ReCreating Europe

Rethinking digital copyright law for a culturally diverse, accessible, creative Europe

Grant Agreement No. 870626

Deliverable Title	**D3.7 Final report on the role of EU copyright law in relation to training models for machine learning purposes.**
Deliverable Lead:	CREATe, University of Glasgow
Partner(s) involved:	
Related Work Package:	WP3 - Authors and performers
Related Task/Subtask:	T3.3 AI, machine learning and EU copyright law
Main Author(s):	Martin Kretschmer, Thomas Margoni (CiTiP KU Leuven & CREATe), Pinar Oruc
Other Author(s):	
Dissemination Level:	Public
Due Delivery Date:	30.06.2022
Actual Delivery:	21.10.2022
Project ID	870626
Instrument:	H2020-SC6-GOVERNANCE-2019
Start Date of Project:	01.01.2020
Duration:	36 months

Version history table

Version	Date	Modification reason	Modifier(s)
v.01	24.06.2022	Full draft	Martin Kretschmer, Thomas Margoni
v.02		Formatting	Thomas Margoni, Martin Kretschmer

Legal Disclaimer

The information in this document is provided "as is", and no guarantee or warranty is given that the information is fit for any particular purpose. The above referenced consortium members shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

Table of Contents

List of Figures

Abbreviation list

AI	Artificial Intelligence
API	Application Programming Interface
CDSM	Copyright in Digital Single Market (Directive (EU) 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market
CJEU	Court of Justice of the European Union
CNN	Convolutional neural network
DMA	Digital Markets Act (Proposal for a Regulation on contestable and fair markets in the digital sector Brussels, 15.12.2020 COM(2020) 842 final, adopted in 2022)
GAN	Generative adversarial network
ISD	Information Society Directive (Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society)
ML	Machine learning
PSI	Public Sector Information
TDM	Text and Data Mining
TPM	Technological Protection Measures

Executive summary

There is global attention on new data analytic methods. Artificial Intelligence (AI) is seen as a critical technology, often relying on machine learning (where an algorithm is trained on data to recognise and predict patterns). Data scraping, the acquiring and structuring of information from online sources, is a typical first step for machine learning.

The technologies of scraping, mining and learning are often conflated, as are the legal regimes under which they are regulated. One regulatory lever under one legal regime will not deliver policy aims, such as innovation, personal dignity, Open Science, or the currently popular 'data sovereignty'. The legal issues involved in the governance of data range from proprietary approaches (copyright, database rights) to privacy and data protection.

In addition, there are a wide range of public law instruments, for example relating to public sector data governance[1], access to and use of user facilitated data[2] or the right to non-discrimination.[3] Competition law again (which may be both privately and publicly enforceable) increasingly prescribes conduct in relation to data, such as in merger or acquisition cases, or in transparency provisions (Art. 17 CDSM; and centrally in the proposed DMA and AI Regulation).

The scope of our enquiry in this report is within private law, specifically on the attempt to assert quasi-proprietary control of information and data, or vice versa limit such attempts, for example by exempting desired activities via copyright exceptions, such as the exception for text and data mining in Arts. 3 and 4 CDSM.

The copyright regime offers a template with a centuries old tradition of exclusive rights, supplemented in the EU since 1996 by a *sui generis* database right.[4] While data or information are not subject matter within copyright law, almost all materials used to construct so-called corpora for new data analytic methods are protected by copyright law: scientific papers, images, videos, and so on.

The research design we adopt for this study is a reverse inductive strategy. We focus on case studies of three technological processes to explore in detail possible descriptions that would allow legal analysis, and an assessment of the need for a harmonisation of rights and connected exceptions under copyright law.

The case studies were selected in consultation with stakeholders, reflecting a need by scientific researchers and technology companies for a better legal understanding of what they do. They are designed to reflect a range of techniques and processes that underpin advanced data analytics, responding to the EU policy objective of supporting innovation in this field.

The three case studies are:

(1) Data scraping for scientific purposes
(2) Machine learning, in the context of Natural Language Processing (NLP)
(3) Computer vision, in the context of content moderation of images

[1] Directive 2019/1024/EU of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information

[2] Proposal for a Regulation of the European Parliament and of The Council on harmonised rules on fair access to and use of data (Data Act), Brussels, 23.2.2022 COM(2022) 68 final

[3] Charter of Fundamental Rights of the European Union Art.21, as reflected in the Proposal for a Regulation laying down harmonised rules on artificial intelligence

[4] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases

In parallel, we offer a thorough analysis of the policy rationale and legal context for the introduction of the two exceptions for text and data mining in the CDSM Directive (Art. 3 Text and data mining for the purposes of scientific research; Art. 4 Exception or limitation for text and data mining) which includes an analysis of how the right of reproduction (Art. 2 ISD) and its limitations (mainly Art. 5(1) ISD) interface with the overall regulatory framework of data analytics. This part of the report was written as a self-contained scientific paper (authored jointly by Thomas Margoni and Martin Kretschmer) and has been published in the peer reviewed journal GRUR International (Journal of European and International IP Law) under a CC-BY license and is available at this link https://academic.oup.com/grurint/article/71/8/685/6650009 . The paper is entitled "A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology". It is appended to this Report as Annex A.

We maintain a resource page that will be regularly updated with project results (workshops and outputs): https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/

1. Methodology

Legal research on data analytic techniques typically starts with an identification of relevant legal regimes and proceeds to doctrinal analysis of the scope of certain concepts and rules. The analysis is then evaluated against practical implications, often using particular factual constructions (scenarios) to illuminate potential effects of interpretations or interventions.

There are dangers to this legal approach to policy making. The analysis often lags technological developments. Scenarios may be filtered via professional representations or trade bodies that were constituted in a different context, perpetuating past discussion. In a wider sense, policy making may be anecdotally driven, by examples that surface through lobbying processes.

The research design adopted for this project aims to reverse this direction of travel. We adopt an inductive approach, attempting to get close to the "real world" of data analytics. Through a detailed empirical description of a selection of cases (in a social science sense) we seek to explore legal issues that are implicated.

The selection of sites for case analysis poses its own generalisability challenge. In case study research, we need to reflect on why selected empirical settings are more or less reflective of the phenomenon under investigation, i.e. rapidly evolving data analytic technologies.[5] In consultation with scientific researchers and technology companies, we identified three case studies that together reflect a range of techniques and processes that underpin advanced data analytics. The selection takes account of the EU policy objective to support innovation in this field.[6]

The three selected cases are:

(1) Data scraping for scientific purposes;
(2) Machine learning, in the context of Natural Language Processing (NLP);

[5] For a classic account of the selection problem in case study research, see Seawright J. and Gerring J. (2008) Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options, Political Research Quarterly 61(2):294-308. doi:10.1177/1065912907313077

[6] Cf. Recital 8, CDSM Directive.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

(3) Computer vision, in the context of content moderation of images.

In researching the cases in a legal context, there is a further tension between an unstructured approach that offers rich descriptions inductively from multiple sources (such as public documents, observations, conferences, or interviews) and the need to capture the empirical world in a form recognisable for subsequent legal analysis. In Law, this challenge of "fact-finding" is discussed under the concept of evidence.[7] In legal disputes, there is an assumption that a representation of facts can be settled (typically in first instance cases). It is then the application of rules to the facts that can be the subject of appeals. The case studies presented in this report offer such a possible description of facts that will aid the development of legal analysis and policy recommendations.

Legal implications are addressed in a self-contained article (attached as an Appendix) focussing on the introduction of the exceptions for text and data mining in the CDSM Directive (Art. 3 Text and data mining for the purposes of scientific research; Art. 4 Exception or limitation for text and data mining). What was the policy problem the interventions sought to solve? Which were the legal hurdles to the development of certain data analytic technologies in the EU that needed to be addressed? What are unresolved issues and shortcomings of the legal approach chosen in the CDSM Directive?

We highlight:

- that the definition of text and data mining may be too broad, making the entire field of data-driven AI development dependent on exceptions;
- that the scope of the exceptions is limited to the right of reproduction;
- that the limitation of the Art. 3 to certain beneficiaries remains problematic;
- that the requirement of lawful access is difficult to operationalize;
- that rightholders *de facto* can override the exceptions by technological interventions.

2. Cases for the study of copyright and *training data* in selected AI environments

These case studies have been prepared by Dr Pinar Oruc, under the supervision of Profs. Martin Kretschmer and Thomas Margoni.

2.1 Introduction to the case studies and delimitation of the area of enquiry

Copyright law has a direct impact on the processes of data scraping, mining and learning.[8] So called "corpora", i.e. collection of information needed for "training" purposes could include works protected by copyright, other related subject matter, or simple facts and data. When copyright or a related right are present, any digital copy, temporary or permanent, in whole or in part, direct or indirect, has the potential to infringe that right, in particular the economic right of reproduction. Furthermore, the changes made in the collected material can amount to an 'adaptation' within the scope of the exclusive right. The relevant exceptions, such as for research or text and data mining, might not sufficiently cover the activities of the researchers and firms in this area. This report presents three case studies to provide an in-depth exploration of the complex

[7] Ho, Hock Lai (2015) The Legal Concept of Evidence, Stanford Encyclopedia of Philosophy, available at <https://plato.stanford.edu/entries/evidence-legal/> (accessed 1 June 2021)

[8] For an analysis of the link between EU copyright law and machine learning form an input data point of view, see Margoni, Thomas (2018), Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?, AIDA (Annali Italiani del Diritto d'Autore della Cultura e dello Spettacolo); 2018; Vol. XXVII; iss. 1; pp. 281 – 304 available at: https://zenodo.org/record/2001763.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

technological processes involved in some of the most popular AI applications. The results of the case analysis will be functional to a proper legal classification and assessment of the relevant regulatory framework.

Three different case settings were selected: web scraping, natural language processing, computer vision. The case studies rely on publicly available sources (e.g. published scholarly analysis, official information issued by companies for instance on their websites and policy documents) and expert feedback.[9] Within the cases, the unit of analysis for comparison is the technological process.

2.1.1 Terminology

This part introduces the technological distinctions that will be employed in the report.
The first distinction relates to data collection methods. "Scraping" involves manually or automatically collecting data from websites. There are different kinds of such data collection, such as web scraping, web harvesting and web crawling, which will be addressed specifically under Case Study 1.

It is useful to clarify that data scraping and data mining are not the same, but the terms are sometimes used interchangeably in the literature. Data scraping is the collection/extraction of the necessary information to build a set of data. Data mining, which does not necessarily include data collection, is the analysis of these datasets and sometimes this analysis requires machine learning to reach more complex purposes. Case Study 1 uses the term 'scraping' but will also touch on the stages for mining and outputs.

A second distinction relates to the type of machine learning, addressed in Case Studies 2 and 3. Supervised learning uses training data labelled by humans. In unsupervised learning, the algorithm uses unlabelled data and detects similarities and patterns. In reinforcement learning, the algorithm relies on trial and error to reach the maximum reward for an activity.

There are also more specific technological developments that might be relevant across the case studies. Deep learning is a form of machine learning (ML), where multiple artificial neural networks[10] carry and interpret complex raw data. With more layers, it becomes more likely to solve complex problems but it also means less clarity on why the AI system decides one way over the other, which reduces the accountability. Although neural networks have been proposed as early as 1943, the research on neural networks and the use of deep neural networks have increased in recent years with the availability of cheaper computational power and resources.[11]

Generative adversarial networks (GAN) have two deep learning networks (one generator and one discriminator) and they learn by competing with each other.[12] GANs can be supervised and unsupervised.
There is also "transfer learning", which is not a ML technique but a way to design a research methodology where there is not enough training data. A pre-trained model for a similar task is taken and adopted into the project at hand.[13] It is used for both NLP and Computer Vision.

[9] As part of the validation process for the case studies, a workshop was held at the University of Glasgow on 27 May 2021. It is documented here:<https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning/>

[10] "The network is a connected framework of many functions (neurons) working together to process multiple data inputs. The network is generally organized in successive layers of functions, each layer using the output of the previous one as an input." WIPO Technology Trends 2019 — Artificial Intelligence, WIPO, 2019, <https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf> 146; See also OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, Adopted on 22/05/2019.

[11] Seifert et al, Visualizations of Deep Neural Networks in Computer Vision: A Survey' in Tania Cerquitelli, Daniele Quercia and Frank Pasquale. *Transparent Data Mining in Big and Small Data* (Springer 2017) 123.

[12] Arthur I. Miller, *The Artist in the Machine: The World of AI-Powered Creativity* (MIT Press 2019) Chapter 10.

[13] Niklas Donge, 'What is Transfer Learning? Exploring the Popular Deep Learning Approach' (2020) https://builtin.com/data-science/transfer-learning , Orhan G Yalcin, '4 Pre-Trained CNN Models to Use for Computer

How to categorise the stages in the process? When categorising these activities for the purpose of our legal analysis, it becomes apparent that the stages are not the same for data analysis and ML training.

For scraping, our legal analysis would require focusing on the data collection and data processing stages. There is no annotation or training, but there are outputs based on data analysis. (3 stages)
For machine learning projects, researchers usually start with defining the problem, choosing the data sources and algorithms, and the trained model will then be released (deployment stage).[14] For the stages in between, both Natural Language Processing (NLP) and Computer Vision stages are to be categorised similarly: data collection (which can also be achieved through scraping), data processing, training (different process if it is supervised or unsupervised) and then the output stage – depending on what the algorithm is for, such as language understanding or audiovisual content moderation. (4 stages)

Accordingly, data scraping as such may be likewise seen as a form of data collection that then leads to different possibilities of data analytics, including those identified in Case Studies 2 and 3, therefore as a preliminary step for Natural Language Processing and Object Recognition. However, given its relevance and complexity we decided to offer it as a self-standing case, seeking feedback during expert consultations.

