CREATe Working Paper 2018/12 (December 2018)

Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?

Authors

Thomas Margoni
CREATe
University of Glasgow

CREATe Working Paper Series DOI: 10.5281/zenodo.2001763

This release was supported by the RCUK funded *Centre for Copyright and New Business Models in the Creative Economy (CREATe)*, AHRC Grant Number AH/K000179/1.

Artificial Intelligence, Machine learning and EU copyright law:

Who owns AI?

Thomas Margoni

10 Nov 2018

Table of Contents

I. Introduction

Within the broad field of Artificial Intelligence (AI), Machine Learning (ML) looks at improving the performances of computers in executing tasks for which they were not specifically pre-programmed. Applied to the field of Natural Language Processing (NLP), ML helps computers to autonomously learn tasks such as the recognition, understanding and generation of natural language (i.e. the language spoken by humans). In other words, ML applied to NLP refers to the ability of humans to interact with computers in the same way in which humans interact among themselves. On the part of the computers this implies being able to understand human language, to understand its meaning, and to interact with it thorough the generation of new language.

Examples of these applications are very common in the current information society. Digital devices including phones, tablets, watches and an increasing number of home furniture, are nowadays equipped with personal assistants (generally called "AI") which can be activated and can communicate through voice. More often than what one may think, when calling the consumer support service of a growing number of companies – or more commonly when contacting them via social networks – it is not a human who answers the phone call, replies to the tweets or other notifications. It has been estimated that as much as 40% of these "answers" come from AI bots which have learned to speak a human language (e.g. English, Italian or any other language).

This study focuses specifically on this element, i.e. how computers learn a language. The reason is straightforward: when humans learn a new language they usually store the training information (e.g. the text book used to learn it) as an electrochemical trace in the area of the brain dedicated to language. Humans do not need a copyright exception in order to store that copy. Traditional copyright law and theory (in addition to common sense) have that this activity is outwith the copyright realm. However, it is far from clear that when a computer makes the corresponding digital copy of training material in order to learn a language this activity is likewise excluded from the copyright domain. On the contrary, normally any digital copy, temporary or permanent, in whole or in part, direct or indirect, has the potential to infringe copyright. Let's just recall here briefly and for the sake of the argument that the temporary copy of a webpage made in the cache memory of computers and tablets is only possible thanks to an exception. Otherwise, browsing the internet would most likely be a copyright infringement (or need to be justified under other theories and legal doctrines such as implied licence or estoppel).

Normally, computers learning natural languages need to "train models" using specific ML algorithms. The trained models represent the "memory" of a machine which has learned a language. The machine will use this memory to learn more or better linguistic skills and will use it to formulate its own statements. But how is this memory created? Or in NLP parlance, how are the models trained? Usually, models are trained on corpora, that is to say on literary works often "available on the internet". In more precise terms, a typical ML/NLP workflow is as follows: a certain amount of corpora are identified on the basis of their relevance (language, topic, register, etc). Once identified, they have to be converted into a suitable file format and annotated (unless the original corpora are already annotated, an increasingly valuable resource). Following annotation, a ML algorithm analyses the text. In order to do this, normally one or more temporary copies of the original text are made. These copies are analysed by the algorithm and depending on the type of algorithm certain elements (syntactic and grammatical rules, statistical recurrences, words, semantic correlations, sentiment tags, etc) are inferred from the corpora and recorded in the model. The model is then used by other algorithms whose scope is not to learn a language but to speak it or write it.

The question thus becomes the following: is the act of training a model for ML purposes a copyright relevant activity? The answer to this question is not only relevant in terms of copyright law and theory, but more broadly in terms of innovation policy as it has the potential to determine who has to ask whom for what permission in order to perform ML functions. In other words: who owns AI? In more precise terms, the research question of this short contribution will focus on the act of training a model for ML/NLP purposes and attempts to answer the question of whether this act infringes copyright and in particular the right of reproduction. In addition to this, the contribution also intends to explore whether there are other rights that may be infringed, in particular the right of adaptation, and

thus determine whether a ML trained model can be considered a creative adaptation of the original corpora. The reference legal framework will be EU copyright law, with occasional reference to domestic law when necessary.

II. Technological basics

This section will briefly identify the steps required in a typical NLP workflow. It will only focus on the main features, in particular those relevant from a legal point of view. However, it should be borne in mind that different fields of ML as well as different algorithms in the same field of ML, operate differently and thus the technical steps and the legal consequences may vary (1).

In NLP, as well as in most text analytic fields, algorithms "learn" abstract probabilistic models from texts annotated with labels (e.g. named entities, part-of-speech tags, sentiment tags, etc.) in order to predict such labels on unseen text. They do this by storing the relevant information in a separate file, the "trained model". Models are constructed through a training process involving a learning algorithm and training data to learn from. The model captures abstract probabilistic characteristics from the training data, which can then be used to predict the learned labels on unseen data. In general, constructing a model consists of the following steps: (1) corpus compilation (2) corpus preprocessing, (3) corpus annotation, and (4) training of the model (2). A fifth step can be identified in the permanent creation of a file (the trained model).

During the corpus compilation (step 1) a set of linguistic resources (often literary works, but also databases, dictionaries, thesauri, etc.) are identified and collected in order to capture specific aspects of real world language. For best results, the ML algorithm must be trained on a set of texts that is close to the corpus to which it is later applied, for example it must be of the same language and domain and annotated with the appropriate labels. This set of texts are selected and obtained from one or more sources such as publishers repositories, journals, web sites, etc.

During the second phase (step 2), the corpus is preprocessed, which involves – usually automated – processes to convert the textual content acquired from resources that are usually in a format that cannot be directly processed by the ML algorithms (such as PDF or HTML) into formats that can be processed, usually plain text file formats. During this phase images, tables and other non textual elements are likewise removed.

The third is the annotation phase (step 3), where the plain text is "enriched" with labels relevant to the task that needs to be performed. These labels are usually classified in categories (grammatical, morphological, syntactic, etc.). The annotator assigns to text segments (e.g. words, phrases) the appropriate labels. There are inventories of labels that define the type and content of annotations and annotators usually follow detailed instructions on how to select and apply the correct annotation type and label to the identified units.

The fourth phase is the training of the model properly conceived (step 4). The training software programme implements a ML algorithm. This algorithm analyses the annotated corpus and extracts a set of defined probabilistic, statistical, grammatical and syntactical characteristics which are eventually saved in a file.

The fifth and final phase (step 5) is the creation of the trained model as a permanent file. The trained model can be seen as a sort of abstraction of the annotated corpus based on statistical observations which can then be used with a second software tool to predict the learned labels on unseen text (3).

---

1 The technical knowledge herein contained is based on the work of the H2020 OpenMinTeD project (2016 – 2018) and on the results produced, among which is particularly relevant here: Richard Eckart de Castilho, Giulia Dore, Thomas Margoni, Penny Labropoulou, Iryna Gurevych, A Legal Perspective on Training Models for Natural Language Processing, Conference Paper, March 2018.

2 ID.

3 See fn 1 above.

These five steps are representative of a normal workflow in the field of ML applied to NLP. Their clear identification is particularly useful as these five steps will be employed during the legal analysis to determine when a copy (temporary or permanent) is created and how. They will additionally offer the possibility to compare a typical ML workflow with a similar five-fold categorisation developed by the EUCJ in its case law. It should be kept in mind, however, that while the process and the steps are usually similar, at least within the same field of NLP, different algorithms and even more different fields can behave differently and could potentially lead to different results.

