable Commodity – Implications for Contract Law, Drexl (Ed.), Proceedings of the 18th EIPIN Congress: The New Data Economy between Data Ownership, Privacy and Safeguarding Competition, Edward Elgar.

[104] Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, repealing Directive 2003/98/EC, as amended by Directive 2013/37/EU.

charge (Art. 14). As the same Commission puts it, "these datasets … have a high commercial potential and can speed up the emergence of value-added EU-wide information products. They will also serve as key data sources for the development of Artificial Intelligence".[105] A final element of the Directive is found in Art. 1(6) and reads: "The right for the maker of a database provided for in Article 7(1) of Directive 96/9/EC", which corresponds to the aforementioned SGDR "shall not be exercised by public sector bodies in order to prevent the re-use of documents or to restrict re-use beyond the limits set by this Directive". The ambit of application of the PSI Open Data Directive is limited to Public Sector Bodies and, since the new Directive, to certain public undertakings. It is perhaps not a purely provocative exercise to consider whether a proper regulatory framework would be one where similar rules in relation to the training of AI should apply generally to any type of data or works.[106] Whereas there would certainly be strong opposition to such a framework, it cannot be accepted that choices affecting both the public and private elements of the life of individuals be made by an AI developed without the guarantees of openness, transparency and accountability. We would not accept laws made by a parliament if it operated in secrecy, we would not accept the determinations of a court of justice or an administrative authority if they were not supported by the reasons exposed in publicly available decisions or through accountable decision-making processes, why should we adopt a different standard when the same decisions are made through the use of a technology that we are only starting to understand?

IV. Conclusions

This paper endeavoured to offer a novel insight into some of the least apparent but far-reaching implications of Arts. 3 and 4 CDSM. In doing so, the paper followed a double approach. In the first instance, it focused on the legal changes affecting the regulation of mere facts and data brought by the new CDSM, critically assessing their fitness for the task. Subsequently, the paper also attempted to offer a complementary conceptual and normative reading of the copyright theory behind a TDM exception. In this process, the paper pointed out that Arts. 3 and 4 are agnostic to any theoretical element, an aspect we term "theory-less law making" which is recurrent in EU copyright law. The paper also attempted to identify some of the reasons (seeing digitisation as a threat) at the basis of this approach. The paper argues that technology is not exogenous to

[105] See https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information (last visited 1 July 2021).

[106] For a competition law argument supporting a possible obligation to open privately held databases in cases of anticompetitive behaviours, see Drexl, Designing Competitive Markets for Industrial Data – Between Propertisation and Access, 8 (2017) JIPITEC 257.

(copyright) law,[107] on the contrary law and technology are in a dialogic relationship constantly shaping and being shaped by each other. This intimate relationship with the law has an often decisive impact on how technology will evolve, not only in strictly technological terms, but also in the sense of how that technology will be governed, who will have access to it, at what costs and under which conditions.[108] When this technology is AI, with its endless potential applications, the risk of getting it wrong is likewise immense. In other words, paraphrasing a famous expression, digital artefacts have politics, and AI perhaps more than others.[109] Ignoring this dimension, as in the CDSM, won't solve the problem.

An overall assessment of the situation portrayed in this paper cannot be optimistic. Whereas a good amount of attention in scholarship has been (rightly) dedicated to critically evaluate recent proposals to create a data producer right, this paper shows that the EU legislator, probably even beyond its own intentions, has taken a very drastic position on a complementary and highly relevant matter, the ownership of *pre-existing* mere facts and data. As demonstrated, this position is not functional to a proportionate, fair, and accountable regulatory framework for copyright, for technology and for the EU as an economic, social and political institution.

Is this the end of the story for the protection of mere facts and data contained in protected works and of the connected far-reaching technological, cultural and innovation policy implications? Or are there other areas that could be explored further and that could possibly offer some prospect for a balanced, proportionate and theory-based EU copyright law? There seem to be at least three levels where some residual "flexibility" may still be found and which will form the basis of future investigation. There is an EU level, an EU Member States level and an extra-EU level.