2.2. Data Scraping

Scraping involves manually or automatically collecting data from websites. Screen scraping involves scraping the data that is displayed on users' screens. Web-scraping or web-harvesting is collecting all underlying data from a website, including website scripts. Web-crawling is "accessing web content and indexing it via hyperlinks; thus, only the URL but no specific information is extracted".[15]

There are multiple ways of categorising scraping tasks: (1) accessing the web pages, (2) finding specified data elements, (3) extracting them, (4) transforming them and (5) saving these as a structured data set.[16] Alternatively, they can also be further divided: '(1) information identification, (2) choice of strategy, (3) data retrieval, (4) information extraction, (5) data preparation, (6) data validation, (7) debugging and maintenance, (8) generalisation'.[17]

For our purposes, we will merge some stages that are similar for legal analysis: (1) data collection, (2) data processing and, (3) data analysis and outputs.

Vision with Transfer Learning' (2020) https://towardsdatascience.com/4-pre-trained-cnn-models-to-use-for-computer-vision-with-transfer-learning-885cb1b2dfc.

[14] Richmond Alake, '10 Stages Of A Machine Learning Project In 2020 (And Where You Fit)' (2020) https://towardsdatascience.com/10-stages-of-a-machine-learning-project-in-2020-and-where-you-fit-cb73ad4726cb.

[15] Fiona Campbell, 'Data scraping - what are the privacy implications?' (2019) Privacy & Data Protection 20(1), Frank Jennings and John Yates, 'Scrapping over data: are the data scrapers' days numbered?' (2009) JIPLP 4(2) 120, Judith Hillen, 'Web scraping for food price research' (2019) British Food Journal 121(2) 3350.

[16] Geoff Boeing and Paul Waddell, 'New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings' (2017) Journal of Planning Education and Research 37(1) 457, 459.

[17] Tasks are identified as '(1) information identification, (2) choice of strategy, (3) data retrieval, (4) information extraction, (5) data preparation, (6) data validation, (7) debugging and maintenance, (8) generalisation' Simon Munzert, Christian Rubba, Peter Meißner and Dominic Nyhuis. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (John Wiley & Sons 2015).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

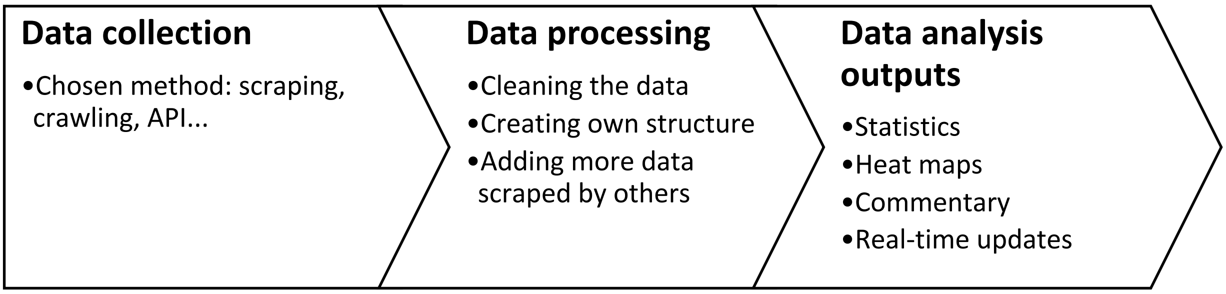

Figure 1: Data scraping workflow

This part will rely on the example of a property website (such as AirBnB) being scraped for research on short/long-term letting market effects.

2.2.1. Collection stage

Most property listing websites such as AirBnB do not create a new page for every listing. Instead, a template exists, and it is automatically filled with data for that specific property as entered by the users/property hosts. The data available on the website includes property descriptions, user reviews, photographs of the property (only saved as hyperlinks), location, longitude and latitude of the property, neighbourhood ID, available dates, maximum and minimum price, place type and number of guests, user scores.

If the collection purpose is unknown, it is often useful and possible to collect all available information. At this stage, no distinction may be made whether particular data is created by AirBnB or uploaded by the property hosts.

As introduced before, screen scraping is limited to what is available to the visitors, web-harvesting targets and collects all data, and web-crawling follows and indexes all links (those can be visited and scraped). Since scraping relies on how data is displayed, even the small changes in the display of the website can disrupt the collection stage.[18]

Another method of making the process faster and more efficient is using the application programming interface (API) scraping. The stages of using an API for data collection can be summarised as follows: (i) finding the API and becoming familiar, (ii) registering for API use and retrieving keys, (iii) calling the API to collect data and then (iv) processing the data.[19] API scraping does not mean access to previously inaccessible data, but it speeds up the process by circumventing the rendering stage.

While it makes it easier for scrapers, APIs require substantial resources for hosts to develop and maintain'.[20] In fact, some websites do not make their API openly available to stop the competitors from scraping data from them in order to ensure their competitive edge remains.[21] Although not directly competitors, any researchers interested in this data and unable to collect it via the API, then have to come up with their own strategies and/or rely on different scraping strategies.

[18] Jeffrey Hirschey 'Symbiotic Relationships: Pragmatic Acceptance of Data Scraping' (2014) Berkeley Technology Law Journal, Vol. 29, 906

[19] Simon Munzert, Christian Rubba, Peter Meißner and Dominic Nyhuis. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining* (John Wiley & Sons 2015) 277

[20] Brett Massimino, 'Accessing Online Data: Web-Crawling and Information-Scraping Techniques to Automate the Assembly of Research Data' (2016) Journal of Business Logistics 37(1) 34.

[21] Janet Williams, 'Web Scraping better than a Data API' https://www.promptcloud.com/blog/web-scraping-better-alternative-to-api/

In the example of AirBnB, their API is not openly available to the general public, but may be requested by developers, certain groups of users, such as hosts wanting to use their own interface to add multiple listings at once[22] or external partners such as travel companies and Groupon.[23]

In addition to the concerns about loss of control over the data and its devaluation reported by website operators, they have to make sure that the scraping does not cause system overload.[24] Examples in this sense are blocks of excessive requests from the same IP range to ensure stability of the servers. Since data hosts can detect unusually high or repetitive tasks from the same IP address (even easier to detect if it comes from the same user account, if the scraping is performed after the login page), the scrapers usually use proxies to distribute their requests to avoid exceeding this threshold and being blocked.[25]

Additionally, as common practice many websites have terms and conditions that restrict the collection and analysis of their data. Under the AirBnB Terms of Service (both for European Users and non-European Users), there are terms that limit the ways and purposes of using the platform. Under 12.1 of their Terms of Service, the following is not allowed: "scraping, hacking, reverse engineering, compromising or impairing the platform, using bots, crawlers, scrapers or other automated means, attempts to circumvent any security or technological measure, taking any action that could damage or adversely affect the performance or proper functioning of the platform". Furthermore, the Content cannot be used without the permission of Content owner and can only be used as necessary to enable to use of the website as a Guest or Host.[26]

2.2.2. Processing stage:

After the targeted data is collected, it is then structured in a manner that is more suitable for the upcoming data analysis. Researchers typically store this raw data in a way that is structured by them, that is more in line with their research purposes and internal structure. As the computational power and storage costs are constantly getting more effective, it is suggested that scrapers are now able to scrape more data and can choose to be less conservative.[27] But that also means holding more data to be filtered and cleaned.

As the property information in this example are added by the users, it can be messy and the researchers might have to go through substantial wrangling and validation to make the data usable.[28] It requires identifying and removing duplicate listings (by relying on things as the Property ID, location and the size of the property) or identifying other mistakes such as typos in the rental price.[29] As part of validation, the researchers have to ensure that the new data is reliable and usable for their purposes. Depending on the purpose of each research output, the necessary data is then pulled from these databases.

[22] https://www.airbnb.co.uk/partner

[23] Ingrid Lunden, 'Airbnb eyes expansion with affiliate program for sites with 1M+ users, new API' (2017) https://techcrunch.com/2017/10/16/airbnb-eyes-expansion-with-affiliate-program-for-sites-with-1m-users-new-api/

[24] Frank Jennings and John Yates, 'Scrapping over data: are the data scrapers' days numbered?' (2009) JIPLP 4(2) 120

[25] Jeffrey Hirschey 'Symbiotic Relationships: Pragmatic Acceptance of Data Scraping' (2014) Berkeley Technology Law Journal, Vol. 29, 918; Manthan Koolwal, '10 Tips to avoid getting Blocked while Scraping Websites' (2020) https://www.codementor.io/@scrapingdog/10-tips-to-avoid-getting-blocked-while-scraping-websites-16papipe62

[26] https://www.airbnb.co.uk/help/article/2908/terms-of-service#EU12

[27] Judith Hillen, 'Web scraping for food price research' (2019) British Food Journal 121(2) 3350, 3354, Zachary Gold and Mark Latonero, 'Robots Welcome: Ethical and Legal Considerations for Web Crawling and Scraping' (2018) 13 Wash J L Tech & Arts 275, 281.

[28] Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining: Concepts and Techniques* (2012) The Morgan Kaufmann Series in Data Management System, 592.

[29] Geoff Boeing and Paul Waddell, 'New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings' (2017) Journal of Planning Education and Research 37(1) 457, 460.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

It is also possible to enrich the scraped data with the data from other sources. For example, there are websites and analytics companies based in the United States that collect and aggregate AirBnB data, such as AirDNA and SmartHost, to guide the hosts and nearby businesses. These are not prevented to do so by AirBnB so far. There are also US sources that provide scraped data together with own analysis, such as Tom Slee (tomslee.net) and InsideAirBnB (insideairbnb.com).[30] Researchers, in both inside and outside United States, often rely on such scraped datasets, commentary and research outputs by such third parties.

2.2.3 Analysis and outputs stage

The collected data can be one-off and shows the exact situation at a certain time of it can allow real-time updates (such as price comparison websites).[31] It is up to the researcher to choose which data will be collected and analysed, to solve the problem at hand.

The results of the analysis are then shared in formats chosen by the researcher (such as reports, journal articles, heat maps or blog posts). The extent of the data used in these outputs is determined case by case. These outputs are not a replacement of the website; however they can convince policymakers about changes that indirectly affect websites like AirBnB.

Restructured datasets based on the scraped data may or may not be shared with other researchers. Parties might contact AirBnB for permission.

There is a growing body of academic literature based on AirBnB. A wide range of issues are addressed, such as the extent to which neighbourhoods are vulnerable to the switch from long-term letting to short-term letting.

2.3 Natural Language Processing

Natural language processing (NLP) is in the intersection of computer science/AI and linguistics. It is a form of machine learning where the purposes can range from analysing larger texts to computers generating realistic texts. The applications of NLP include information extraction, machine translation, natural language generation and sentiment analysis.[32]

NLP can be supervised or unsupervised. Supervised learning requires labelled/tagged text data, so they have an "annotation" stage in their workflow. On the other hand, unsupervised NLP uses unlabelled data and instead detects patterns, but it requires large datasets to achieve that and is therefore not suitable for all research projects. If some labels are from humans and others are not, then it will be classified as semi-supervised machine learning – which is useful for projects holding small annotated datasets together with large amount of raw data found online.[33]

[30] "Other companies include, but are not limited to, Beyond Pricing (beyondpricing.com), SmartHost (smarthost.co.uk), Everbooked (www.everbooked.com) and PriceLabs (www.pricelabs.co)". Teresa Scassa, 'Ownership and control over publicly accessible platform data' (2019) Online Information Review 43(6) 986, 991.

[31] "Once the script is written, it is up to the user whether it should run and extract prices and other data monthly, weekly, daily or even at a higher frequency" Judith Hillen, 'Web scraping for food price research' (2019) British Food Journal 121(2) 3350, 3353.

[32] WIPO Technology Trends 2019 — Artificial Intelligence, WIPO, 2019, https://www.wipo.int/edocs/pubdocs/en/wipo_pub_1055.pdf

[33] Ben Dickson, 'What is semi-supervised machine learning?' (2021) https://bdtechtalks.com/2021/01/04/semi-supervised-machine-learning/

NLP research focuses on achieving and improving various tasks.[34] Some tasks have direct applications, such as translation or summarisation. Other tasks such as segmentation or named entity recognition are used to inform other tasks and turn the texts into machine-readable data.

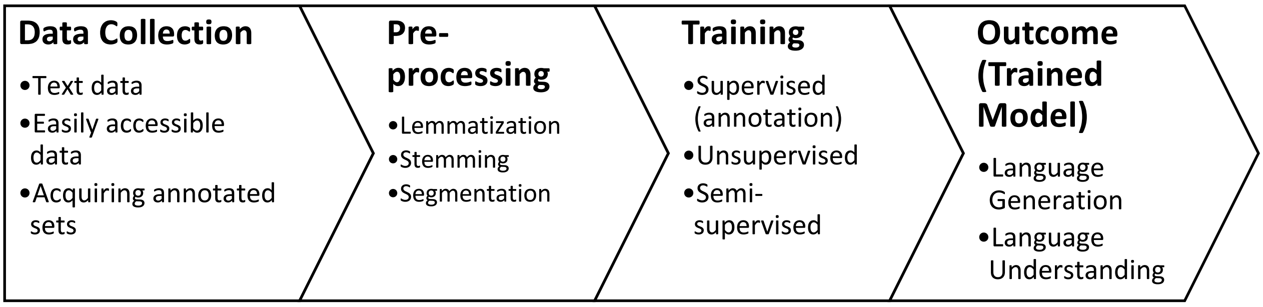

Figure 2: NLP workflow

2.3.1 Collection Stage

The first step for Natural Language Processing is the compilation of the necessary data. The data can come from anywhere, ranging from user comments to ancient philosopher corpus.

The data collection stage is similar to the scraping process described above: the necessary data is identified in line with the research purpose and then it will be targeted with the appropriate data collection methods (scraping or otherwise).

It is also possible to find freely available datasets online, such as the books from Project Gutenberg or the Spoken Wikipedia Corpora – depending on the task at hand.[35] The NLP researcher can also choose to focus on licensed corpora[36] or scholarly literature held in databases that they have access to.[37]

While it can be possible to build the models in a way to reduce access to data and keep it as temporary as possible, it would both be impractical and very straining on the resources of third parties.

2.3.2. Pre-processing

The data then goes through pre-processing. This part involves different tasks to understand the texts.[38] The collected material goes through some changes at this stage, which will be important for the legal analysis later.