III. Copyright basics

This section first explores the copyright status of the resources used in ML/NLP activities and then analyses the relevance of selected rights of exploitation during the ML/NLP identified steps. In particular, it will be necessary to identify the copyright nature of the resources used in ML/NLP (what kind of protected subject-matter are involved), the level of originality (not just in relation to the entire work but also of segments thereof), and what kind of acts (rights of economic exploitation) are involved.

III.1 Literary works and original databases

Most corpora employed in NLP consist of web pages, publications, articles, newspaper texts, blog posts or even tweets, annotated or not. All of these resources possess the potential to be protected by copyright law. To be eligible for copyright protection a work must be original (4). A brief analysis of the originality standard in international and EU law is developed in the next section with the goal of determining whether full texts or parts of texts can be considered original (and if yes, how short that part of text can be).

III.1.a. Originality in international law

Originality is an essential requirement of copyright law; only works that show some minimum amount of this attribute attract protection. However, generally speaking, originality lacks a precise statutory definition (5). In the Berne Convention, the reference international copyright instrument, the requisite of "intellectual creations" possesses a central yet implicit role. Intellectual creations are required not

4 Other qualifying requirements (e.g. fixation in a material form where relevant, sufficient connection to a Berne territory, absence of exclusion for public policy grounds), are not relevant here and will not be discussed.

5 See Ricketson S., Ginsburg J., International Copyright and Neighbouring Rights – The Berne Convention and Beyond, OUP, 2005, 8.05; Bently L., Sherman B., Gangjee D., Johnson P., Intellectual Property Law, OUP, 5th Ed., 2018, 93; Gervais D., Judge E., Intellectual Property: The Law in Canada, Carswell (1st Ed.), 2005, 16; Goldstein P., Hugenholtz B., International Copyright – Principles, Law, and Practice, 3rd Ed., OUP, 2013, 192; Ginsburg J., No 'Sweat?' copyright and other protection of works of information after Feist v. Rural Telephone, in Columbia Law Review, 1992, 92:338–388; Gervais D., Feist Goes Global: A Comparative Analysis of the Notion of Originality in Copyright Law, in 49 Copyright Soc'y U.S.A. 949, 2002; Gravells N., Authorship and originality: the persistent influence of Walter v. Lane, in I.P.Q., 2007, 3:267; Schricker, Farewell to the "Level of Creativity" (Schöpfungshöhe) in German Copyright Law?, in IIC 1995, 41.

only in relation to article 2(5), i.e. to collections of literary and artistic works (6), but they are a necessary element of all the subject-matter covered by article 2 ("Protected Works") (7). It has been said that an explicit identification of originality in "intellectual creations" was indispensable only for the case of article 2(5), because the originality inherent in collections, as opposed to that in the works collected, "may not be as readily discernible" (8). Accordingly, not only collections but also any other scientific or literary work such as books, lectures, musical compositions, songs, works of photography, and sketches, have to possess the required type of originality in order to comply with Berne standards (i.e. they have to be intellectual creations). This doctrinal reconstruction of the meaning of article 2 also corresponds to the view the EUCJ expressed in the Infopaq I case (9). Nevertheless, what this exactly entails, how high – or low – the level of originality is, and what the tests, standards, and elements are that can fill-up the concept of "intellectual creation" remains a matter for national legislatures and courts (10).

III.1.b. Originality in the *Acquis Communautaire*

As said, the standard of originality has historically been a matter of domestic law. This was and still is true at the international level (e.g. within the Berne Convention, which only requires works to be an intellectual creation) and was true for the EU. However, at the EU level things started to change as far back as 1991 with the first Directive in the field of copyright – the Computer Programs Directive (11) – which harmonised the originality standard at the level of the author's own intellectual creation. However, in that legal instrument, as well as in the successive Term of Protection and Database Directives (12), the originality standard was harmonised only "vertically", i.e. only with regard to the specific subject-matter regulated by the aforementioned acts of EU secondary legislation.

The reason for this sectorial harmonisation of copyright law can probably be found in the absence of a clear and direct attribution of powers to the EU to regulate copyright (13). Until recently the main basis for EU intervention in the field of copyright were articles 26 and 114 of the Treaty on the Functioning of the European Union (TFEU), which have given the EU the competence to respectively adopt measures with the aim of establishing or ensuring the functioning of the internal market and the approximation of the laws of Member States. The absence of a clear attribution of powers to regulate copyright contributed to the fragmentary and subject-matter specific approach taken by EU copyright directives, especially during 1990s (14).

---

6 "... which, by reason of the selection and arrangement of their contents, constitute intellectual creations"; see Berne Convention for the Protection of Literary and Artistic Works, of September 9, 1886, Paris Text, Art. 2(5).

7 "A line therefore seems to run from article 2(5) through article 2(3) to article 2(1) as follows: ''original translations, adaptations, etc.'' under article 2(3) and collections of works that are ''intellectual creations'' under article 2(5) are to be protected as ''literary and artistic works'' under article 2(1), suggesting that both originality and intellectual creation are correlative and implicit requirements for literary and artistic productions that otherwise fall under article 2(1)", see Ricketson S., Threshold Requirements for Copyright Protection under the International Conventions, in 2009, W.I.P.O. Journal, No. 1, 51 – 62 2009, 57.

8 ID.

9 Case C-5/08 *Infopaq International A/S v. Danske Dagblades Forening* [2009] E.C.R. I-06569.

10 See Ginsburg J., No 'Sweat?' copyright and other protection of works of information after Feist v. Rural Telephone, in Columbia Law Review, 1992, 92:338–388.

11 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).

12 Respectively, Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights (codified version), as amended and Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

13 See generally Benabou V., Droits d'Auteur, Droit Voisins et Droit Communautaire, Brussels, 1997.

14 This can be observed in different documents of the EC. In the *Green paper on copyright and the challenge of technology: copyright issues requiring immediate action*, for example, it can be read that the "Commission concluded that a directive on the legal protection of computer programs is a necessary step for the completion of the internal market" and that "the creation of a European information services market, currently divided by juridical and linguistic barriers, is of prime importance" (European Commission 1988, 5.4.1, 6.2.1).

However, this changed in the following decade. In at least five landmark decisions (15) the EUCJ took the opportunity to clarify (or to establish) that the originality standard until then contained in the above mentioned three Directives did not apply only "vertically", i.e. only in relation to the subject-matter harmonised within those directives, but horizontally. In other words, the "author's own intellectual creation" was not limited to software, photographs and databases, but extended "horizontally" to all subject-matter covered by the Information Society Directive (16) or better by the Berne Convention (with the exclusion of registered designs) (17).

Additionally, the EUCJ also developed a set of interpretative elements to be used when determining whether a work is original. EU originality is achieved when authors can exercise free and creative choices and put their personal stamp on the work (18). On the contrary, it is not present when an expression is determined by technical or functional rules, such as when there is only one way to express an idea, or the expression is predetermined by a specific goal or constrained by narrow rules that leave no space for free and creative choices (19).

However, the wording employed by the Court should not be taken as to suggest that the "author's own intellectual creation" requirement is placed at a particularly high level. In fact, a closer look at the facts decided probably indicates a different outcome. The EUCJ recognised protection – or rather determined that protection could not be excluded, something to be verified by the national referring court – of an eleven word extract (Infopaq International v. Danske Dagblades Forening), to a portrait photograph (Eva-Maria Painer v. Standard VerlagsGmbH), to a graphical user interface (Bezpečnostní softwarová asociace v. Ministerstvo kultury) and to a programming language (SAS Institute v World Programming), provided that they constitute the author's own intellectual creation. Excluded from protection are match fixtures (Football Dataco v. Yahoo!) and sports games as such (Football Association Premier League v. QC Leisure) due to the lack of free and creative choices (20).