At the EU level, further work should delve into a clearer and standard interpretation of the conditions of the exception for temporary copies under InfoSoc Art. 5(1). The position of the CJEU seems ambivalent, stating – sometimes within two consecutive paragraphs – that the exception for temporary copies must be interpreted narrowly as it deviates from the general rule; and that the function of 5(1) is to ensure not only users' *rights* but also to allow technological development. Clarity in this area is crucial and for the reasons exposed above, such clarity should be in the sense that Art. 5(1) serves a dual function: it protects users' rights and it allows an open and accountable development of technology. This route seems to be even more essential in

[107] In this sense see Benkler Y., The Role of Technology in Political Economy, LPE Blog July 2018, available at: https://lpeblog.org/2018/07/25/the-role-of-technology-in-political-economy-part-1/.
[108] Id.; Benkler, Power and Productivity: Institutions, Ideology, and Technology in Political Economy (December 2019), available at: http://www.benkler.org/Benkler_Power&Productivity.pdf.
[109] Winner L., *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, Daedalus 109 (1980): 121-36.

the light of recent CJEU case law that appears to establish that any fundamental rights limitation to copyright has to be found within Art. 5.[110]

At the Member States level, a main source of potential flexibility has traditionally been the right of adaptation, the only major economic right not yet object of horizontal legislative harmonising interventions.[111] Despite some initial doubt, the CJEU clarified that the right of adaptation is not harmonised. However, reproductions are and in the light of cases such as *Allposter*[112] and *Pelham*[113], it seems that the space for MS to regulate autonomously an adaptation right (including its limits and exceptions) has shrunk considerably. And yet, it seems that the fundamental function of so-called transformative uses comfortably resides within a right that perhaps more than others determines the external boundary of how far copyright law can and should extend.[114] MS interested in enabling computational uses should consider this option.

The Open Data Directive briefly discussed above and especially the national Open Access guidelines it mandates will likewise represent a fundamental area of intervention to ensure that research data held by public sector bodies fuels innovation. The opportunity to extend similar obligations also to privately held databases seems an essential condition to develop open, transparent and accountable AI. No AI trained on unverifiable data, i.e., "black box" AI, should be used by public authorities. Arguably there seems to be a timid indication of this also in the recent AI Regulation proposal.

Finally, extra EU countries which are not bound by the rigidity of EU law in this area, can be divided in two main categories. Those who have enacted a broad and/or flexible approach (US, Canada, Singapore, South Korea, Japan, Israel[115]), and those who have not yet done so. In the light of the above, a technology enabling exception, or a computational uses provision appears as one of the most urgent additions to national copyright laws that countries concerned with cultural and technological sovereignty should pursue. For the UK which was bound by the InfoSoc Directive until very recently (and will follow the "old" rule until domestic law changes[116]), the future seems a choice between the need to maintain a level playing field with the EU neighbour and

[110] See ECLI:EU:C:2019:624 Case C-476/17, of 29 July 2019 *Pelham v Hütter*; Senftleben M., Flexibility Grave – Partial Reproduction Focus and Closed System Fetishism in CJEU, Pelham, *IIC* 51, 751–769 (2020).
[111] Hugenholtz and Senftleben, Fair Use in Europe: In Search of Flexibilities, Amsterdam Law School Research Paper No. 2012-39, Institute for Information Law Research Paper No. 2012-33.
[112] ECLI:EU:C:2015:27 Case C-419/13, of 22 January 2015, *Allposters v Pictoright*.
[113] ECLI:EU:C:2019:624 Case C-476/17, of 29 July 2019 *Pelham v Hütter*.
[114] Margoni T. (2015) The digitisation of cultural heritage: originality, derivative works and (non) original photographs, available at ttps://ssrn.com/abstract=2573104.
[115] Flynn et al. (2020) Implementing user rights for research in the field of artificial intelligence: a call for international action, European Intellectual Property Review EIPR 42(7), pp. 393 – 398.
[116] Kretschmer M. (2020) UK sovereignty: A challenge for the creative industries, available at: https://www.create.ac.uk/blog/2020/07/21/uk-sovereignty-a-challenge-for-the-creative-industries.

the attractiveness of regulatory competition and of a modern, dynamic and accountable regulation of AI. Perhaps, an updated TDM exception (Section 29A, UK Copyright, Designs and Patents Act 1988) not limited to non-commercial uses and not limited to certain rights or to certain sources may nudge the EU copyright legislator to escape the technological determinism of the CSDM Directive.

CREATe

UK Copyright and Creative Economy Centre
School of Law / University of Glasgow
10 The Square
Glasgow G12 8QQ
www.create.ac.uk

2021/7 DOI: 10.5281/zenodo.5082012

In collaboration with:

University of Glasgow · KU LEUVEN CiTiP · ReCreating Europe