First, formats such as PDF of MS Word need to be converted into text for the NLP tasks that follow.[39]

[34] For an overview of different tasks in NLP: https://paperswithcode.com/area/natural-language-processing

[35] ODSC - Open Data Science, '20 Open Datasets for Natural Language Processing' (2019) https://medium.com/@ODSC/20-open-datasets-for-natural-language-processing-538fbfaf8e38 ; Jason Brownlee, 'Datasets for Natural Language Processing' (2020) https://machinelearningmastery.com/datasets-natural-language-processing

[36] Eckart de Castilho et al, 'A Legal Perspective on Training Models for Natural Language Processing' (2018)

[37] Przybyła et al, 'Text mining resources for the life sciences' (2016) Database, Volume 2016, 2016, baw145

[38] Mirantha Jayathilaka, '25 NLP tasks at a glance' (2020) https://medium.com/@miranthaj/25-nlp-tasks-at-a-glance-52e3fdff32e2; '.Natural Language Processing (NLP) Guide – What Is NLP & How Does it Work?' https://monkeylearn.com/natural-language-processing/

[39] Bruce H Cottman, 'Converting PDF and Gutenberg Document Formats into Text: Natural Language Processing in Production' (2020) https://towardsdatascience.com/natural-language-processing-in-production-converting-pdf-and-gutenberg-document-formats-into-text-9e7cd3046b33

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

Tokenization is when text is separated into smaller units in a way that can be read by the machine. This smaller unit can be word pieces or characters.[40] Parts of speech (POS) tagging is when words are tagged as noun, verb, or prepositions.

Normalization is when a more normalized version of the text is created by removing variations that are not important for the final research target. Through normalization, the text becomes more standard and is easier for the machines to "read". It includes tasks such as lemmatization, stemming or spelling correction, which all change the text.[41] Stemming removes the end of the word, while lemmatization changes the word into its base or dictionary form.[42] Such tasks are sometimes performed by an algorithm, but humans can be consulted as well, at least while developing these methods or applying it to a new application domains.

2.3.3. Training:

As mentioned earlier, the stage after pre-processing then differ according to the type of the learning.

(a) Supervised:

If the project relies on supervised learning, then pre-processed data is annotated by humans and the human input then helps the development of AI. The data that was previously unreadable to the machine becomes something usable through the annotation stage.

During the annotation process, it is possible to both add the annotations to the original text or create a separate file for annotations.[43] The former has the advantage of keeping both the text and annotations in a single file – such as XML file – so the NLP algorithms have access to both.

(b) Unsupervised:

If unsupervised, then learning requires no human input once the data is collected. There is no annotation stage. The project could involve multiple tasks that support each other by creating annotations, but as long as the NLP rely only on pre-trained models and the final task does not involve humans, it would still count as unsupervised training.

Although unsupervised learning is possible and is a growing field in NLP, it is also not widely accessible to smaller groups due to the need for computer power and large amount of data. Companies that have such resources, such as Google or OpenAI, use it to create pre-trained models. As long as they make these models available, smaller projects can then obtain these pre-trained models and use it on their datasets.

Use of embeddings and language models

Pre-trained embeddings and models mentioned here are trained on a large corpus in an unsupervised manner (by Google, OpenAI and similar companies with such resources), then fine-tuned in a supervised manner.[44] These are then made available for other users, so that they can use them to support their other supervised and semi-supervised learning projects. This means that as long as these pre-trained versions are available, other researchers can skip some stages or reap the benefits of the collection and pre-processing done by other companies. But this also creates a monopoly over language modelling.[45]

[40] 'Tokenization' <https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html>

[41] Dan Jurafsky and James H. Martin, *Speech and Language Processing* (3rd edn, 2020) https://web.stanford.edu/~jurafsky/slp3/; Tiago Duque, 'Text Normalization' (2020) https://towardsdatascience.com/text-normalization-7ecc8e084e31

[42] 'Stemming and lemmatization' https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[43] Przybyła et al, 'Text mining resources for the life sciences' (2016) Database, Volume 2016, 2016, baw145.

[44] Edward Ma, 'Combining supervised learning and unsupervised learning to improve word vectors: Introduction to Generative Pre-Training' (2019) 'https://towardsdatascience.com/combining-supervised-learning-and-unsupervised-learning-to-improve-word-vectors-d4dea84ec36b ; OpenAI, ' Improving Language Understanding with Unsupervised Learning' (2018) https://openai.com/blog/language-unsupervised/

[45] Taylor Soper, ''OpenAI should be renamed ClosedAI': Reaction to Microsoft's exclusive license of OpenAI's GPT-3' https://www.geekwire.com/2020/openai-renamed-closedai-reaction-microsofts-exclusive-license-openais-gpt-3/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

The paragraphs below will explain where embeddings and models sit within the developments of NLP. It is useful to take such developments into consideration for our legal analysis, as the approaches determine the amount and type of data that is used and the parties' involvement.

- In earlier NLP projects, 'bag of words' approach assigns a unique token to words, so a text can be displayed in numbers. While transforming words to numerical representations (vectors), the basic method is to count how many times a word occurs in a text, without paying attention to the order of the words. Since this approach would determine words like "the" or "is" as the most common and therefore the most important, the weights of the words need a separate adjustment (TF-IDF encoding).[46] N-grams extracts a consecutive n-number of words from the text to analyse.[47] These methods are still used, but are now supported by the others below.
- Word embeddings (2013 onwards): Embedding models mean giving vectors that show the connection between words. This allows the machines to understand which words go together, which helps in tasks like prediction or translation. There are word embedding models like *word2vec* (by Google) and *GloVe* (by Stanford).

The researchers then have the option of either (i) relying on in pre-trained word embeddings (based on the training done by their developers) such as *word2vec* trained on Google News corpus[48] or (ii) train the embeddings themselves to make sure that it assigns numerical values based on their specific dataset/research topic - so that it can be used on later NLP tasks with greater accuracy.

Since the first option is trained on generic texts, they are not overly helpful for using on very specialist texts, for example legal documents.[49] This means that researchers of specific topics still might prefer to train their own word-embedding models with their own training data.

The fact that pre-trained embeddings rely on easily found text material also leads to bias problems. For example, it was determined that *word2vec* carries the same the sexist biases present in the news corpora it was trained on.[50] But since the researchers can only view the trained *word2vec*, and not the news corpus it was trained on, it is also hard to pinpoint the reasons of this bias or make it less biased.[51]

- Language models (2018 onwards): Most recent ones rely on deep learning. They also excel in analysing the whole document, but here the 'vectors' are dynamic and adapt to the context. This means that transformers will be better at understanding the difference when same word is used in different context.[52]

[46] Rostyslav Neskorozhenyi, 'Word embeddings in 2020. Review with code examples' (2020) https://towardsdatascience.com/word-embeddings-in-2020-review-with-code-examples-11eb39a1ee6d , Antonio Lopardo, Word2Vec to Transformers' <https://towardsdatascience.com/word2vec-to-transformers-caf5a3daa08a>

[47] Timothy Tan, 'Evolution of Language Models: N-Grams, Word Embeddings, Attention & Transformers' (2020) https://towardsdatascience.com/evolution-of-language-models-n-grams-word-embeddings-attention-transformers-a688151825d2

[48] https://code.google.com/archive/p/word2vec/

[49] Ilias Chalkidis and Dimitrios Kampas 'Deep learning in law: early adaptation and legal word embeddings trained on large corpora' (2019) Artificial Intelligence and Law 27, 171, 174.

[50] Tommaso Buonocore, 'Man is to Doctor as Woman is to Nurse: The Gender Bias of Word Embeddings' (2019) https://towardsdatascience.com/gender-bias-word-embeddings-76d9806a0e17

[51] Amanda Levendowski, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem' (2017) 93 Wash. L. Rev. 579, 582-583.

[52] Lavanya Gupta, 'Differences between Word2Vec and BERT' (2020) https://medium.com/swlh/differences-between-word2vec-and-bert-c08a3326b5d1

These rely on deep neural networks, which are better at detecting and predicting 'complicated linguistic structures along with their long-distance relationships, as humans do'.[53] Another difference of transformers is that they can process words "in parallel", instead of "sequentially one by one" like the former methods, which makes it faster at going through large amounts of data.[54]

Transformer models found online are also trained on unlabelled data, for example Google's BERT trained on Wikipedia and Brown Corpus.[55] They can then be tweaked for the task/corpus at hand in other projects. One of the drawbacks is they do not exist for all languages, so not all researchers will have the same advantage. Additionally, the pre-trained versions might still require some fine-tuning. So, they might not be sufficient on their own, but they can make the other smaller projects easier.

2.3.4. Trained Model

The final stage is the creation of the trained model (a permanent file). Once the researchers have a trained model, they can use it on previously unseen datasets or use it to inform and support other larger tasks. What the trained model achieves depends on what task it was trained for. As mentioned above, some tasks have direct applications, while the others mainly help other NLP tasks.

Algorithms developed for Natural Language Understanding aim to determine the meaning of a sentence. Through syntactic and semantic analysis, AI applications manage to "read" the text. Document classification, sentiment analysis or named entity recognition are some of the examples such "understanding" tasks. Algorithms that "write" or "speak" are for Natural Language Generation.[56] For example, machine translations or chat bots answering questions achieve both understanding and generation through the multiple NLP tasks.

As a final note for the trained model, both for NLP and Computer Vision, it is not possible to remove some of the data after the model is trained. So, if a small part of the data needs to be removed (due to copyright or another reason, following an injunction), then the whole model needs to be retrained from the beginning.

2.4. Computer Vision

The third case study will focus on computer vision. The developments in this field have been largely driven by industry uses, such as facial recognition or self-driving cars.[57] The discussion here will rely on the example of using object recognition technology for content moderation.

In supervised learning, AI is trained with annotated datasets and also gets human feedback when it wrongly classifies something based on the features. In unsupervised learning, AI learns by looking at the different images and recognising the similarities, like the way humans do by observation.[58]

As mentioned earlier, the use of deep neural networks has developed together with the increase in computing power. Although they can be used in language processing (as illustrated earlier), their earliest

[53] Ilias Chalkidis and Dimitrios Kampas 'Deep learning in law: early adaptation and legal word embeddings trained on large corpora' (2019) Artificial Intelligence and Law 27, 171.

[54] Rostyslav Neskorozhenyi, 'Word embeddings in 2020. Review with code examples' https://towardsdatascience.com/word-embeddings-in-2020-review-with-code-examples-11eb39a1ee6d

[55] Sejuti Das, 'Top 8 Pre-Trained NLP Models Developers Must Know' (2020) https://analyticsindiamag.com/top-8-pre-trained-nlp-models-developers-must-know/

[56] Eda Kavlakoglu, 'NLP vs. NLU vs. NLG: the differences between three natural language processing concepts' https://www.ibm.com/blogs/watson/2020/11/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts

[57] Taylor Arnold, Lauren Tilton and Annie Berke, 'Visual Style in Two Network Era Sitcoms' (2019) Journal of Cultural Analytics.

[58] Arthur I. Miller, *The Artist in the Machine: The World of AI-Powered Creativity* (MIT Press 2019) Chapter 10.

application was in the field of computer vision.[59] An example of using deep learning is the use of generative adversarial networks (GAN) in creating art. In this unsupervised form of learning, the generator continuously tests the discriminator with a realistic works, that are not very different from what is currently perceived as art. In addition to requiring large datasets of images of paintings,[60] such practices lead to questions about the copyright status of the AI-created works, that is outside the scope of this paper.

Although computer vision tasks vary widely, the process also starts with the collection of input data, followed by the processing of the data (which are different from NLP pre-processing tasks), followed by the training and leading to the outputs (which could range from a simple yes/no classification decision to a detailed, AI-generated response).

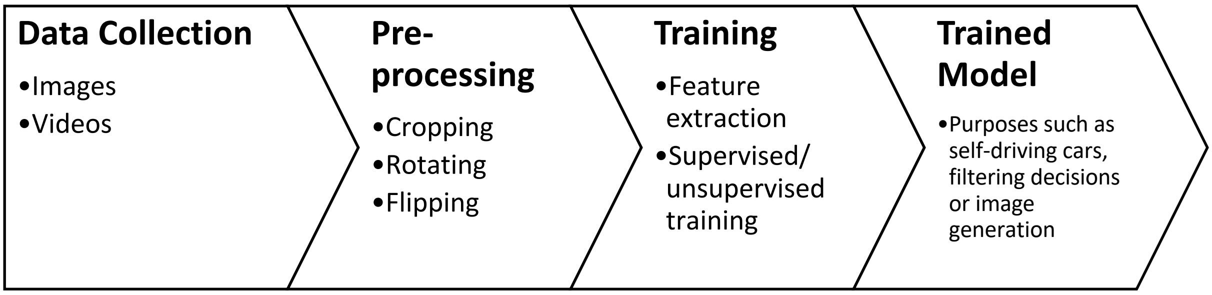
Data Collection
•Images
•Videos
Pre-processing
•Cropping
•Rotating
•Flipping
Training
•Feature extraction
•Supervised/ unsupervised training
Trained Model
•Purposes such as self-driving cars, filtering decisions or image generation

Figure 3: Computer vision workflow

2.4.1. Data Collection

The images or videos can come from various sources, such as phone cameras or medical devices. When training a computer vision model, it is important to use a dataset that is similar to the data it will be used for.[61]

For common objects, there are open datasets of labelled images online.[62] As one of the earlier projects of computer vision, ImageNet was launched in 2007 and holds over 14 million images labelled by participants.[63] But easily accessible datasets are not sufficient for very specific research problems[64] and does not give any competitive edge if everyone trains their AI systems with the same images.[65]

Another option is using own image data or even a digitally generated dataset (synthetic data).[66] If the collected data is too small, it can be augmented (see below).

[59] Seifert et al, Visualizations of Deep Neural Networks in Computer Vision: A Survey' in Tania Cerquitelli, Daniele Quercia and Frank Pasquale. *Transparent Data Mining in Big and Small Data* (Springer 2017) 125.