It has been effectively argued that the new standard created by the Court gives much more emphasis to the qualitative rather than the quantitative type of authorial contributions (21). Accordingly, it cannot be excluded that even single sentences, if original, can be the object of copyright protection, something that for example was excluded under UK law before EU harmonisation, at least with regard to titles (22). In conclusion, it can be assumed that most corpora used for ML/NLP, especially those of a literary and scientific character, such as scholarly articles, are protected by copyright. Additionally, parts of those literary works, sometimes as short as eleven consecutive words if original in their own right, could likewise be considered protected and their reproduction reserved.

III.2. Original and non-original databases

Under EU law, as well as under the law of many other jurisdictions, databases are defined as collections of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means (23). Copyright exists if originality is found in

15 See *Infopaq International v. Danske Dagblades Forening* [2009]; *Bezpečnostní softwarová asociace v. Ministerstvo kultury* [2010]; *Football Association Premier League v. QC Leisure* and *Karen Murphy v. Media Protection Services* [2011]; *Eva-Maria Painer v. Standard VerlagsGmbH* [2011]; *Football Dataco v. Yahoo!* [2012]).

16 See 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.

17 For an analysis of the originality standard in relation to registered and unregistered designs see Margoni, T., Design rights and 3D printing in the UK: Balancing innovation and creativity in a (dis)harmonised and fragmented legal framework, in Mendis, D., Lemley, M. and Rimmer, M. (eds.), *3D Printing and Beyond: The Intellectual Property and Legal Implications Surrounding 3D Printing and Emerging Technologies,* Edward Elgar Publishing, 2018.

18 E.g. (*Football Dataco v. Yahoo!* [2012], 38; *Infopaq International v. Danske Dagblades Forening* [2009], 45; *Bezpečnostní softwarová asociace v. Ministerstvo kultury* [2010], 50; *Eva-Maria Painer v. Standard VerlagsGmbH* [2011], 89, 92).

19 E.g. *Football Association Premier League v. QC Leisure* and *Karen Murphy v. Media Protection Services*, 98 [2011]; *Bezpečnostní softwarová asociace v. Ministerstvo kultury* [2010], 49; *Football Dataco v. Yahoo!* [2012], 39).

20 MARGONI, T. (2016). The harmonisation of EU copyright law: The originality standard. In Mark Perry, editor, Global Governance of Intellectual Property in the 21st Century, pages 85–105. Springer, Switzerland.

21 See Bently L., Sherman B., Gangjee D., Johnson P., Intellectual Property Law, OUP, 5th Ed., 2018, 102.

22 ID., at 64.

23 Directive 96/9/EC, OJ L 77, 27.3.1996, Article 1.

the selection or arrangement of the content, i.e., the "intellectual creation" has to be found in the database structure. Consequently, copyright in databases protects only the structure and does not extend to the content. The content, in turn, can be autonomously protected by copyright (a database of scholarly articles), related rights (a database of sound recordings), or be in the public domain (a database of unprotected facts or of e.g. medieval texts)

In addition, EU law, unlike the law of most other countries, has introduced a new right protecting non-original databases when a substantial investment has been put in the obtaining, verification or presentation of the data. This is an important element of the protection afforded to non-original databases: if the investment has gone into the creation and not into the obtaining, verification or presentation of data, then this is not enough to trigger protection (24). In this case, the database maker (usually the person or entity who bears the financial risk) enjoys a *sui generis* database right (SGDR), which protects the content of the database from substantial extractions. In other words, even databases of unprotected facts could become the object of a proprietary right that extends to the database content in the light of the aforementioned substantial investment (25). Therefore, certain collections of corpora (e.g. the database of Institute X that over the years has collected public domain corpora investing substantial time and work in the process) could be protected by the SGDR and restrict the reuse of substantial parts (or repeated reuse of insubstantial parts). Copyright and database rights are probably the two most relevant rights potentially covering the annotated and unannotated corpora forming the basis for any ML training activity (26).

III.3. Adaptations, translations and other creative modifications

Adaptations, translations and other alterations have received some attention at the international level but have been excluded from the harmonisation process under EU copyright law, with the limited exception of the Computer Program and of the Database Directives.

An example of this international attention can be seen, for instance, in the Berne Convention where a few articles indicate that some works, although based on other works, deserve autonomous (yet derivative) protection. At the EU level, despite the absence of a general right of adaptation, some useful insight may be acquired by the interplay between a broad right of reproduction and the unharmonised right of adaptation (e.g. which adaptations are not, at least in part, a reproduction?) and by a relatively recent decision by the EUCJ.

III.3.a. The international landscape

---

24 Hugenholtz, BP, Something Completely Different: Europe's Sui Generis Database Right, in Frankel S. and Gervais D. (Eds.), The Internet and the Emerging Importance of New Forms of Intellectual Property, Information Law Series, Vol. 37, Kluwer Law International, 2016, 205 – 222

25 Hugenholtz, BP, Something Completely Different: Europe's Sui Generis Database Right, in Frankel S. and Gervais D. (Eds.), The Internet and the Emerging Importance of New Forms of Intellectual Property, Information Law Series, Vol. 37, Kluwer Law International, 2016, 205 – 222; Guibault L. et al. (Eds.), (2013). Safe to be open. Study on the protection of research data and recommendations for access and usage. Universitätsverlag Göttingen.

26 Stamatoudi I., Torremans P., (Eds), Copyright in the New Digital Environment: V. 8 The Need to Redesign Copyright, Sweet & Maxwell, 2000; Borghi, M. and Karapapa, S., Copyright and Mass Digitization: A cross-jurisdictional perspective. OUP, 2013; Truyens M., Van Eecke P., Legal aspects of text mining. Computer Law & Security Rev., 2014, 30(2):153–170; Triaille, J.-P., et al, Study on the legal framework of text and data mining (tdm), 2014, Luxembourg: Publications Office;

Adaptations, translations and other alterations (also known as derivative works) are those works which are based on pre-existing ones. From this point of view derivative works are not "primary" works, such as those listed in article 2(1) Berne Convention, but "secondary" works (27).

When an adaptation is sufficiently original and contains an "intellectual creation" additional to that of the original work, the protection is assimilated to that afforded to original works by article 2(1) Berne Convention. Nevertheless, this is without prejudice to the copyright in the pre-existing work. Accordingly, in order to create an adaptation the authorisation of the right holder of the primary work is necessary to avoid liability for copyright infringement, unless the use is covered by a specific exemption or the pre-existing work has fallen into the public domain. If the derivative work is created in absence of authorisation or outside the cases admitted by law, and therefore constitutes an unauthorised use, it generally still attracts protection (28). The U.S., however, have a peculiar provision whereby a derivative work unlawfully created does not benefit from copyright protection (29).

Not every case of creation of a work based on another work constitutes an act of adaptation or alteration requiring authorisation. In order to constitute a secondary work the elements constituting the intellectual creation in a primary work need to be reproduced, adapted or altered in the secondary work. Consequently, if a work is only inspired by the idea expressed in a previous work, there is simply no act of derivation and accordingly no authorisation is required. In these cases, the resulting work, if an intellectual creation in its own right, is protected as an original (primary) work.

Three types of derivative works are specifically regulated by the Berne Convention: translations, arrangements of music, and adaptations and other alterations. Translations commonly refer to changing a literary or dramatic work from one language into another (30). Whether the term language includes only "traditional" human languages, or also includes modern forms of "artificial" languages such as computer programming languages is ultimately a matter to be decided by domestic law, but in principle not incompatible with Berne's broad definition (31). Arrangements of music generally involve skills such as adaptation and transcription of a musical part for one instrument into that for another, or the addition of rhythmic parts to a melody (32). The third category, adaptations and other alterations, represents a sort of residual class whose scope is to cover all the elaborations "considered to fall within the scope of adaptation" such as "dramatisations and choreographic or mime adaptations, the making of prose versions of dramatic works, the rendition of a literary or dramatic work into a dramatic-musical form and so on" (33). This open-ended definition, however, encounters a precise limit. Only the adaptations and alterations that involve new authorial contributions deserve protection. Changes of small sections, additions or omissions of material not accompanied by new original contributions do not trigger the protection afforded to original adaptations (34).

Furthermore, it must be noted that translations, adaptations and other alterations not only constitute protectable subject matter in their own right as established by article 2(3). The Convention explicitly recognises to authors of literary or artistic works the enjoyment of the exclusive right of authorising adaptations, arrangements and other alterations of their works (article 12) and that authors of literary and artistic works shall enjoy the making and authorising of the translation of their works throughout the term of protection of the original works (article 8). Moreover, authors of dramatic, dramatic-musical and literary works enjoy, during the full term of their rights in the original works, the rights of authorising the public performance/recitation and communication to the public of the translations of their works (articles 11-2 and 11ter-2). Finally, authors of literary or artistic works have the exclusive right of authorising the cinematographic adaptation and reproduction of their works, and the distribution, public performance and communication to the public of the works so adapted or reproduced. The adaptation into any other artistic form of a cinematographic production derived from literary or artistic works shall, without prejudice to the right in the cinematographic production, remain subject to the authorisation of the authors of the original works (article 14).

27 See Ricketson & Ginsburg, 2006, at 8.75.

28 Goldstein & Hugenholtz, 2013, at 6.1.2.7.

29 See U.S. Copyright Act 1976 Sec. 103(a).

30 Ricketson & Ginsburg, 2010, at 8.78. Goldstein & Hugenholtz, 2013, at 6.1.2.7.

31 ID.

32 ID., at 8.79.

33 ID., at 8.81; Masouyé C., Guide to the Berne Convention, WIPO, 1978, at 76-7.

34 Goldstein & Hugenholtz, 2013, at 8.81.

III.3.b. The *Acquis Communautaire*

As seen above, many of the EU directives in the field of copyright regulate rights of economic exploitation such as the reproduction right, the distribution right and the right of communication to the public. Traditionally, this was done through a "vertical" approach such as in the case of the Software and Database directives. The Information Society Directive took instead a broad horizontal approach and offered fully harmonised definitions of the right of reproduction, communication to the public and distribution.

However, the adaptation right has remained untouched by these vertical and horizontal harmonising interventions. Unique exceptions to this lack of harmonisation are found in the Computer Program and Database Directives where both adaptations and translations are explicitly mentioned (35).

III.3.b.1. The Software and Database Directives

Recital 15 of the Software Directive states that translations, adaptations or transformations of the computer program code constitute an infringement of the exclusive rights of the author, unless these acts are necessary to achieve interoperability. Preparatory design materials are also protected, since all subsequent programming steps can be considered adaptations of the preceding stages (36). Furthermore, article 4(b) states that the rights vested in the author of a computer program include the translation, adaptation, arrangement and any other alteration of a computer program, subject to the exceptions listed in articles 5 and 6 (37).

Regarding the Database Directive, article 5(b) establishes that in respect of the expression of the database which is protected by copyright, the author has the exclusive right to translate, adapt, arrange and perform any other alteration. It must be noted that article 5 deals with copyright protection and therefore translations and adaptations refer to the selection or arrangement of the database and not of the data itself. It has been stated that the translation of the structure of a database is "hardly imaginable" (38).

It follows that, from a EU perspective – and with the exception of software and databases – translations, adaptations and transformations of works into new expressive forms are left to Member States' discretion. The reasons for such a gap in the full harmonisation of the rights of economic exploitation can be attributed to the "borderline" nature of the right of adaptation. The adaptation of a work often requires its, at least partial, reproduction. Yet, there is more: as seen above, a simple reproduction of a work accompanied by small non-creative modifications, does not lead to the creation of a protected adaptation. A derivative work requires an authorial original contribution of the intervening author; however, until recently the concept of originality was not harmonised. Hence, a proper harmonisation of the adaptation right without simultaneously harmonising the threshold of originality could have caused unpredictable consequences (39).

35 See Art. 4(1)(b) Software Directive and Art. 5(b) Database Directive.

36 Walter & von Lewinski, 2010, at 5.1.39

37 See Recital 15 and Art. 4 of the Software Directive; Walter & von Lewinski, 2010, at 5.1.39; Samuelson, Vinje & Cornish, 2012.

38 See Walter & von Lewinski, 2010, at 9.5.9.

39 See van Eechoud M., Hugenholtz B., van Gompel S., Guibault L., Helberger N., Harmonizing European Copyright Law – The Challenges of Better Lawmaking, Kluwer, 2009.

III.3.b.2 Adaptations and integrity

Moreover, the economic right of authorising adaptations is intimately connected with the moral right of integrity. Moral rights, similarly to the concept of originality until 2009, are not object of EU law harmonisation. As usual, a common reference can be found in article 6-bis of the Berne Convention which establishes that authors shall have the right to claim authorship of the work and to object to any distortion, mutilation or other modification or derogatory action which would be prejudicial to their honour or reputation (40). These rights are independent from the economic rights, shall remain with the author even after the transfer of the rights of economic exploitation and shall be maintained at least until the expiry of the economic rights (41). At the Member State level additional moral rights may be recognised. Usually, the right of disclosure, the right to retract (or withdraw) the work and the right of access are commonly found in EU Member States (42).

Of particular relevance for the present purpose is the second right recognised by the article 6-bis(1), also known as the right of integrity. The strong connection between the (economic) right that regulates the creation of adaptations and creative elaborations and the (moral) right that protects the integrity of the work is clear; also clear is the potential conflict caused by the transferable nature of the former and the non-transferable nature of the latter.

In conclusion, it can be said that the absence of a harmonised right of adaptation was probably justified by the concomitant absence of harmonisation of the originality requirement and of the moral right of integrity. Such an absence was partially compensated by a broadly defined – and harmonised – right of reproduction. Whereas the latter does not explicitly include a right of adaptation, it must be noted that at the Member State level some Copyright Acts systematically classify the right of adaptation as a form of reproduction (43).

III.3.b.3. The EUCJ and the right of adaptation

As seen above, starting in 2009 the first of these two justifications started to fade away and currently it can be affirmed that the concept of originality is completely harmonised at the EU level.

The same, however, does not hold true for moral rights. As the EUCJ has recently confirmed, moral rights are a matter of Member State law (as clearly stated in recital 19 of the Information Society Directive) and must be exercised in compliance with the provisions of the Berne Convention (44).