[60] 75753 paintings were used to train the Generative Adversarial Network in the project where creative adversarial networks were proposed for the first time: Elgammal et al, 'CAN: Creative Adversarial Networks Generating "Art" by Learning About Styles and Deviating from Style Norms' (2017). See also Arthur I. Miller, *The Artist in the Machine: The World of AI-Powered Creativity* (MIT Press 2019) 113-118.

[61] Lila Mullany, 'Introduction to Computer Vision Model Training' (2020) < https://towardsdatascience.com/introduction-to-computer-vision-model-training-c8d22a9af22b>

[62] Meiryum Ali, '20 Free Image Datasets for Computer Vision' https://lionbridge.ai/datasets/20-best-image-datasets-for-computer-vision/

[63] http://www.image-net.org/about

[64] Appen, How to Create Training Data for Computer Vision Use Cases (2019) https://appen.com/blog/how-to-create-training-data-for-computer-vision-use-cases/

[65] https://www.dynam.ai/computer-vision-projects-management-part-1/

[66] Lila Mullany, 'Introduction to Computer Vision Model Training' (2020) < https://towardsdatascience.com/introduction-to-computer-vision-model-training-c8d22a9af22b>

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

2.4.2. Pre-Processing

Once the data is collected, the images or videos go through pre-processing tasks, which are relevant for the legal analysis.

One of the tasks in pre-processing is the resizing of the image, so that all images in the dataset are the same size. Converting colour images to grayscale reduces the computation complexity, for research problems where the colour does not matter.[67]

Another task is noise reduction where the background features are smoothed and removed, so that the machine can focus on a single feature.[68]
One way to increase the dataset and preparing the AI application for recognising the same objects in different environments is data augmentation. This can be achieved by rotating, scaling, cropping or flipping the image.[69] While augmentation follows similar steps as above, it is only applied to the training data sets and not to the test sets.[70]

2.4.3. Training stage

Similar to NLP, Computer Vision also has supervised, semi-supervised and unsupervised training options. Supervised and semi-supervised requires annotated datasets. In unsupervised learning, computer vision is able to recognise common features in images (cluster analysis), so it works without annotations.[71]

Annotation is performed by assigning a label to the selected part of the image, or a single label for the entire image.[72] Feature extraction can be included under this stage – or alternatively be seen as a separate stage in the computer vision process. A feature is defined as "a measurable piece of data in your image that is unique to that specific object…a distinct color or a specific shape such as a line, edge, or image segment".[73] The features can be extracted manually or automatically. The training then occurs based on the extracted features.

Some steps here can be merged due to the technological developments in deep learning. Convolutional neural networks (CNN) are used for image classification and recognition problems.[74] Prior to CNNs, the standard ML training process (for videos) included (i) extracting the features, (ii) combining the features into

[67] Mohamad Elgendy, *Deep Learning for Vision Systems* (Manning Publications 2020).
[68] Sharath Kumar and Manjula Hosurmath, 'Multiclass image classification of yoga postures using Watson Studio and Deep Learning as a Service' (2019) 'https://developer.ibm.com/technologies/artificial-intelligence/tutorials/image-preprocessing-for-computer-vision-usecases/
[69] Appen, How to Create Training Data for Computer Vision Use Cases (2019) https://appen.com/blog/how-to-create-training-data-for-computer-vision-use-cases/; Mohamad Elgendy, *Deep Learning for Vision Systems* (Manning Publications 2020)
[70] Joseph Nelson, 'Why Image Preprocessing and Augmentation Matter' (2020) https://blog.roboflow.com/why-preprocess-augment/
[71] Appen, How to Create Training Data for Computer Vision Use Cases (2019) https://appen.com/blog/how-to-create-training-data-for-computer-vision-use-cases/
[72] Lila Mullany, 'Introduction to Computer Vision Model Training' (2020) https://towardsdatascience.com/introduction-to-computer-vision-model-training-c8d22a9af22b
[73] Mohamad Elgendy, *Deep Learning for Vision Systems* (Manning Publications 2020)
[74] Andrej Karpathy et al, 'Large-scale Video Classification with Convolutional Neural Networks' (2014) IEEE Conference on Computer Vision and Pattern Recognition

a fixed-sized video level description and (iii) a classifier is trained on 'bag-of-words' level descriptions - CNNs combine all these stages.[75]

CNNs have layers of "small computational units that process visual information hierarchically in a feed-forward manner", so each layer works as an image filter and extracts a feature from the image and the image becomes increasingly more explicit along this hierarchy.[76] The process is a slightly different for videos. When used for a video, AI technology has to detect key images which are the most relevant images in the video and eliminate redundant or blurry images. Doing so simplifies the analysis work afterwards.[77] CNNs can be used both supervised and unsupervised, and although widely used for image classification, they can also be used for text classification.[78]

2.4.4. Models for Content Moderation

Trained models can be used in tasks such as image classification (used for example in medical diagnosis or reading traffic signs), object detection and localisation, generating images, face recognition and image recommendation.[79] Some tasks of computer vision are more suitable for unsupervised methods (such as image classification), while others might require more human input.
When using AI for content moderation, it is also possible to combine computer and human moderation: for example, when determining if the user generated content is harmful; the AI application can flag some as "uncertain", which then goes to human moderators whose decisions can be fed back as training data for the AI to learn how to address similar images or videos.[80] Trained on datasets for recognising things like nudity, violence or drugs, there are various companies that are using AI technology for content moderation.[81]

As a final note, in the example of using computer vision for content moderation, AI is only one of the methods. There are also methods called hashing and fingerprinting. Hashing works by generating unique identifiers for files and then comparing it with reference databases for detecting things like terrorist content or viruses.[82] Fingerprinting is similar to hashing, but the unique identifier is not based on the file, but for the characteristics of its content.[83] While it is easier to match content found online to previously flagged content, training AI to make decisions on new content is more difficult. Furthermore, the reasoning for AI decisions is more obscure.[84]

[75] Andrej Karpathy et al, 'Large-scale Video Classification with Convolutional Neural Networks' (2014) IEEE Conference on Computer Vision and Pattern Recognition; Cambridge Consultants, Use of AI in Online Content Moderation 2019 Ofcom Report, 51-52

[76] Leon S Gatys, Alexander S Ecker and Matthias Bethge, 'A Neural Algorithm of Artistic Style' (2015)

[77] 'Mission Report: Towards more effectiveness of copyright law on online content sharing platforms: overview of content recognition tools and possible ways forward' (English version) Joint Report by CSPLA, CNC and HADOPI (January 2020).

[78] Joris Guérin, Olivier Gibaru, Stéphane Thiery, and Eric Nyiri, 'CNN features are also great at unsupervised classification' (2018) 8th International Conference on Computer Science, Engineering and Applications.

[79] Mohamad Elgendy, *Deep Learning for Vision Systems* (Manning Publications 2020).

[80] European Parliament Study, 'The impact of algorithms for online content filtering or moderation' Policy Department for Citizens' Rights and Constitutional Affairs (2020) 23.

[81] Examples include Clarifi, Amazon Rekognition, Valossa and Sightengine. EUIPO Automated Content Recognition – Discussion paper Phase 1 (2020) https://euipo.europa.eu/tunnel-web/secure/webdav/guest/document_library/observatory/documents/reports/2020_Automated_Content_Recognition/2020_Automated_Content_Recognition_Discussion_Paper_Full_EN.pdf.

[82] EUIPO Automated Content Recognition – Discussion paper Phase 1 (2020) 7.

[83] EUIPO Automated Content Recognition – Discussion paper Phase 1 (2020) 15.

[84] 'Mission Report: Towards more effectiveness of copyright law on online content sharing platforms: overview of content recognition tools and possible ways forward' (English version) Joint Report by CSPLA, CNC and HADOPI (January 2020).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

At this stage, in areas such as copyright content moderation, i.e., where the goal is to identify infringing copies, the principle technological tools in place seem to be fingerprinting and matching. However, when it is crucial to determine contextual uses, such as in the case of determining whether a certain use falls within exempted areas like parody or criticisms, properly trained AI tools appear to the the only viable solution. This discussion crosses the bridges between the work conducted in this task of WP3 and the work on intermediaries developed in WP. In the light of this development, a specific – initially not planned – collaboration between the two WPs has taken place and lead to a first online publication titled: "Algorithmic propagation: do property rights in data increase bias in content moderation"[85].

3. Legal implications

The case studies seek to capture legally pertinent stages of three technological processes. They represent an underlying factual set of assumptions for legal analysis.

It is important to note that the legal framework in this specific field is in a phase of transition, with new exceptions entering into force with the adoption and transposition of the CDSM Directive. In particular, Arts. 3 and 4 significantly reshape the *acquis* applicable to text and data mining but also more generally to the broader field of data analytics. These two articles introduce two mandatory exceptions that will exempt respectively acts of reproduction for the purpose of text and data mining made by research organisations and cultural heritage institutions for the purpose of scientific research (Art. 3) or by anyone for any purposes but with the possibility of "contract-out" (Art. 4).

In the paper attached as an Appendix we focus on the underlying legal framework in which exceptions for text and data mining are set, i.e. the right of reproduction contained in Art. 2 ISD. In doing so, proper consideration is given to Art. 5(1) ISD which creates a limited but key exception for certain acts of temporary reproduction which has been tasked by the EUCJ with the crucial role of enabling technological development. Consequently, the legal analysis concentrates on the relationship between the aforementioned legal provisions (Arts. 2, 5(1) ISD; 3 & 4 CDSM) and the technological processes identified in the case studies (e.g. data acquisition, data (pre-)processing, and data analysis).

The analysis covers the key aspects specified for the work package:

- Identification of the relevant *acquis* and the systematic classification of the new TDM exceptions;
- The right of reproduction;
- The exception for temporary reproductions (Art. 5(1)) as interpreted by the CJEU;
- Temporary and permanent reproductions in TDM and data analytics;
- The nature of data analytics as a copyright relevant act
    - International and EU law considerations;
- Detailed analysis of the provisions of Art. 3&4
    - Definitions
    - Rights
    - Contractual and technological overridability
    - Lawful access to original sources
    - Storage of copies for verifiability

[85] EUIPO Automated Content Recognition – Discussion paper Phase 1. See further work conducted by Margoni, Quintais and Schwemer as a WP3 - WP6 spin-off collaboration: 'Algorithmic propagation: do property rights in data increase bias in content moderation?' (part I & II), Kluwer Copyright Blog (8, 9 June 2022) http://copyrightblog.kluweriplaw.com/2022/06/08/algorithmic-propagation-do-property-rights-in-data-increase-bias-in-content-moderation-part-i/.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

For future research and policy direction on the role of the TDM exceptions in data analytics and Artificial Intelligence systems, we recommend moving towards a different paradigm in data regulation, from exceptions to proprietary property rights to a focus on access and use for the purposes of innovation. The proposed EU Data Act[86] already points in this direction.

Our case studies show that a property based approach is problematic where AI applications are based on machine learning, i.e. where an algorithm needs to be trained on data. Property rights create issues of access (authorization to use) and establish conditions (availability at what price). Consequently, access to data for AI (such as in the case studies of data scraping, natural language processing and computer vision) may be limited to those who can pay the price, train outside the EU, or train in the EU and hide their sources – not a desirable mix of incentives for innovation.

The proposed Data Act takes a different approach. In regulating data generated by Internet of Things (IoT) products or related services, the Data Act establishes rules to make data generated by the use of IoT available to the user of that product or service, to make data available by data holders to data recipients, and to make data available by data holders to public sector bodies or where there is an exceptional need, for the performance of a task carried out in the public interest. Reasons include that '[s]uch data are potentially valuable to the user and support innovation and the development of digital and other services protecting the environment, health and the circular economy, in particular though facilitating the maintenance and repair of the products in question' (Rec. 14). Arts. 4 and 5 in particular articulate 'The right of users to access and use data generated by the use of products or related services' and a 'Right to share data with third parties'.

These interventions seem to reflect a very different approach in the governance of data than we have observed and studied in the field of Intellectual Property Law. Art. 35 of the proposed Data Act directly addresses the tension with the proprietary premises of Intellectual Property legislation. It states that in order to avoid that data holders illegitimately shield under the *sui generis* database right (SGDR)[87] to escape their obligations under Arts. 4 and 5 of the Data Act, the SGDR does not apply to IoT data.[88]

Our case studies and analysis of the existing text and data mining provisions suggest that the approach of Data Act is superior. Value generated from data by private investment does not necessarily mean that a regime of private rights should follow.

Annexes

Annex A – Legal analysis of rights and exceptions for input data in machine learning environments:
Thomas Margoni and Martin Kretschmer (2022) 'A deeper look into the EU Text and Data Mining exceptions: Harmonisation, data ownership, and the future of technology' (full paper).

[86] Proposal for a Regulation of the European Parliament and of The Council on harmonised rules on fair access to and use of data (Data Act), Brussels, 23.2.2022 COM(2022) 68 final

[87] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases

[88] Note that there remains an ambiguity in the proposed text of the Data Act whether the *sui generis* database right does not apply for all data generated by IoT devices and services, or only when obligations under Arts. 4 and 5 may be affected.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870626

GRUR International, 71(8), 2022, 685–701

https://doi.org/10.1093/grurint/ikac054
Advance Access Publication Date: 26 July 2022
Article

THOMAS MARGONI[*]/ MARTIN KRETSCHMER[**]

A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology

This paper focuses on the two exceptions for text and data mining (TDM) introduced in the Directive on Copyright in the Digital Single Market (CDSM). While both are mandatory for Member States, Art. 3 is also imperative and finds application in cases of text and data mining for the purpose of scientific research by research and cultural institutions; Art. 4, on the other hand, permits text and data mining by anyone but with rightholders able to ‘contract-out’ (Art. 4). We trace the context of using the lever of copyright law to enable emerging technologies such as AI and the 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; we 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 training data that offer the wrong incentives. To avoid licensing, it may be economically attractive for EU-based developers to train their algorithms on older, less accurate, biased data, or import AI models already trained abroad on unverifiable data.