Nevertheless, by stressing the importance of the function of parody under EU law and in particular its relationship with the right of freedom of expression, the Court seemed to imply that there are EU limits to moral rights protection (45). Whether these two recent events point to the fact that the right of adaptation will soon be, or perhaps has already been at least in part, harmonised is not clear. The

---

40 The honour and reputation requirement is set in the Berne Convention. Some MS state however do not explicitly include the harm to the honour and reputation as a requirement; see Salokannel M., Strowel A., Final report: Study contract concerning moral rights in the context of the exploitation of works through digital technology, Luxemburg, 2000, at 16.

41 See Art. 6-bis Berne Convention.

42 See Salokannel & Strowel, 2000.

43 See van Eechoud, Hugenholtz, van Gompel, Guibault, Helberger, 2012, at 84.

44 See Opinion of the AG Pedro Cruz Villalón of 22 May 2014 in Case C-201/13 (Deckmyn v Vandersteen), at 4 and 28.

45 In the following decision (Case C-201/13) the CJEU offered its interpretation of what constitutes a parody under EU law stressing the strict relation it possesses with the fundamental right of freedom of expression and identifying the multiple implications of the use of modified version of a work for transformative purposes such as parody. The Court, or more correctly the AG, also clarified that the decision could not address the issue of moral rights. At the national level, see for example the decision of the Brussels Court of Appeal of 8 June 1978 (Tin-Tin and Suisse, rejecting the parody defence), cited in Salokannel & Strowel, 2000; Court of Appeal of Amsterdam of 13 September 2011 BS 7825 in favour of parody for the use of Nijntje).

best view, in absence of an explicit legislative intervention, is that the right of adaptation has to date not been the object of EU harmonisation (46).

The situation remains partially unclear, nevertheless. This was in part confirmed by the European Commission in an unofficial draft document, which should obviously be treated as such, where it was purported the idea that the adaptation right might in fact have been the object of EU harmonisation (47). Discussing the issue of user-generated content, the Commission recognised that "contrary to the reproduction right and communication to the public/making available right, there is no express rule with respect to adaptations in the Information Society Directive". The document continues and opines that:

"the broad manner in which the reproduction right in article 2 is formulated and the [EUCJ]'s jurisprudence on the scope of the reproduction right notably in Infopaq and Eva-Maria Painer seem to cover adaptations which give rise to a further reproduction within the meaning of article 2" (48).

Opportunely, the Commission leaves the door open to possible different interpretations, making explicit reference to a case – pending at the time of the document – decided by the EUCJ. All the same, it should be also noted that the Commission intervened in that case and supported the broad interpretation of the right of reproduction.

III.3.b.4. Allposter v Stichting Pictoright

The case in question is Allposter v Stichting Pictoright, a reference from the Dutch Supreme Court (Hoge Raad) which examined the transfer of images of paintings from posters to canvas. The Hoge Raad referred four questions to the EUCJ. It asked: 1) to define the scope of article 4 Information Society Directive (i.e. distribution right) and whether it includes a right to distribute modified copies; 2a) whether the fact that the redistribution happens in a modified form has any consequence on article 4(2) regulating the exhaustion of the right of distribution; 2b) which type of modifications can avoid the exhaustion of the distribution right; and finally 2c) whether a national rule, such as the Dutch's, which excludes exhaustion when the reseller modifies the work and then distributes it (Poortvleet doctrine), is allowed under EU law (49). Or, as the Advocate General Cruz Villalón summarised in their opinion "[c]an the right holder of a pictorial work, who authorised the sale of posters based on that work, prevent the commercialisation of the same images transferred on canvas" (50)?. The EUCJ borrowed the Advocate General's reformulation of the referred questions and re-proposed it in a slightly more articulated form (51). As the Advocate General clarifies in paragraphs 51-53, the question asked by the referring Court is limited to the right of distribution and does not consider the right of reproduction. Consequently – the Advocate General continues – even though the question could be relevant in terms of the right of reproduction (and its relationship with the right of adaptation) their opinion will disregard that right and only focus on the right of distribution. This is both comprehensible and disappointing since one of the most interesting questions related to the case, and which was elaborated to some extent by the intervening parties (the French Government, the British Government and the European Commission), was going to be avoided.

The Court, however, perhaps realising the logical hardship to solve the referred questions without addressing, at least in part, the issue of the adaptation right and its relation to the right of reproduction,

---

46 Only the AG opinion expressly dealt with the issue of moral rights, the Court did not mention them in its decision.

47 The document is the Commission's draft Impact Assessment on the modernisation of the EU acquis made originally available on the Statewatch website. The specialised IP blog "The IPKat" gives a detailed account here http://ipkitten.blogspot.co.uk/2014/04/breaking-news-draft-impact-assessment.html.

48 See draft Impact Assessment, at 99.

49 See Case C-419/13, of 22 January 2015 Art & Allposters International BV v Stichting Pictoright, (Allposters).

50 See Opinion of the Advocate General Pedro Cruz Villalón of 11 September 2014 in Case C-419/13 Allposters.

51 "the questions referred, which should be considered together, must be understood to mean that the referring court is asking, in essence, whether the rule of exhaustion of the distribution right set out in Article 4(2) of Directive 2001/29 applies in a situation where a reproduction of a protected work, after having been marketed in the European Union with the copyright holder's consent, has undergone an alteration of its medium, such as the transfer of that reproduction from a paper poster onto a canvas, and is placed on the market again in its new form", See Case C-419/13 (Allposters), at 23.

offered some interesting insight. Accordingly, the Court confirmed that article 12 of the Berne Convention confers on authors of literary and artistic works an exclusive right of authorising adaptations, arrangements and other alterations of their works and that there is no equivalent right in the Information Society Directive (52). Nevertheless, the Court held that – without having to interpret the concept of "adaptation" within the meaning of Berne, an exercise carried out by the Advocate General in their Opinion – it is suffice to state that both the paper poster and the canvas transfer contain the image of a protected artistic work and thus fall within the scope of the right of reproduction (`article 4 Information Society Directive) (53). That being said, it is noteworthy for present purposes to briefly look at the qualification of the right of adaptation present in the words of the Advocate general in the remainder of their Opinion:

"one of the essential elements of 'adaptation' as a process of adjustment of the subject-matter of an artistic creation to the methods of expression peculiar to different types of art lies in the diversity of languages and artistic techniques. Another of its essential elements concerns adaptation as a technique of creative expression which seeks to intervene in the work itself rather than to adjust the work to the expressive characteristics of another artistic language, making the work, in its own language, a different work in so far as it is only vaguely recognisable in its original expression" (54).

On the basis of this formulation of the adaptation right the Advocate General concludes that the facts at issue in the main proceedings do not constitute an adaptation since there is no different artistic language, nor are there are additions or modifications to the creative expression. In the present case – the Advocate General states – the objective of the elaboration is to reach the highest possible grade of identity with the original (55).

The Court, avoiding entering into an analysis of the right of adaptation, substantially followed the Advocate General's Opinion on this point. In particular, with regard to the transfer of the image from poster to canvas, the Court noted that such a replacement of the medium results in the creation of a new object incorporating the image of the protected work, whereas the poster itself ceases to exist (56). Such an alteration of the copy of the protected work, which provides a result closer to the original, is actually sufficient to constitute a new reproduction of that work, within the meaning of article 2(a) Information Society Directive (57). Crucially, the Court rejected Allposters's argument that no act of reproduction is performed, since there is no multiplication of copies of the protected work (the transfer of the ink from the poster to the canvas not only reproduces the image on the new medium, but erases it from the old one) (58):

"The fact that the ink is saved during the transfer cannot affect the finding that the image's medium has been altered. What is important is whether the altered object itself, taken as a whole, is, physically, the object that was placed onto the market with the consent of the right-holder. That does not appear to be the case in the dispute in the main proceedings" (59).

Consequently, the Court concludes that the consent of the copyright holder does not cover the distribution of an object incorporating his work if that object has been altered after its initial marketing in such a way that it constitutes a new reproduction of that work (60).

III.3.b.5 Some final considerations on the right of adaptation

The Court clearly established that, differently from the right of reproduction and the right of distribution, the right of adaptation is not present in the European *aquis* (61). The Court, however, also stated

52 ID., at 26.

53 ID., at 27 – 28.

54 See AG Opinion in Case C-419/13, at 58.

55 ID., at 59.

56 The Court notes that such a technique "increases the durability of the reproduction, improves the quality of the image in comparison with the poster and provides a result closer to the original of the work" an expression that seems to imply that the modified work is directed to a new public; See Case C-419/13, at 42.

57 See Case C-419/13, at 43.

58 ID., at 44 and 45.

59 ID.

60 ID., at 46.

61 ID., at 26.

that certain kinds of alterations, such as the one at stake in the main proceedings, are covered by the harmonised right of reproduction (62). The question that the Court did not address is how to distinguish non-harmonised adaptations from those other adaptations included in the harmonised right of reproduction.

In accordance to previous EUCJ case law, a reproduction, in order to be covered by article 2 of the Information Society Directive (i.e. reproduction by any means and in any form, in whole or in part) needs to reproduce the author's own intellectual creation. The EUCJ clearly stated that even a small or short reproduction – such as eleven words – can constitute an infringement of the reproduction right provided that the reproduced excerpt constitutes the author's own intellectual creation (63). The facts under analysis in Allposter related to a perfect and complete reproduction of the original image onto the new medium, and therefore the reproduction certainly included the author's own intellectual creation present in the original. Consequently the question is: does the same infringement test apply also to the right of adaptation? Admittedly, a mere extension of the Infopaq infringement test to the right of adaptation will lead to the assimilation of the latter into the already broadly defined right of reproduction. Nonetheless, a similar course of action would deny the idiosyncrasy of a right to create secondary works. This specialty resides in the fundamentally different function carried out by the right to create adaptations. This function is to delimit the boundaries between infringing activities and permitted uses, and basically defines the scope of copyright in modern legal systems (64). In particular, in knowledge-based and digitally enhanced societies this function is fundamentally connected to the role that transformative uses possess. More and more often, the creation of new forms of expression employing existing works – very often to an extent that certainly satisfies the Infopaq test – constitutes a central element of new cultural and economic practices, particularly in on-line environments. This can be seen, for instance, in phenomena such as user-generated content, as well as in the one covered in this analysis (ML and other forms of textual analysis such as text mining). The more transformative a work is, the less likely a finding of infringement should be, irrespective of how much of the original author's own intellectual creation has been taken. Conversely, a rigid test that only looks at the amount taken, therefore reducing the right of adaptation to a particular form of reproduction, would deny the economic and cultural value that the concept of transformative uses embraces. Likewise, such a rigid test would also stifle innovation. In other words, the question that should be asked in order to determine whether highly transformative uses should be authorised or not by right holders is whether and to what extent right holders should be able to control the development of technological innovation. The answer seems to belong much more to the Parliaments than to the Courts.

In conclusion, a flexible test that compares the amount reproduced with the added creative elements and evaluates how distant the final transformative use is from the original work would certainly strike a fair balance between creativity, innovation and freedom of expression. Furthermore, it must be reiterated that, if it is accepted that the concept of adaptation is ontologically different from that of reproduction as it is here suggested, then a broad exception for the case of transformative works, that is to say an exception to the right of adaptation, would not be limited by the closed list of article 5 Information Society Directive (65).

III.3.c Annotations, models and original adaptations

62 ID., at 43.

63 See Infopaq, at 39.

64 In this sense Gervais D., The Derivative Right: Or Why Copyright Law Protects Foxes Better than Hedgehogs, 15:4 Vanderbilt Journal of Entertainment and Technology Law 785-855, 2013.

65 See van Eechoud & all, 2012, 84; Hugenholtz B., Senftleben M., Fair Use in Europe. In Search of Flexibilities, IViR research papers, Amsterdam 2011. In the same sense a Gervais, 2013. See also the Report prepared by the Irish Copyright Review Committee, Modernising Copyright, Dublin 2013, at 55 and 72 (available at http://www.enterprise.gov.ie/ en/News/2013/October/-Copyright-report-published-aimed-at-supporting-digital-industry-%E2%80%93-Minister-Bruton.html).

In the light of the above, examples such as those of user-generated remix of videos available on popular video sharing platforms often constitute creative adaptation of the original videos (sometimes exempted from authorisation as a parody or criticisms). This is on the basis of the original contribution (e.g. the creative selection, the artistic message) that the author of the remix is adding to what would otherwise be a mere (partial) reproduction of others' works. However, it is doubtful whether an annotated text can be considered a derivative work of the original unannotated corpora or whether the model trained from an annotated corpora constitutes an original adaption of the latter.

The act of annotation is certainly time consuming and can be a quite complex activity. Nevertheless it is an activity that, if properly and diligently carried out following the provided instructions, will look very similar if not identical regardless of whether it was executed by highly trained annotator A, or an equally skilled annotator B. In fact, annotations will often be limited to the correct categorisation of certain words or phrase segment (e.g. verbs, nouns, conjunctions) or expressions (rules, conditions, statements, etc). This type of systematic classificatory activity that has to follow rigid and predetermined instructions, i.e. the "rules of the game", does not commonly allow the free and creative choices through which the personality of the author can be expressed. That said, in highly complex annotations (e.g. the annotation of poetry where the annotator has to identify which sentences express a certain sentiment or moral condition) it cannot be excluded *a priori* that a certain, even small, space is present for choices and personality.

Similarly, the process of training a model from an annotated text is usually an automated process where the ML researcher has no general control or possibility to perform free and creative choices in order to input their personality. There are a – usually limited – number of settings and other parameters that can be adjusted in the ML software but these are normally quite standardised. Of course, also in this case it cannot be excluded that in certain highly advanced sectors, where the ML/NLP workflow is composed not by a single process, but where models are trained multiple times following specifically selected corpora at different stages of the analysis, once again there may be space for choices. Whether these are free and creative enough is a matter that can only be ascertained on a case by case basis, and while originality cannot be excluded *a priori*, it does not seem like an easy result to achieve. As a matter of fact, it seems that – absent clear case law in this area – the relevant industry sectors appear to rely on other forms of protection, such as trade secrets, in order to protect what appears to be much closer to know-how than to authorial contributions.

IV. Training models: temporary and permanent reproductions

In the EU legal order, the right of reproduction is defined as "any direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part" and is reserved to the right holder of copyright works and other protected subject-matter by article 2 Information Society Directive (66). The right of reproduction can be seen as the cornerstone of copyright law protection, both historically and functionally. Accordingly, it can be observed that the definition of the right of reproduction is broad enough as to ensure legal certainty within the internal market (67) and to extend to every act of reproduction, however transient or irrelevant it may be from an economic perspective (68).