I. Introduction

The Directive on Copyright in the Digital Single Market (CDSM)[1] contains 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. 3-7), 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 immediately attracted scholarly and media attention and were the 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]).

[*] Research Professor of Intellectual Property Law, Centre for IT & IP Law (CiTiP), Faculty of Law and Criminology, University of Leuven (KU Leuven), Belgium; thomas.margoni@kuleuven.be.

[**] Professor of Intellectual Property Law, University of Glasgow, United Kingdom, and Director of CREATe (UK Copyright & Creative Economy Centre).

[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) [2019] OJ L130/92 (CDSM).

[2] The Directive was advertised as addressing cross-border issues, more inclusive exceptions and fairer markets. The original pitch is preserved on the Internet Archive’s Wayback Machine, available at: European Commission, ‘Modernisation of the EU copyright rules’ (*Archive-It*, 4 March 2021) <https://wayback.archive-it.org/12090/20210304045117/https://ec.europa.eu/digital-single-market/en/modernisation-eu-copyright-rules> accessed 16 March 2022.

[3] For a thorough review of the Directive see Severine Dusollier, ‘The 2019 Directive on Copyright in the Digital Single Market: Some progress, a few bad choices, and an overall failed ambition’ (2020) 57(4) CMLR 979; João Pedro Quintais, ‘The New Copyright in the Digital Single Market Directive: A Critical Look’ (2020) 42 EIPR 28; for a linguistic analysis of the formation of the Directive, see Ula Furgał and others, ‘Memes and Parasites: A discourse analysis of the Copyright in the Digital Single Market Directive’ (2020) CREATe working paper 2020/10 <https://doi.org/10.5281/zenodo.4085050> accessed 18 March 2022.

[4] European Copyright Society, ‘General Opinion on the EU Copyright Reform Package’ (*European Copyright Society*, 24 January 2017) 7 <https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf> accessed 16 March 2022; assessments of the Implementation of the CDSM Directive followed in 2020: The European Copyright Society, ‘Comment of the European Copyright Society Selected Aspects of Implementing Article 17 of the Directive on Copyright in the Digital Single Market into National Law’ (2020) 11(2) Journal of Intellectual Property, Information Technology and Electronic Commerce Law 115.

[5] Academic interventions relation to the press publishers’ right (then art 11) were catalogued by CREATe in 2017, see CREATe, ‘Article 11

At least during the drafting phase, the provisions contained in Arts. 3 and 4 of the Directive which are dedicated to 'text and data mining' (TDM) have attracted far less attention,[6] although Art. 4 was introduced quite late in the legislative process following a proposal by the Dutch delegation.[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 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. Article 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; however it can be expressly reserved by rightholders – in other words it may be overridden by 'opt-out' or 'contract-out'.

This article 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 overly 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 article further argues that the implications of accepting the principle that AI in the EU 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 the role of copyright is often seen as secondary. However, the recognition of property rights in data, i.e. in the essential building blocks necessary for 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 when contained in protected subject matter 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 years to come.[11] These rules, today, in the EU, 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 rightholders to develop AI. Article 4 only partially recalibrates this situation by changing the default rule from an opt-in to an opt-out: important, but not sufficient. Outside the EU, in an increasing number of jurisdictions there is simply no need to obtain equivalent

Research' <https://www.create.ac.uk/policy-responses/eu-copyright-reform/article-11-research/> accessed 16 March 2022; Ula Furgał, 'The EU Press Publishers' Right: Where Do Member States Stand?' (2021) 16 Journal of Intellectual Property Law & Practice 887.

[6] Although see Christophe Geiger and others, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects: In-Depth Analysis' (Policy Department for Citizens' Rights and Constitutional Affairs, Directorate General for Internal Policies of the Union, February 2018) <https://www.europarl.europa.eu/RegData/etudes/IDAN/2018/604941/IPOL_IDA(2018)604941_EN.pdf> accessed 16 March 2022; Christophe Geiger and others, 'Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?' (2018) 49 IIC 814; Rossana Ducato and Alain Strowel, 'Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility"' (2019) 50 IIC 649; Begoña Gonzalez Otero, 'Machine Learning Models Under the Copyright Microscope: Is EU Copyright Fit for Purpose?' [2021] GRUR International 1043; Eleonora Rosati, 'An EU Text and Data Mining Exception for the Few: Would It Make Sense?' (2018) 13 JIPLP 429; Andres Guadamuz and Diane Cabell, 'Data Mining in UK Higher Education Institutions: Law and Policy' (2014) 4 Queen Mary Journal of Intellectual Property 3.

[7] P Bernt Hugenholtz, 'The New Copyright Directive: Text and Data Mining (Articles 3 and 4)' (*Kluwer Copyright Blog*, 24 July 2019) <http://copyrightblog.kluweriplaw.com/2019/07/24/the-new-copyright-directive-text-and-data-mining-articles-3-and-4/> accessed 16 March 2022; Christophe Geiger and others, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/Eu' (2019) Centre for International Intellectual Property Studies Research Paper No 2019-08 <https://www.ssrn.com/abstract=3470653> accessed 16 March 2022.

[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 the InfoSoc (Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L167/10 (InfoSoc Directive)) and Database (Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L77/20 (Database Directive)) Directives. The reference in art 3 to the InfoSoc Directive 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 (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) (Text with EEA relevance) [2009] OJ L111/16 (Software Directive)) is considered *lex specialis* with regards to the general InfoSoc Directive (eg Case C-128/11 *UsedSoft GmbH v Oracle International Corp.* ECLI:EU:C:2012:407, paras 51, 56), therefore the reference of art 3 CDSM to the general right of reproduction *ex* art 2 InfoSoc Directive 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 a 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] Thomas Margoni, 'Computational Legal Methods: Text and Data Mining in Intellectual Property Research' in Irene Calboli and Maria Lillà Montagnani (eds), *Handbook of Intellectual Property Research* (Oxford University Press 2021); Josef Drexl and others, 'Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective' (2019) Max Planck Institute for Innovation & Competition Research Paper No 19-13 <https://papers.ssrn.com/abstract=3465577> accessed 16 March 2022.

[10] Although some early warnings were raised; Geiger and others, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects: In-Depth Analysis' (n 6); Martin Kretschmer and Thomas Margoni, 'Data Mining: Why the EU's Proposed Copyright Measures Get It Wrong' (*The Conversation*, 24 May 2018) <http://theconversation.com/data-mining-why-the-eus-proposed-copyright-measures-get-it-wrong-96743> accessed 16 March 2022.

[11] Pamela Samuelson, 'Regulating Technology Through Copyright Law: A Comparative Perspective' (2020) 42 EIPR 214; Yochai Benkler, 'The Role of Technology in Political Economy: Part 1' (*LPE Project*, 25 July 2018) <https://lpeproject.org/blog/the-role-of-technology-in-political-economy-part-1/> accessed 16 March 2022.

authorisations thanks to statutory or judicial developments supportive of technological innovation. What this means for cultural and innovation policies, regulatory competition and the future of democracy is a complex question that 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. It appears at least arguable that, being 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] Another plausible scenario is where developers simply purchase pre-trained models, i.e. models already trained, usually in jurisdictions where this is permitted by law. This latter dynamic, however, may further propagate biased, opaque and unaccountable AI given the fact that there will be little or no transparency of the underlying data used for training.

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 theory, but also 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 from a paper poster onto a canvas[17]); a right protecting non-original databases (Art. 7 Database Directive); and a closed non-mandatory list of exceptions that must be interpreted narrowly (Art. 5 InfoSoc Directive) which, at the same time, represents all the limits to which EU copyright law's exclusive rights can be subjected, including those connected to fundamental rights (the *Funke Medien, Pelham* and *Spiegel Online* cases[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 the 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 and 4 CDSM: a non-formalistic 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. This definition reaches far beyond the taxonomy employed and captures most activities where digital technologies are utilised to analyse information and extract meaning. Nowadays, one of the most widespread approaches to perform this task is a sub-class of AI termed machine learning (ML).[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 AI/ML now known or developed in the future as long as they rely on data analytics.

However, when such a broad definition is inserted into a narrowly construed exception, such as the one under analysis, the result may not be that of opening up new technological and cultural practices, as was arguably the original intention of the drafters, but rather

[12] Martin Senftleben and others, '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 and the Opportunity Arising from Article 17 of the CDSM Directive' (2022) 13 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 67.

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

[14] The paper does not discuss the related but systematically distinct issue of property rights in generated data; for an insightful analysis see P Bernt Hugenholtz, 'Against "Data Property"' in Hanns Ullrich and others (eds), *Kritika: Essays on Intellectual Property* (Edward Elgar 2018); Wolfgang Kerber, 'A New (Intellectual) Property Right for Non-Personal Data? An Economic Analysis' [2016] GRUR Int 989; Francesco Benterle, 'Data Ownership in the Data Economy: A European Dilemma' in Tatiana-Eleni Synodinou and others (eds), *EU Internet Law in the Digital Era* (Springer 2020). The paper is also not directly concerned with whether AI outputs may or should be protected by copyright, see Reto M Hilty and others, 'Intellectual Property Justification for Artificial Intelligence' in Jyh-An Lee and others (eds), *Artificial Intelligence and Intellectual Property* (OUP 2021); Joint Institute for Innovation Policy and IViR – University of Amsterdam, *Trends and Developments in Artificial Intelligence: Challenges to the Intellectual Property Rights Framework: Final Report* (Publications Office of the European Union 2020) <https://data.europa.eu/doi/10.2759/683128> accessed 16 March 2022; Mauritz Kop, 'AI & Intellectual Property: Towards an Articulated Public Domain' (2020) 28 Texas Intellectual Property Law Journal 297. Finally, the paper does not cover the issue of personal data, see eg Paolo Guarda, 'Free data? open science in the age of personal data protection' in Jacob H Rooksby (ed), *Research Handbook on Intellectual Property and Technology Transfer* (Edward Elgar 2020).

[15] Case C-5/08 *Infopaq I* ECLI:EU:C:2009:465 and Case C-302/10 *Infopaq II* ECLI:EU:C:2012:16.

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

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

[18] Case C-469/17 *Funke Medien* ECLI:EU:C:2019:623; Case C-476/17*Pelham v Hütter* ECLI:EU:C:2019:624; Case C-516/17 *Spiegel Online* ECLI:EU:C:2019:625.

[19] For common usage, see 'Machine Learning' (*Wikipedia* 2021) <https://en.wikipedia.org/wiki/Machine_learning> accessed 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'.

the opposite. Not only TDM *stricto sensu* has been limited to research uses by research and cultural organisations or private ordering, 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. These interventions, which have clear *prima facie* potential for anti-competitive effects, have not been addressed from an 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 regard.[22]

This short-sighted approach to law-making may further contribute to the already existing problem of 'opaque' AI systems or AI 'black boxes',[23] an expression referring to a type of automated decision-making tool (e.g. AI) 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, a 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.[25]

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.[26] 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.[27] 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, or credit or health-related requests accepted or refused), but it is not possible for human observers to *understand* why.[28]

This reflects the characteristic of any ML-based AI system to be a black box, but there are ways to mitigate this situation. One route to get closer to 'understand' how the learning algorithm has reached certain decisions is access to the original training data. This would not necessarily explain the complex statistical process leading to those decisions but would make it possible to scrutinize 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.[29]

The appropriateness 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 this emerging field.

In conclusion, it may be argued that under the misleading label of 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 a potentially far-reaching legislative development which may have a

[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, 'Commission Staff Working Document on the Modernisation of EU Copyright Rules Brussels' (SWD(2016) 301 final PART 1/3, s 4.3.

[21] ibid 114.

[22] European Copyright Society (n 4) 5; Kretschmer and Margoni (n 10); Geiger and others, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects: In-Depth Analysis' (n 6).

[23] Levendowski (n 13).

[24] Jonathan Zittrain, 'Intellectual Debt: With Great Power Comes Great Ignorance' (*Berkman Klein Center Collection*, 24 July 2019) <https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c> accessed 16 March 2022.

[25] Margoni (n 9).

[26] Jared Kaplan and others, 'Scaling Laws for Neural Language Models' [2020] arXiv:2001.08361 [cs, stat] <http://arxiv.org/abs/2001.08361> accessed 16 March 2022.

[27] Thomas Margoni, 'Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?' (2018) 27(1) AIDA 281-304; for a case scenario analysis see the documentation collected at <https://www.create.ac.uk/legal-approaches-to-data-scraping-mining-and-learning> accessed 10 May 2022.

[28] Zittrain (n 24).

[29] Levendowski (n 13).

profound impact on the relationship between law and technology, and the future of the EU legal order.

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[30] or approximately one new paper every 30 seconds.[31] 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 read by anyone other than their authors, referees and journal editors.[32] 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 rightholders, something not completely clear under previous law.[33]

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 NLP, the ability to develop automated translation tools is currently limited mostly to the official documents produced by the European Union,[34] which are openly available and reusable.[35] 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 very large platforms, such as Google, Facebook, Amazon, Microsoft and Twitter which can benefit from copyright laws that permit *them* to engage in this type of activity without prior authorisation, therefore significantly reducing the cost of certain 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 contribute to stop fake news.[36]

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). In all these cases proper authorisation is needed to avoid infringement.

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' in Art. 2 of the InfoSoc Directive.[37] 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 the remit of copyright.[38] 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, whereas there is no explicit general statutory recognition of the idea/expression doctrine, it can nonetheless be found for instance in the Software Directive (Recital 11 and Arts. 1 and 5.3) with a wording that reveals 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

[30] Mark Ware and Michael Mabe, 'The STM Report: An Overview of Scientific and Scholarly Journal Publishing' (2009) International Association of Scientific, Technical and Medical Publishers 7 <https://www.stm-assoc.org/2009_10_13_MWC_STM_Report.pdf> accessed 16 March 2022. See generally OpenMinTeD ('Home') (*OpenMinTeD*) <http://openminted.eu/> accessed 16 March 2022, for examples of how TDM techniques can be used.