It is important to note here that not only permanent but also temporary reproductions are protected, such as those made in the random-access memory (RAM) of computers and other digital devices when browsing the internet and visualising websites (69). Thus, it is only thanks to the mandatory exception of article 5(1) that acts such as internet browsing are allowed under EU law (the relevance of doctrines such as implied licence and estoppel are not considered here). Article 5(1) establishes that temporary acts of reproduction, which are transient or incidental and an integral and essential

66 Directive 2001/29/EC, OJ L 167, 22.6.2001

67 See Recital 21 InfoSoc Directive.

68 See von Lewinski S., Walter M., European Copyright Law – A Commentary, OUP, 2013, p. 968.

69 See Recital 33 Information Society Directive.

part of a technological process and whose sole purpose is to enable a transmission in a network between third parties by an intermediary, or a lawful use of a work or other subject-matter, and which have no independent economic significance, are exempted from the reproduction right provided for in article 2.

Interestingly, the EUCJ clarified that certain acts of temporary reproduction carried out during a "data capture" process fulfil the requirements of the exception for temporary copies and offered some interpretative guidance on the conditions listed in article 5(1) and recital 33. This is relevant, as the "data capture" process described by the Court is not too different from the steps carried out in modern ML processes and identified above. However, before proceeding to a comparison between the "data capture" and the ML/NLP steps, it is useful to to spend some words on how the Court qualified and interpreted further certain conditions of article 5(1).

As indicated above, one of the requirements of article 5(1) is that the acts of temporary reproduction must constitute an integral and essential part of a technological process. The Court clarified that, whereas temporary acts of reproduction need to be carried out entirely in the context of the implementation of the technological process and 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 (70). Secondly, acts of temporary reproduction must pursue a sole purpose, namely to enable the lawful use of a protected work, i.e. a use that is permitted by the right holder or that it is not restricted by law. Thirdly, temporary reproductions must not have an independent economic significance, a condition which is achieved when those acts do not enable the generation of an additional profit going beyond that derived from the lawful use of the protected work and do not lead to a modification of that work (71).

IV.1. *Infopaq I & II* and data capture processes

A brief description of the facts of the Infopaq cases may be helpful. The decisions (technically a judgement and an order of the Court) refer to the compilation, extraction, indexing and printing of newspaper articles and keywords operated by a media monitoring business (Infopaq) and the request of an association representing Danish publishers to obtain a licence for this type of activity. The Court starts its analysis by identifying five phases in the process of data capture: (1) newspaper publications are registered manually in an electronic registration 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 their 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 the computer and all data are saved as a text file, whereas the TIFF file is then deleted; (4) the text file is processed to find a search word defined beforehand, identifying possible matches and capturing five words before and after the search word (i.e. a snippet of eleven 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. In the figure below (Fig. 1) an example of the outcome of the Infopaq service can be seen, as reported by the Court in its order of 2012 at paragraph 16.

'4 November 2005-Dagbladet Arbejderen, page 3:
TDC: 73 % "forthcoming sale of the telecommunications group TDC, which is expected to be bought".'

Fig. 1 - An example of the snippet provided by the Infopaq service

The Court found that the exception of article 5(1) only covers the activities listed in points 1) to 4) above, whereas the activity of point 5), i.e. printing, constitutes a permanent act of reproduction

70 Infopaq II, 29 and ss.

71 See Case C-5/08 Infopaq I and C-302/10 Infopaq II.

which is therefore not covered by the exception for temporary copies. It should further be noted that, in point 5), what is printed is not the entire literary text, but only eleven consecutive words. Only if these 11 consecutive words constitute a "reproduction in part" of the original work, copyright would be infringed. In this regard, the EUCJ found that "it cannot be excluded" that eleven consecutive words constitute the author's own intellectual creation and therefore represent a partial (and thus infringing) permanent reproduction (72). The eleven words threshold should obviously not be taken as a strict parameter. As explained above, the real test is that of the author's own intellectual creation, introduced in Infopaq I and further developed in the Court jurisprudence (73). Accordingly, there could be shorter extracts that meet such a condition (although as a rule of thumb it could be said that the shorter the extract the harder for it to be original), and longer extracts that do not meet it (for example, there are many standard expressions in a given field or sentences that are mere descriptions of facts which hardly will be considered original). How to assess on a case by case basis whether the extracted sentence fragment meets, or does not meet, such a threshold can be quite problematic. However, this problem can be avoided, as long as the reproductions are temporary and meet the conditions described in article 5(1).

IV.2. Data capture processes and ML model training

Regarding the first step, the registration of newspaper publications in an electronic database seems substantially equivalent to the corpus compilation phase where corpora are selected on the basis of their relevance and obtained from a variety of sources. Regarding the second step, the creation of a TIFF file from the newspapers and the transferring to an OCR server appears substantially identical to ML/NLP preprocessing where textual content is converted into a format that can be further processed by the NLP tools (e.g. PDF to plain text). The third step, the translation of the image of each letter into a character code recognisable by the computer, appears again equivalent to the ML/NLP annotation phase where words and sentences are enriched in order to make them understandable by the ML algorithm. The fourth step, i.e. finding a search word defined beforehand, identifying possible matches and capturing five words before and after the search word, appears again very similar to the actual training performed by the ML algorithm, where the latter analyses and extracts grammatical, syntactic, probabilistic and other similar features from the text. Finally, the permanent record of the key word including the five preceding and five successive words, although not relevant for the purpose of article 5(1) analysis, seems substantially similar to the model creation (the file) that contains the extracted probabilistic features. The only difference in this instance is that, depending on what is contained in the trained model, the latter may not constitute an infringing partial copy of the former.

It seems furthermore plausible that the temporary copies created during the ML process are transient or incidental if they are only kept for the amount of time justified by the proper completion of the technological process and are automatically destroyed at the and of the process. It is also arguable that the act of reproduction is an integral and essential part of a technological process (the conversion of the text into data) which is necessary to enable a lawful use. For example, ML statistical analysis is arguably as lawful as the preparation of summaries and is not a right reserved to the right holder by EU copyright law, however if the right holder contractually limits this operation and domestic law allows it, this condition could not be met (see infra). The requirement of absence of independent economic significance is probably harder to assess. Independent economic significance is present if the author of the reproduction is likely to make a profit out of the economic exploitation of the temporary copy. This profit has to be distinct from the efficiency gains that the technological process allows (see Infopaq II, 51). In the present case, it seems that even if an economic gain may be derived, it will be connected to the lawful use (the resulting ML analysis) and not directly to the temporary copy.

72 See Infopaq I.

73 See supra Sec. III.1.b

Accordingly, it seems plausible, if not probable, that the temporary copies necessary to perform model training activities as those described above, are covered by the exception of article 5(1), as long as the conditions described here are met.

Moving to the final step, the storage of the results (the eleven consecutive words or the storage of certain probabilistic rules into a digital database, i.e. the trained model) is clearly not covered by article 5(1) as this act is permanent by definition. Therefore, the last point cannot be exempted on the basis of acts of temporary reproduction. It must be assessed however, whether the model constitutes a "reproduction in part" within the meaning of article 2 Information Society Directive. If it does not, there is simply no copyright relevant activity and thus no need to rely on an exception. In certain surveyed ML/NLP scenarios, the trained model contains three consecutive words of the original corpora. While the test to be applied is not eleven vs. three consecutive words, but that of the "author's own intellectual creation", it seems plausible if not highly possible that three consecutive words are too insubstantial to constitute a "reproduction in part" of the original corpora. Whether the repeated extraction of three consecutive words to the point where the entire original text is being reproduced, just in a different "shuffled" format is tantamount to a reproduction in part (or perhaps to an adaptation) is another aspect that should be ascertained.