[31] Scott Spangler and others, 'Automated Hypothesis Generation Based on Mining Scientific Literature' (Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM 2014) <https://dl.acm.org/doi/10.1145/2623330.2623667> accessed 16 March 2022.

[32] Lokman I Meho, 'The Rise and Rise of Citation Analysis' (2007) 20 Physics World 32.

[33] Jean-Paul Triaille and others, 'Study on the Legal Framework of Text and Data Mining (TDM)' (Publications Office of the European Union 2014) 41 <https://data.europa.eu/doi/10.2780/1475> accessed 16 March 2022.

[34] OpenMinTeD Communications, 'TDM Stories: How Zalando Links Languages With TDM' (*OpenMinTeD*, 5 February 2018) <http://openminted.eu/tdm-stories-zalando-links-languages-tdm/> accessed 16 March 2022.

[35] art 4 Commission Decision of 12 December 2011 on the reuse of Commission documents (2011/833/EU) OJ L330/39.

[36] OpenMinTeD Communications, 'TDM Stories: A Text & Data Miner Talks About Analysing The Recent Past' (*OpenMinTeD*, 2 February 2018) <http://openminted.eu/tdm-stories-text-data-miner-talks-analysing-recent-past/> accessed 16 March 2022.

[37] InfoSoc Directive (n 8).

[38] eg Sean Flynn and others, 'Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action' (2020) 42 EIPR 393; Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning' (2019) 66 Journal of the Copyright Society of the USA 291; Carys J Craig, 'Globalizing User Rights-Talk: On Copyright Limits and Rhetorical Risks' (2017) 33 American University International Law Review 1.

European Union (CJEU) has endorsed these doctrines, both by direct confirmation of their operativity,[39] 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 Agreement and the WCT, both containing an explicit recognition of the doctrines.[40] Therefore, there should be no doubt about the general validity of an idea/fact/expression doctrines under EU copyright law.

Nevertheless, the effect of the dispositions contained in Arts. 3 and 4 CDSM is to formalise an interpretation that significantly reduces the ambit of application of the idea/fact/expression doctrines. 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 broadly defined rights corrected by specific carve-outs, is emblematic of a more general trend which reached its peak with the InfoSoc Directive of 2001. As is well known in EU copyright scholarship, this trend is characterized by the full harmonisation of copyright's exclusive rights through broad and all-encompassing definitions (Arts. 2-4 InfoSoc Directive), and by the systematic and semantic classification of any area not covered by copyright's exclusivity as an exhaustively listed 'exception' (Art. 5 InfoSoc Directive) – a concept that in the general theory of law derogates from a rule and therefore is subject to conditions such as that of strict interpretation.[41]

Consequently, the introduction of an exception establishing that in very specific cases TDM can be freely performed, leads to the exact opposite effect: all uses that cannot be subsumed within the narrow construction of Arts. 3 and 4 are reserved. Had the legislative technique been 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 an illustration, one could look at the path taken in Art. 14 CDSM. That article plainly clarifies that the digitization 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, and with no ambition of being exhaustive, the following main approaches can be identified: open and flexible standards,[42] the judicial construction of users' rights,[43] or a dedicated TDM limitation for *any purpose*.[44]

In relation to the effects of the broad definition of the right of reproduction in Art. 2 InfoSoc Directive, it is insightful to note that already during the phase that led to its adoption in 2001 this approach was met with criticism. As P. Bernt 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[d] 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

[39] eg *Brompton Bicycle* (n 16) para 27; Case C-683/17 *Cofemel* ECLI:EU:C:2019:721, para 29; Case C-393/09 *BSA* ECLI:EU:C:2010:816, para 49.

[40] See eg, Case C-306/05 *SGAE* ECLI:EU:C:2006:764, para 35.

[41] As an example, 'quotations' are classified as 'free uses' under art 10 Berne Convention for the Protection of Literary and Artistic Works (Paris Act of 24 July 1971), as amended on 28 September 1979 (Berne Convention, BC), but as 'exceptions and limitations' under art 5(3)(d) InfoSoc Directive.

[42] This is the US approach, but it has been adopted by other countries among them Singapore, South Korea, Malaysia, Israel, Taiwan. See Niva Elkin-Koren and Neil Weinstock Netanel, 'Transplanting Fair Use across the Globe: A Case Study Testing the Credibility of U.S. Opposition' (2020) 72 Hastings Law Journal 1121.

[43] 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 eg, Michael Geist, '5. Fairness Found: How Canada Quietly Shifted from Fair Dealing to Fair Use', *The Copyright Pentalogy: How the Supreme Court of Canada Shook the Foundations of Canadian Copyright Law* (Les Presses de l'Université d'Ottawa | University of Ottawa Press 2017) <http://books.openedition.org/uop/969> accessed 17 March 2022. A perhaps similar development could be seen – albeit still in an embryonic from – in some CJEU decisions, see Martin Senftleben, 'Bermuda Triangle – Licensing, Filtering and Privileging User-Generated Content Under the New Directive on Copyright in the Digital Single Market' (2019) 41 EIPR 480, 481; Dusollier (n 3); Caterina Sganga, 'A New Era for EU Copyright Exceptions and Limitations?' (2020) 21 ERA Forum 311.

[44] This is the course taken more than ten years ago by Japan, see Tatsuhiro Ueno, 'The Flexible Copyright Exception for "Non-Enjoyment" Purposes – Recent Amendment in Japan and Its Implication' [2021] GRUR International 145.

interpretation of the reproduction right, as advanced by the Commission, would mean carrying the copyright monopoly one step too far. [...].'[45]

The advice of the Legal Advisory Board seems to have been largely ignored in the adopted text. However, its message should not be completely lost. The rational way to rebalance the amplitude currently enjoyed by the right of reproduction would be to redefine it, i.e. a modification of Art. 2 InfoSoc Directive. However, this seems a highly unlikely course of action at present time.[46] 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 a workable compromise. This would need to be broader than the current CDSM's Arts. 3 and 4 and broader than what was known as 'option four', a TDM exception not limited to research organisations for research purposes.[47] However, also this door appears to have been firmly shut after the contentious approval of the CDSM.[48] 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: the exception for temporary acts of reproduction in Art. 5(1) InfoSoc Directive.[49] While not specific to computational uses, Art. 5(1) was implemented with the goal of enabling certain technological uses (mainly internet browsing[50]) 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 the Database and InfoSoc Directives, for uses or fields covered by the exceptions or limitations provided for in this Directive'.[51] For present purposes this means that Member States may maintain or introduce a new TDM exception usually on the basis of Art. 5(3)(a) InfoSoc Directive (i.e. illustration for non-commercial teaching and scientific research). Beside the non-commercial *versus* research-purposes-by-research-institutions discussion, as the Art. 5(3)(a) exception is not mandatory, it therefore does not represent an EU-wide solution to the problem addressed in this article. It should be pointed out, however, that this exception, like all the exceptions listed under Art. 5(3) InfoSoc Directive, covers both reproductions and communications to the public thereby offering an opportunity to Member States interested in implementing a wider exception.[52]

3. Article 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.[53]

Article 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'.[54]

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 that act. This condition is satisfied notwithstanding the fact that initiating and terminating that process involves human intervention.[55]

[45] P Bernt Hugenholtz, 'Copyright and Freedom of Expression in Europe' in Rochelle Cooper Dreyfuss and others (eds), *Innovation Policy in an Information Age* (OUP 2000) 9.

[46] Proposing a different interpretation of the relationship 'right-infringement' for art 2 InfoSoc Directive which relies *inter alia* on the CJEU 'recognizability' test expressed in the *Pelham* case in relation to art 2(c), see Rossana Ducato and Alain Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (2021) 43 EIPR 322.

[47] In the Impact Assessment, the EC identified as 'Option four' a TDM exception not limited to research organisations for research purposes (Commission (n 20)).

[48] Martin Senftleben, 'The Perfect Match: Civil Law Judges and Open-Ended Fair Use Provisions' (2017) 33 American University International Law Review 231; P Bernt Hugenholtz, 'Flexible Copyright: Can EU Author's Rights Accommodate Fair Use?' in Irini A Stamatoudi, *New Developments in EU and International Copyright Law* (Kluwer Law International BV 2016).

[49] Recital 9 CDSM. See also Triaille and others (n 33); Margoni (n 27).

[50] Recital 33 InfoSoc Directive.

[51] art 25 CDSM.

[52] Some Member States took full advantage of this opportunity (eg France, Estonia, Germany), whereas others did not (eg UK); see Geiger and others, 'The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects: In-Depth Analysis' (n 6); European Commission, Directorate-General for Communications Networks, 'Content and Technology, Study on copyright and new technologies: copyright data management and artificial intelligence' (2022) <https://data.europa.eu/doi/10.2759/570559> accessed 10 May 2022.

[53] *Infopaq I* (n 15) and *Infopaq II* (n 15). For a detailed analysis please refer to Margoni (n 27).

[54] See *Infopaq I* (n 15) paras 61-64; Theodoros Chiou, 'Copyright Lessons on Machine Learning: What Impact on Algorithmic Art?' (2020) 10 Journal of Intellectual Property, Information Technology and E-Commerce Law 398, 411.

[55] *Infopaq II* (n 15).

(4) Temporary acts of reproduction must pursue a sole purpose, namely, to enable […[56]] 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.[57]
(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.[58]

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* cases may be helpful to properly situate these conditions within a data capture process which shares many logical steps with more modern TDM approaches. 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: 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 Directive, 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 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 Art. 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 perspective 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).

a) Eroding lawful uses

Additionally, it should be briefly contemplated whether the new Arts. 3 and 4 CDSM may in fact have contributed to narrow even further the scope of Art. 5(1) InfoSoc Directive. This may be 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 I* and *II* the Court states that '[…] the parties in the main proceedings do not dispute that in itself [genuinely independent] summary writing is lawful', that 'such an activity is not restricted by European Union legislation' and finally that 'it is apparent from the statements … that the drafting of that summary is not an activity which is restricted'. These statements need closer scrutiny as the availability of Art. 5(1) entirely rests on the lawful nature of this final use, in this case, summary preparation.

Remarkably, in the case under scrutiny, the Court appears satisfied with the fact that parties in the main proceedings do not dispute the issue of summary preparation which allows the Court to avoid, on a procedural ground, a potentially challenging legal question. Certainly, it may be argued that the 'genuinely independent' parameter,[63] whatever its concrete meaning may be, should be safe enough a standard to draw a clear line between independent and derivative works or adaptations. However, it would be interesting (albeit beyond the scope of this

[56] '… either the transmission of a protected work or a protected subject-matter in a network between third parties by an intermediary or …'

[57] *Infopaq II* (n 15).

[58] ibid.

[59] *Infopaq II* (n 15).

[60] *Infopaq I* (n 15) paras 56 and 57; Joined Cases C-403/08 and C-429/08 *Football Association Premier League and Others (FAPL)* ECLI:EU:C:2011:631, para 162; Case C-360/13 *NLA* ECLI:EU:C:2014:1195, para 23.

[61] *NLA* (n 60) para 24.

[62] Recital 33 InfoSoc Directive; *Infopaq II* (n 15), para 68; *FAPL* (n 60) para 168.

[63] *Infopaq I* (n 15) para 23.

paper) to verify whether it is domestic law which does not provide for a right of adaptation apt to cover the creation of summaries which reproduce in part (e.g. 11 consecutive words) the original work or whether other factual or legal considerations played a role in reaching this conclusion. Plausibly, this aspect of the decision was at least in part intentionally evaded by the Court in the light of the fact that, as the same AG notes, the facts of the case as referred by the national court and in particular the relationship between the eleven word extracts and the summary preparation is not clear.[64] Regardless of the reasons that allowed 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 a right not reserved to rightholders under applicable law. A statement that however finds minimal examination in the decision and which leaves open the possibility for domestic legal orders to deviate from this rule, especially when the summaries are not genuinely independent, such as when they reproduce the author's own intellectual creation (and are not excused by other exceptions and limitations).

This concise analysis intends to stress the narrow grounds 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 broad right of reproduction reserves virtually any type of use save for when an exception applies, then the situations where Art. 5(1) finds application are logically limited to those cases where another exception already applies or when the use of a work does not trigger the right of reproduction and/or adaptation (such as the preparation of summaries in the above example).

It follows, that if Art. 5(1) is only available when a certain use is not restricted by applicable legislation as described, the recognition that TDM is a restricted use of rightholders (excused by Arts. 3 and 4 when performed by research and cultural organisations for research purposes or when it is not contracted out) means that temporary acts of reproductions performed for TDM purposes outside the scope of Arts. 3 and 4 are not permitted any longer as they do not meet the condition of lawful use (as TDM is now an act restricted by law). 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 which can arguably be correlated to a lack of adherence to copyright's theoretical framework in the drafting process and in the reported 'exceptionalist' approach in EU copyright law.

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 that sees in Art. 5(1) only a narrow exception needs to be abandoned in favour of a teleological approach to EU copyright law able to strike a fair balance between the public and private rights. The CJEU has shown acquaintance with both approaches, the unambiguous adoption of the latter over the former will likely prove decisive for the future of EU technological development from a property rights point of view.

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

Retaining permanent copies represents a crucial tool to mitigate the black box effect of AI (discussed at the beginning of section II). Greater transparency enables 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 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,[65] 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 representations of the original data. This is especially the case with more sophisticated approaches to ML, 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 these permanent copies (i.e. the trained model 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 purposes. Not recognising this possibility may lead to the situation where, if the trained model contains a reproduction in part or is an adaptation of the training (protected) data, it cannot be distributed or communicated to the public, thereby rendering the whole TDM process, and the related exception, useless. As we will see, Arts. 3 and 4 CDSM have failed to fully address

[64] Case C-5/08 *Infopaq I* ECLI:EU:C:2009:89, Opinion of AG Trstenjak, para 109.

[65] Eckart de Castilho and others, 'A Legal Perspective on Training Models for Natural Language Processing' (2018) Proceedings of the Eleventh International Conference on Language Resources and Evaluation <https://aclanthology.org/L18-1202.pdf> accessed 10 May 2022.

this first type of permanent copies as they only excuse 'reproductions'.

A second type of permanent copy is one necessary for verification purposes. For something to be called 'scientific', it must 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 and it is interesting that in the last decades the field has suffered from a so-called 'reproducibility crisis'.[66] 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.[67] 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.[68] 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 Member States' 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 lives of millions of people should be open, accountable and verifiable. There seems to be little space, if any, in European 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 may agree that market dynamics should regulate these decisions, either with or without public interventions to correct certain distortions in strategic sectors, but public authorities should follow a different, higher standard.[69]

As seen above, in order to be able to 'understand' which determinations are being made by AI systems, perhaps even more important than the algorithm itself, is the data used to train those algorithms. To fulfil this scope, such data must be available to public scrutiny. 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.[70] While this type of permanent copy is not covered by an exception for temporary uses, some limited but important recognition of this aspect is present in Arts. 3 and 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 operate confidently.[71]

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

The main criticisms against the current formulation of Arts. 3 and 4 of the CDSM Directive 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.

1. Definitions

As discussed in the first part of this article, a broad TDM definition inserted in a narrow exception has the effect of subjecting a wide array of text and data analytics activities, including entire fields such as AI and ML, to the strict requirements of Arts. 3 and 4. We refer to the considerations developed in the first part of this article for a full analysis.

2. Beneficiaries

Article 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 can a university acting for any other purpose than scientific research. Other purposes commonly accepted as fundamental in democratic societies also appear to be excluded, such as journalism, criticisms or review.[72]

[66] Monya Baker, '1,500 Scientists Lift the Lid on Reproducibility' (2016) 533 Nature 452.

[67] John PA Ioannidis, 'Why Most Published Research Findings Are False' (2005) 2(8) PLOS Medicine e1 24.

[68] This is an explicit priority of the European Commission, see European Commission, 'Open Science' <https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science_en> accessed 17 March 2022.

[69] In the EU see the proposal for an AI Regulation: Commission, '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.

[70] Zittrain (n 24).

[71] Triaille and others (n 33).

[72] Dusollier (n 3).

In the opinion of the drafters of the Directive, the current wording is thought to be less restrictive than the 'non-commercial' limitation.[73] It seems, however, that the 'double limitation' of Art. 3 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 limitation to the efficacy of the exception that excludes important economic sectors and small and medium-sized enterprises (SMEs) from benefiting from a critically important tool to compete on global markets. This limitation appears in tension 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.[74]

Article 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 rightholders, 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, as early data seems to show at least linguistic divergences in its transposition (see Fig. 1).[75]

3. 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 where the results of an act of TDM include a protected part of the original 'mined' work (and as seen above excerpts, a passage 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. 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. With outside textual sources, e.g. in the case of audiovisual works or software, it may be even more difficult to establish when this threshold has been reached. Whereas it seems that the direction of technological development is towards forms of analysis that reach higher levels of abstraction in the mined texts or data (e.g. neural networks), thus reducing the relevance of this aspect, current statistical ML will likely remain available for a number of years and with it the uncertainty connected with the presence of protected parts in the trained models.

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 at the domestic level.

4. Contractual and technological overridability

Article 3 states that contractual provisions intended to limit the TDM exception shall be unenforceable. This is an important rule, as often 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 is no equivalent rule safeguarding the enjoyment of the exception. The approach taken by the CDSM is convoluted at best. Article 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 other words 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 subpara. 4 of Art. 6(4) does not find application in this case. Subparagraph 4 establishes that the reported mechanism (the obligation on Member States 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 subpara. 4, it is the entire mechanism of Art. 6(4) InfoSoc Directive 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 complaint

[73] Commission (n 20) 108-109.

[74] Commission, 'Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market (Text with EEA relevance)' COM(2016) 593 final, 9; See Geiger and others, 'Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?' (n 6).

[75] For instance, the omission of the 'express' element the 'express reservation' mechanism. CREATe, 'CDSM Implementation Resource Page' <https://www.create.ac.uk/cdsm-implementation-resource-page/> accessed 18 March 2022.

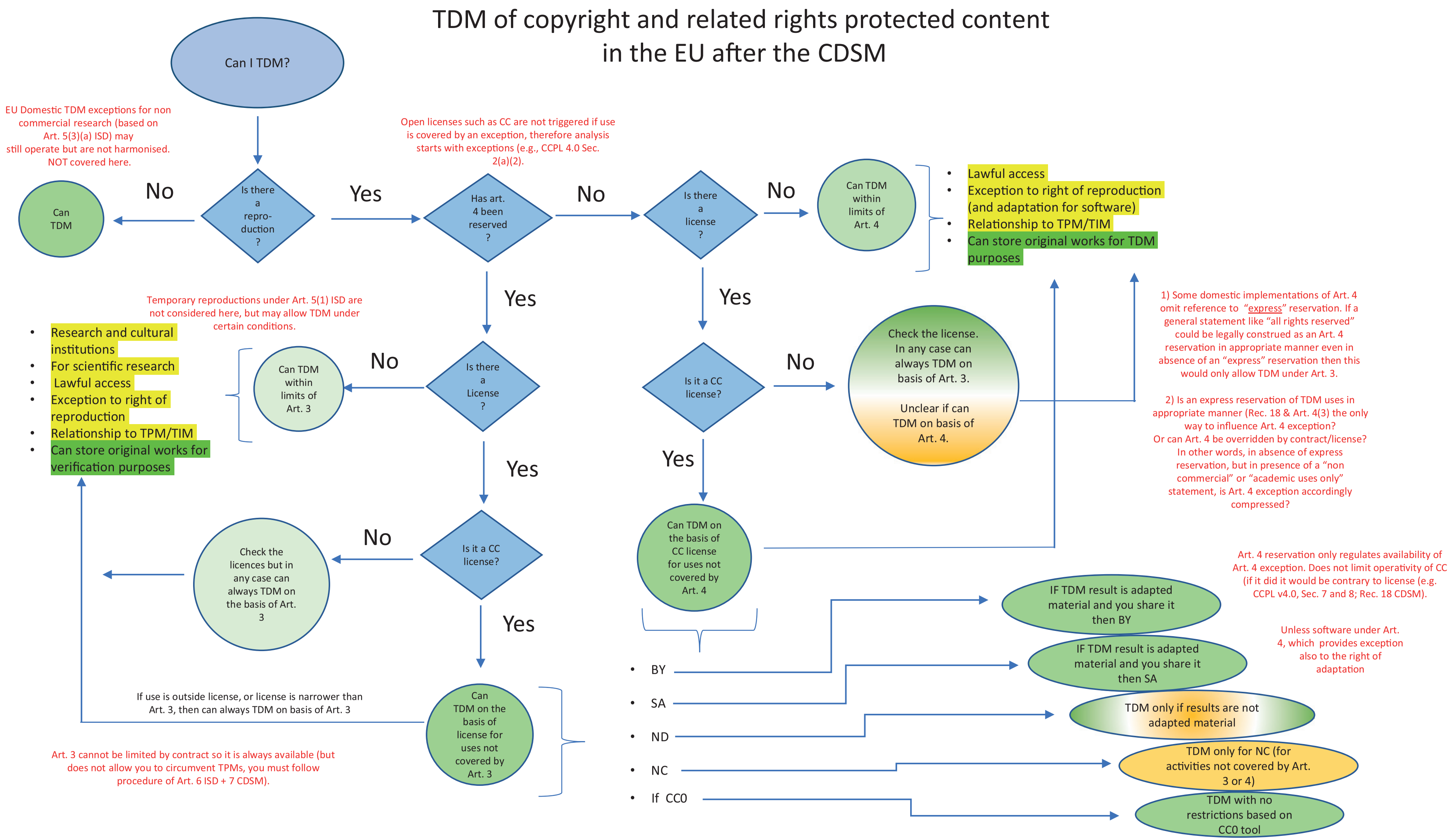
TDM of copyright and related rights protected content in the EU after the CDSM
Can I TDM?
EU Domestic TDM exceptions for non commercial research (based on Art. 5(3)(a) ISD) may still operate but are not harmonised. NOT covered here.
Is there a repro-duction ?
No
Can TDM
Yes
Has art. 4 been reserved ?
Open licenses such as CC are not triggered if use is covered by an exception, therefore analysis starts with exceptions (e.g., CCPL 4.0 Sec. 2(a)(2).
No
Is there a license ?
No
Can TDM within limits of Art. 4
Lawful access
Exception to right of reproduction (and adaptation for software)
Relationship to TPM/TIM
Can store original works for TDM purposes
Yes
Yes
Temporary reproductions under Art. 5(1) ISD are not considered here, but may allow TDM under certain conditions.
Is there a License ?
No
Can TDM within limits of Art. 3
Research and cultural institutions
For scientific research
Lawful access
Exception to right of reproduction
Relationship to TPM/TIM
Can store original works for verification purposes
Is it a CC license?
No
Check the license. In any case can always TDM on basis of Art. 3.
Unclear if can TDM on basis of Art. 4.
1) Some domestic implementations of Art. 4 omit reference to "express" reservation. If a general statement like "all rights reserved" could be legally construed as an Art. 4 reservation in appropriate manner even in absence of an "express" reservation then this would only allow TDM under Art. 3.
2) Is an express reservation of TDM uses in appropriate manner (Rec. 18 & Art. 4(3) the only way to influence Art. 4 exception? Or can Art. 4 be overridden by contract/license? In other words, in absence of express reservation, but in presence of a "non commercial" or "academic uses only" statement, is Art. 4 exception accordingly compressed?
Yes
Yes
Is it a CC license?
No
Check the licences but in any case can always TDM on the basis of Art. 3
Yes
Can TDM on the basis of CC license for uses not covered by Art. 4
Art. 4 reservation only regulates availability of Art. 4 exception. Does not limit operativity of CC (if it did it would be contrary to license (e.g. CCPL v4.0, Sec. 7 and 8; Rec. 18 CDSM).
BY
SA
ND
NC
If CC0
IF TDM result is adapted material and you share it then BY
IF TDM result is adapted material and you share it then SA
TDM only if results are not adapted material
Unless software under Art. 4, which provides exception also to the right of adaptation
TDM only for NC (for activities not covered by Art. 3 or 4)
TDM with no restrictions based on CC0 tool
Can TDM on the basis of license for uses not covered by Art. 3
If use is outside license, or license is narrower than Art. 3, then can always TDM on basis of Art. 3
Art. 3 cannot be limited by contract so it is always available (but does not allow you to circumvent TPMs, you must follow procedure of Art. 6 ISD + 7 CDSM).

Fig. 1. The law of TDM in the EU under Arts. 3 and 4 CDSM.

procedure,[76] 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 the subpara. 4 mechanism, and 1 led to a voluntary solution.[77]

5. Lawful access to original sources

Article 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 about this requirement. Some more details about the role of the 'lawful access' can be found 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'.[78]

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.[79] The arguments to support such a position are numerous. As Michael 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'.[80]

Whereas such argument is developed within the US copyright framework which operates quite differently in relation to some of the elements of EU copyright law scrutinised here, it seems that the same rationale could also find application under EU law. Furthermore, it has been pointed out how the lawful access limitation could subject TDM research to private ordering[81] 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.[82]

6. Storage of copies for verifiability

Article 3(2) 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'.[83] 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 communication or making those copies available to the public, whereas, as mentioned, Art. 3 (and Art. 4) are 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.

IV. EU Copyright law and data enclosures

Having discussed the functional constraints imposed by Art. 5(1) InfoSoc Directive and the practical scope of the TDM interventions under Arts. 3 and 4 CDSM Directive, we now turn to the broader question of the status of data and factual information under EU copyright law.

The position embraced in the CDSM regarding 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 – a category with a strong attractive force under EU law – they become an object of exclusivity. The reason is to be found in the well-known ubiquity of copies in the digital environment. Being that this reason is global, EU copyright law has developed its own idiosyncratic

[76] See Intellectual Property Office, 'Technological Protection Measures (TPMs) Complaints Process' (*GOV.UK*, 3 November 2014) <https://www.gov.uk/government/publications/technological-protection-measures-tpms-complaints-process> accessed 17 March 2022.

[77] See Intellectual Property Office, 'Complaints to Secretary of State under s.296ZE under the Copyright, Designs and Patents Act 1988' (*GOV.UK*, 17 July 2015) <https://www.gov.uk/government/publications/complaints-to-secretary-of-state-under-s296ze-under-the-copyright-designs-and-patents-act-1988> accessed 17 March 2022. The data available on the website was released in 2014. An additional FOI request was sent to the UK IPO by the authors of this article in August 2020 which revealed that since 2015, two additional requests were filed, one of which was rejected (due to para 4 exemption) and the other resolved on a voluntary basis. Ironically, this latter request, the only one that has had a successful outcome in almost two decades, was based on the since repealed UK private copy exception.

[78] Commission (n 20) 114.

[79] See Michael W Carroll, 'Copyright and the Progress of Science: Why Text and Data Mining Is Lawful' (2019) 53 UC Davis Law Review 893; see also Geiger and others, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/Eu' (n 7) 33.

[80] Carroll (n 79) 954.

[81] See Geiger and others, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/Eu' (n 7) 33.

[82] In this sense see Dusollier (n 3) 987.

[83] art 4(2) contains a similarly worded provision 'Reproductions and extractions made pursuant to paragraph 1 may be retained for as long as is necessary for the purposes of text and data mining'.

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 dichotomies, the concept of intellectual creation, exceptions and limitations and ultimately the very same concept of work of authorship – all principles embed in fundamental rights such as freedom of expression, property and economic initiative. Modern copyright law emerged when it expressly departed from the censorial prototypes predating the early statutes of the 18th century. Since that moment, it has never been about controlling the use of information contained in a work. 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 for which does not possess the tools or the procedures. However, if a conscious decision was to be made to move towards this unprecedented function of copyright law, then this should be made explicit and be part of an open and transparent process not of a tacit, possibly surreptitious and probably unintentional effect.