Therefore, it can be concluded that, at least in the ML/NLP workflows used to inform this study, article 5(1) has the potential to cover the temporary copies made during a technological process similar to the one described in this scenario. Additionally, it can further be concluded that the resulting trained (and permanent) model of the above described technological process is not a copy or adaption of the original corpora. However, given the cumulative, strict and partially unclear conditions that qualify article 5(1), a very careful case-by-case assessment should be performed before deciding to rely on this exception given the unavoidable degree of risk involved. Regarding the "fifth step", in cases where the trained model reproduces the entirety of the original textual elements, it should be verified whether the repeated reproduction of insubstantial textual strings can be considered an article 2 reproduction in part or perhaps even an adaptation.

V. Final remarks and the issue of "owning" AI

As a result of the above, it can be argued that EU copyright law in the field of ML and model training permits the temporary copy of copyright works for purposes such as "data capturing" as long as the cumulative conditions of article 5(1) as interpreted by the Court are met (74). Additionally, it seems that the steps described in the "data capturing" process can be considered equivalent to the steps involved in modern ML, particularly in the field of NLP. A few remarks however are necessary. First of all, different algorithms behave differently, therefore it should be ascertained on a case by case basis whether article 5(1) conditions are met. Second, the condition of lawful use should be properly qualified. Recital 33 and article 5(1) indicate that a lawful use is one that is either authorised by the right holder or by the law (e.g. under another exception), and the EUCJ clarified that this includes uses that do not require authorisation (e.g. the creation of summaries) (75). Whether or not this can be "contracted-out", i.e. whether a right holder can contractually condition the lawful access to a resource to the prohibition to proceed to ML or other analytic processes, is not fully clear. On the one hand, a literal reading of the condition seems not to contemplate this eventuality, as the authorisation of the right holder is only needed if the use would otherwise be restricted by law. Nevertheless, the issue of whether exceptions and limitations can be effectively limited by contract under EU law is a relevant and open one, so much so that current national exceptions and EU proposals in the field of Text and Data Mining (TDM) clarify that said exceptions cannot be limited by contract. This is a very

74 Confirming the cumulative nature of Art. 5(1), Infopaq II, 26.

75 Infopaq I, 23.

important aspect in the field of data analytics as often access to corpora is based on contractual agreements with publishers and other resource platforms which may include contractual limitations regarding certain uses. Finally, the absence of independent economic significance is also difficult to assess. On the one hand, gains in efficiency cannot be counted as independent economic elements, nevertheless ML and other forms of Text Mining (TM) often lead to important commercial benefits that certainly go beyond what originally contemplated by the right holder who made available the corpora. It should be noted however that this commercial benefits do not stem directly from the temporary copy, but from the results of the ML process, the temporary copy being an incidental element in such process.

Finally, it must be stressed that the conditions of article 5(1) are not only cumulative (i.e. all must be met) but must be interpreted strictly (76). These considerations have led many commentators to the conclusion that article 5(1) is not suitable as a general solution for TDM purposes (77). Whereas article 5(1) is certainly not the clear and open standard through which EU law could show its innovation-oriented approach, in absence of any other broader exception article 5(1) appears to be the main tool that can currently be employed at the EU level to balance the needs of innovation and the protection of investments in the digital single market.

However, article 5(1) only exempts acts of temporary reproduction. This means that when the final trained model reproduces an article 2 Information Society Directive part, or when the trained model can be seen as a protected adaptation of the original corpora, a specific authorisation is needed. This authorisation can be found in the law, although article 5 Information Society Directive exceptions do not seem to offer a systematic solution to this problem. The reasons for this conclusion have been largely explored by the literature, so they don't need to be recalled here. However, it may be useful to stress that the main problem with the article 5 list of exceptions is that they are not mandatory, meaning that Member States can choose from the closed list of article 5 which exceptions to implement into national law. This solution is clearly unsuitable for the Digital Single Market. Furthermore, it should be noted that the national Copyright Acts that have recently created an exception for TDM as well as the recently proposed EU exception, only apply to the right of reproduction. This means that, if and when the trained model qualifies as an article 2 reproduction, such a model cannot be redistributed or communicated to the public. Similar problems could arise should the trained model qualify as an adaptation.

Another aspect that should be considered is that the role played (or rather, not played) by EU copyright exceptions could be much more effectively played by contractual agreements, that is to say by those licences that permit the exercise of the rights needed for ML and other text analytic processes. Open Access licences have proven to be suitable to this end. In other words, to the extent to which ML algorithms are trained on corpora licensed under Open Access licences (or public domain material), the complexity so far described vanishes. This is particularly relevant in the EU given that many funding organisations require that that results of the funded research be licensed under Open Access compliant licences.

A closing final consideration relates to the question of whom should be allocated the power to determine the future of technological development, or "who owns AI"? In fact, to determine whether training models for ML purposes requires the permission of the copyright holder of the underlying corpora corresponds, from an innovation policy point of view, to determine who "owns" AI, or at least who controls its development, under which conditions and at what prices. In other words, whether ML developers (usually large IT companies but more and more often also innovative start-ups) need the permission of right holders to use certain content in order to train their algorithms is not only a copyright matter, but is an issue that implies broader innovation policy considerations as it will heavily influence the future speed (and cost) of AI development (78). From a copyright point of view, the different rationales or justifications to copyright may lead to different answers (at first superficial sight, a labour law theory could offer more arguments in favour of right holders than what a strict utilitarian interpretation might). This could be an interesting area to conduct further research. But it should be stressed again that the answer to this question exceeds copyright law and theory and

---

76 Confirming the cumulative nature of Art. 5(1), Infopaq II, 27.

77 See Triaille J.-P., de Meeûs d'Argenteuil, J. and de Francquen A., Study on the legal framework of text and data mining (tdm), 2014, Luxembourg Publications Office.

78 At the European level, in 2018 the EU has created a High-Level Expert Group on Artificial Intelligence whose task is to "support the implementation of the EU Communication on Artificial Intelligence published in April 2018". In the Communication, the Commission identifies a number of priority areas that go from governance to ethical aspects of AI, including some references to copyright law; see https://ec.europa.eu/digital-single-market/en/ news/commission-appoints-expert-group-ai-and-launches-european-ai-alliance.

becomes part of a broader analysis of innovation policy. Certain jurisdictions appear to possess, within their copyright laws, certain tools (or open standards, the reference here is to fair use and to broad interpretations of fair dealing) that allow to strike the balance in a more innovation oriented way. After all, U.S. courts have already repeatedly found that, in the field of text and data analytics, to mine books or the web is a transformative use that does not require authorisation. At the moment, it does not appear that this line of judgements have affected the business models of the relevant right holders, even though more data would be needed to support this conclusion. Certainly, what has been shown is that while these more "innovation-oriented" jurisdictions are advancing fast in the field of ML and AI, the EU with its cumulative and narrow interpretation of article 5(1) is "falling behind" (79).

79 Handke C., Guibault L., Vallbé J.-J., Is Europe Falling Behind? Copyright's Impact on Data Mining in Academic Research, in Schmidt B.. et al., (Eds), New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust - Proceedings of Elpub 2015, pages 120–130.

RCUK Centre for Copyright and
New Business Models in the
Creative Economy

College of Social Sciences / School of Law
University of Glasgow
10 The Square
Glasgow G12 8QQ
Web: www.create.ac.uk