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 under discussion here are intimately related to the advent of digital technologies and the EU's reaction to this advent. This reaction was evidenced by the roadmap proposed in the Green Paper of 1988[84] that 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 recent history that throughout the 1980s and 1990s, faced with the paradigm shifting changes brought about 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.[85] 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 which had lost the ability to channel innovation while maintaining incentives for investment and safeguarding 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 laws. This situation is exacerbated by the only partial power that the EU has (had) to regulate copyright, a power which largely relied on internal market attributions as a legal basis. As explained elsewhere,[86] this limited allocation of competences has led to a patchwork of at least 12 Directives (and two Regulations) which, with few exceptions, have harmonised EU copyright law 'vertically', i.e. only in relation to certain rights or certain subject matter.[87] One of the few directives that has taken a 'horizontal' approach (the 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 Member States ample discretion with regards to other aspects (mostly exceptions).[88] This approach has resulted in further fragmentation and uncertainty since having diverging rules within a market that proclaims to be single – as exemplarily illustrated in the *Donner* case[89] – is a natural generator of tensions in the legal, social and economic areas.

It is also in the light of these considerations that the CDSM Directive 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 CDSM Directive is timidly but clearly moving in the right direction regarding the second of the above identified tensions – thanks to the mandatory nature of several 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 order.

[84] Commission, 'Green Paper on Copyright and the Challenge of Technology: Copyright Issues Requiring Immediate Action' COM(88) 172 final.

[85] Dusollier (n 3); See also Commission (n 84); Martin Kretschmer, 'Digital Copyright: The End of an Era' (2001) 25(8) EIPR 333.

[86] *Ex pluris* Thomas Margoni, 'The Harmonisation of EU Copyright Law: The Originality Standard' in Mark Perry (ed), *Global Governance of Intellectual Property in the 21st Century* (Springer 2016); Ana Ramalho, 'Conceptualising the European Union's Competence in Copyright: What Can the EU Do?' (2014) 45 IIC 178.

[87] See Stefan Bechtold, 'Directive 2001/29/EC – on the harmonization of certain aspects of copyright and related rights in the information society (Information Society Directive)' in Thomas Dreier and P Bernt Hugenholtz, *Concise European Copyright Law* (Kluwer Law International 2016).

[88] P Bernt Hugenholtz, 'Why the Copyright Directive is Unimportant, and Possibly Invalid' (2000) 11 EIPR 499; Lucie Guibault, 'Why Cherry-Picking Never Leads to Harmonisation: The Case of the Limitations on Copyright under Directive 2001/29/EC' (2010) 1 Journal of Intellectual Property, Information Technology and E-Commerce Law 55, para 1.

[89] An illustrative case is Case C-5/11 *Criminal proceedings against Titus Donner* ECLI:EU:C:2012:370.

2. Non-original property

In order to portray 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 propertisation of non-personal data. This will offer additional support to the critique developed here concerning the inability (or unwillingness) to address technology as an opportunity. The SGDR naturally stands out as a unique EU device 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 the proprietarisation of data through IP rights was rejected in almost every 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's (at the time) nascent database market.[90] 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.[91] Nonetheless, as has been pointed out, it may be cumbersome to repeal EU legislation, including when, in the words of its drafters, it failed to deliver.[92]

Interestingly, in the recently published Data Act draft, Art. 35 offers some clarifications in relation to certain type of data, i.e. 'data in databases obtained or generated by means of physical components, such as sensors, of a connected product and a related service' (Recital 84), or in other words, machine generated data. Article 35 states that the SGDR does not apply to databases containing data obtained from or generated using a product or a related service. This is a welcome intervention, or better a 'clarification' as stated in Recital 84. However, whereas it has always seemed the correct reading of the SGDR that it cannot offer protection to machine generated data to the extent that it is generated data (and therefore the investment of the maker of the database is not in obtaining, verifying or presenting, but in creating data), machine generated data acquired by a third party through a substantial investment (e.g. payment of money), could become protected by an SGDR that rewards not the one who invested in the creation, but the third party who invested in the obtaining of this data. This has arguably been the proper reading of the creation versus obtaining dichotomy, whereby the 1996 legislator and the subsequent CJEU case law have attempted to avoid anti-competitive situations such as those originated by so-called single-source databases.[93] Therefore, reading that machine generated data do not qualify regardless of whether they are created or obtained – at least in relation to the users' rights to access, use and share their data (Arts. 4 and 5) – and therefore arguably limiting the operativity of SGDR in relation to a specific type of data, is remarkable and certainly an approach that meets the many demands from scholars, industry and creators to domesticate this untameable right.

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

A final element in the account of the EU approach to data propertisation and its implications for technology is, similarly to the proposed Data Act, located 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 (PSBs).[94] It is beyond the scope of this article to explore such an important legislative intervention in detail, but 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 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'.[95] 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 PSBs and, since the

[90] 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.' (Commission, 'Commission Staff Working Document: Executive Summary of the Evaluation of Directive 96/9/EC on the legal protection of databases' SWD(2018) 147 final).

[91] See Case C-30/14 *Ryanair v PR Aviation* ECLI:EU:C:2015:10; for a detailed discussion see Maurizio Borghi and Stavroula Karapapa 'Contractual restrictions on lawful use of information: sole-source databases protected by the back door?' (2013) 37 EIPR 505.

[92] See Martin Husovec, 'The Fundamental Right to Property and the Protection of Investment: How Difficult is it to Repeal New Intellectual Property Rights?' in Christophe Geiger (ed), *Research Handbook on Intellectual Property and Investment Law* (Edward Elgar 2020).

[93] Hugenholtz (n 14); Peter K Yu, 'Data Producer's Right and the Protection of Machine-Generated Data' (2019) 93 Tulane Law Review 859; Josef Drexl, 'Designing Competitive Markets for Industrial Data – Between Propertisation and Access' (2017) 8 Journal of Intellectual Property, Information Technology and E-Commerce Law 257; Herbert Zech, 'Data as a Tradeable Commodity – Implications for Contract Law' in Josef Drexl (ed), *Proceedings of the 18th EIPIN Congress: The New Data Economy between Data Ownership, Privacy and Safeguarding Competition* (Edward Elgar 2017).

[94] 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 [2019] OJ L172/56.

[95] See European Commission, 'European legislation on reuse of public sector information' <https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information> accessed 1 July 2021.

new Directive, to certain public undertakings. However, similar hints, albeit timid, may be seen in the new wave of legislative interventions in the field of data and markets, in particular in the already mentioned Art. 35 of the proposed Data Act, in Art. 10 of the proposed AI Regulation (data quality) and Art. 29 of the proposed DSA (recommender systems).[96] All these approaches, certainly underpinned by different policy considerations and geared towards a plurality of regulatory objectives, seem to share one common principle: data transparency. It is perhaps not a purely provocative exercise to consider whether a proper regulatory framework would be one where similar rules in relation to computational uses should apply generally to any type of data, works or AI.[97] Whereas there would certainly be strong opposition to such a framework, it appears even more controversial 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.

V. Conclusions

This article intends to offer a novel perspective into less apparent but far-reaching implications of Arts. 3 and 4 Copyright in the Digital Single Market Directive (CDSM). First the regulation of data analytic technologies was located in the copyright *acquis*, specifically with respect to the *sui generis* database right (SGDR), the exception for temporary acts of reproduction under Art. 5(1) of the InfoSoc Directive, and the evolving case law of the Court of Justice. Secondly, the practical scope of the new text and data mining exceptions under Arts. 3 and 4 CDSM was assessed with this context in mind. Thirdly, we revealed an underlying proprietary conception of data that seems to emerge, perhaps as an unintended consequence, from the EU's approach to regulating fast evolving digital technologies.

We find a deep tension in the law's treatment of information extracted from copyright works, both theoretically and for the innovation goals of the EU. With respect to foundational concepts of copyright, the law protects the original expression of ideas, not ideas themselves, nor mere facts or data.[98] Accordingly, text and data mining should not be considered a copyright infringement, but a matter external to copyright's scope.[99] It follows that a copyright *exception* is a problematic intervention to regulate the use of unprotected ideas, principles, facts and data, often contained in literary works or other types of texts (text mining) or in structured and/or unstructured datasets (data mining).[100] With respect to the EU's innovation goals, the provisions of the CDSM Directive paradoxically favour the development of biased AI systems due to price and accessibility conditions for training data that offer the wrong incentives. To avoid licensing, it may be economically attractive for developers to train their algorithms on older, less accurate, biased data, or import AI models already trained on unverifiable data.

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 mere facts and data contained in works, including low original works, and in non-original other subject matter (SGDR). 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, or are there other areas 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. 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 Art. 5(1) InfoSoc Directive. 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 Art. 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 the light of recent CJEU case law that appears to establish

[96] Commission, '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; Commission, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC' COM(2020) 825.

[97] 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' (n 93).

[98] eg art 2 WIPO Copyright Treaty (20 December 1996) 2186 UNTS 121, 36 ILM 65 (WCT), art 9(2) WTO Agreement on Trade-Related Aspects of Intellectual Property Rights (15 April 1994) 1869 UNTS 299, 33 ILM 1197 (TRIPS) and Recital 8 CDSM.

[99] European Copyright Society (n 4); Sag (n 38); Craig (n 38).

[100] At the level of international law, WIPO's interpretation of the Berne Convention states that '[t]he 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' (Claude Masouyé, *Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971)* (World Intellectual Property Organization 1978)). Ricketson and Ginsburg similarly state that protection offered by the Berne Convention 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 (Sam Ricketson and Jane Ginsburg, *International Copyright and Neighbouring Rights* (2nd edn, OUP 2010) 407). 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 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' (Case C-406/10 *SAS Institute v World Programming Ltd.* ECLI:EU:C:2012:259, paras 66-67).

that any fundamental rights limitation to copyright must be found within Art. 5.[101]

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 the object of horizontal legislative harmonising interventions.[102] 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*[103] and *Pelham*,[104] it seems that the space for Member States 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.[105] Member States interested in enabling computational uses should consider this option.

The Open Data Directive briefly discussed in section III and especially the national Open Access guidelines it mandates will likewise represent an 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 is a timid recognition of this diagnosis in the recent AI Regulation proposal. Other recent proposals (i.e. the Data Governance Act,[106] Data Act,[107] DSA[108] and DMA[109]) may be more promising as a new approach in this area.[110]

Finally, extra-EU countries which are not bound by the rigidity of EU copyright law, can be divided into two main categories. Those which have enacted a broad and/or flexible approach (US, Canada, Singapore, South Korea, Japan, Israel[111]), and those which have not yet done so (e.g. South American countries[112]). 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 autonomy 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[113]), the future seems a choice between the need to maintain a level playing field with the EU neighbour and the attractiveness of regulatory competition, including a modern, dynamic and accountable regulation of AI.[114]

This paper shows that technology is not exogenous to (copyright) law. 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 becomes part of the technology itself, how it will be governed, who will have access to it, at what costs and under which conditions.[115] When this technology is AI, with its endless potential applications, this poses a legislative conundrum. Paraphrasing a famous expression, 'digital artefacts have politics' and AI perhaps more than others.[116] The CDSM Directive, conceiving of data analytic acts as in need of an exception from proprietary claims, gets it radically wrong.

ACKNOWLEDGEMENTS

This 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 article draws on earlier work presented at the European Policy for IP Association (EPIP Berlin, 7 September 2018; EPIP Madrid, 9 September 2021) and the 5th Global Congress on IP and the Public Interest (Washington, 27 September 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: Thomas Margoni and Martin Kretschmer, 'The Text and Data Mining Exception in the Proposal for a Directive on Copyright in the Digital Single Market: Why It Is Not What EU Copyright Law Needs' (*CREATe*, 25 April 2018) <https://www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/> accessed 17 March 2022.

[101] See *Pelham v Hütter* (n 18); Martin Senftleben, 'Flexibility Grave – Partial Reproduction Focus and Closed System Fetishism in CJEU, Pelham' (2020) 51 IIC 751.

[102] P Bernt Hugenholtz and Martin Senftleben, 'Fair Use in Europe: In Search of Flexibilities' (2012) Amsterdam Law School Research Paper No 2012-39, Institute for Information Law Research Paper No 2012-33.

[103] Case C-419/13 *Allposters v Pictoright* ECLI:EU:C:2015:27.

[104] *Pelham v Hütter* (n 18).

[105] Thomas Margoni, 'The digitisation of cultural heritage: originality, derivative works and (non) original photographs' (2014) <https://ssrn.com/abstract=2573104> accessed 17 March 2022.

[106] Commission, 'Proposal for a Regulation of the European Parliament and of the Council on European data governance (Data Governance Act)' COM(2020) 767 final.

[107] Commission, 'Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act)' COM(2022) 68 final.

[108] Digital Services Act (n 96).

[109] Commission, 'Proposal for a Regulation of the European Parliament and of the Council on contestable and fair markets in the digital sector (Digital Markets Act)' COM(2020) 842.

[110] Julie Baloup and others, 'White Paper on the Data Governance Act' (2021) CiTiP Working Paper; João Quintais and Sebastian Felix Schwemer, 'The Interplay between the Digital Services Act and Sector Regulation: How Special is Copyright' [2022] European Journal of Risk Regulation (forthcoming) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3841606> accessed 17 March 2022.

[111] Flynn (n 38).

[112] Matías Jackson Bertón, 'Text and Data Mining Exception in South America: A Way to Foster AI Development in the Region' [2021] GRUR International 1145.

[113] Martin Kretschmer, 'UK Sovereignty: A Challenge for the Creative Industries' (*CREATe*, 21 July 2020) <https://www.create.ac.uk/blog/2020/07/21/uk-sovereignty-a-challenge-for-the-creative-industries/> accessed 17 March 2022.

[114] Perhaps, an updated TDM exception (UK Copyright, Designs and Patents Act 1988 s 29A) 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.

[115] cf Benkler (n 11); Yochai Benkler, 'Power and Productivity: Institutions, Ideology, and Technology in Political Economy' <http://www.benkler.org/Benkler_Power&Productivity.pdf> accessed 17 March 2022.

[116] Langdon Winner, 'The Whale and the Reactor: A Search for Limits in an Age of High Technology' (1980) 109 Daedalus 121.