Copyright lessons on Machine Learning: what impact on algorithmic art?

by **Theodoros Chiou**[*]

Abstract: Nowadays, Artificial Intelligence (AI) is described as "the new electricity". Current algorithmic innovation allowed the development of software which enables machines to learn and to achieve autonomous decision making, with limited or no human involvement, in a vast number of applications, such as speech recognition, machine translation and algorithmic creation of works (computer generated art), on the basis of a process widely known as Machine Learning (ML). Within the ML context, machines are repeatedly trained by means of specifically designed learning algorithms that use a corpus of examples in the form of data sets as training material. Very often and, especially in the context of algorithmic creativity, the training material is mainly composed by copyrighted works, such as texts, images, paintings, musical compositions, and others.

Machine Learning workflow typically involves the realization of (multiple) reproductions of any protected work used as training material. The present paper aims to assess the extent to which the use of copyrighted works for Machine Learning purposes in the field of algorithmic creativity is controlled by the monopolistic power of the copyright rightholder on that work. The answer to this question will be researched in the context of EU copyright law, by examining the content of reproduction right and exceptions possibly applicable in a typical ML workflow in the field of algorithmic art, before making an overall assessment of the current EU regulatory framework for artistic ML projects, as it is shaped after the DSM Directive 2019/790.

Keywords: artificial intelligence; machine learning; text and data mining; algorithmic art; copyright; copyrighted works; Infosoc Directive; DSM Directive; Big Data; reproduction right; adaptation right; copyright exceptions

Recommended citation: Theodoros Chiou, Copyright lessons on Machine Learning: what impact on algorithmic art? 10 (2019) JIPITEC 398 para 1

A. Introduction

1 **The objective of Making machines intelligent.** Artificial Intelligence (AI) may be seen from different standpoints and receive accordingly different interpretations. From a rather technical point of view[1], Artificial intelligence is the field of

[*] Dr. Theodoros Chiou is Post-Doc Researcher at the University of Athens, School of Law (Department of Private Law) and Attorney-at-law (IPrights.GR). Email: Theodoros.chiou@iprights.gr. This paper is based on a conference presentation delivered by the author during the 9th ICIL Conference, "Psychological and socio-political dynamics within the Web: new and old challenges to Information Law and Ethics", held in Rome, Italy, July 11-13, 2019.

[1] For a different approach, see among others Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach* (3rd ed., Pearson 2010) 1: "the study of agents that exist in an environment and perceive and act".

computer science[2] which focuses on the production of intelligent computational systems, i.e. machines that run software(computers), with or without hardware extension (such as robots), that *mimic* human intelligence and are capable of deploying human cognitive functions, such as problem solving, decision making, object recognition, learning and *creation of works*[3], among others. Nowadays, Artificial Intelligence (AI) is described as "the new electricity", as AI systems that emulate intelligent behavior in terms of computational processes, are (or are about to be) put into daily service of human activity. As of today, AI applications[4] range from autonomous cars to automated language translation, prediction, speech recognition, computer vision, and production of artistic creations; the latter is main subject of the present paper.

2 **A technique to make machines intelligent: Machine learning.** Machine learning (ML) is a subfield of AI that blends mathematics, statistics and computer science[5]. In a nutshell, ML is a self-learning computational process that constitutes a fundamental apparatus for the development AI systems, because it enables machines make 'autonomous' intelligent decisions. The basic idea behind ML is to allow machines learn from thousands of examples of a given phenomenon and build 'mental' models out of these examples that will be used by the machine in order to produce output when confronted with new input. More precisely, ML relies on the creation and implementation of *training or learning algorithms* that "program" machines to learn through the processing and analysis of structured *corpora of (big) training data sets* (so-called *training data*). In addition, these algorithms permit learning from experience and future data input[6], since, via their training, they improve in performance over time[7], *without being specifically programmed*[8]. Obviously, as a technique of automated data analysis, ML implies the deployment of Text and Data Mining methods —TDM[9]. The abundance of available training data (online or elsewhere) in today's big data-driven era[10]along with the available computational power and the algorithmic innovation in the ML field explain, among others, the current rise of AI[11].

3 **(Digital) Works as (Big) training data: Works as data.** In the field of AI-driven creativity or algorithmic creativity, ML algorithms allow machines to "learn" how to autonomously produce *novel creative and artistic output* known as *algorithmic art*[12], such as translated texts, musical compositions,

[2] For some authors, AI is a science by itself. See among others, Aikaterini Georgouli, *Artificial Intelligence, An introductory approach* (Hellenic Academic Electronic Textbooks 2015), available at: <www.kallipos.gr>, accessed 3 December 2019, p. 13.

[3] For the connection between intelligence and creativity see among others Daniel Schönberger, 'Deep Copyright: Up - And Downstream Questions Related to Artificial Intelligence (AI) and Machine Learning (ML)' (2018) SSRN <https://ssrn.com/abstract=3098315> accessed 3 December 2019, pp. 3-4 and references mentioned therein.

[4] For a broader discussion on AI applications see among others Harry Surden, "Artificial Intelligence and Law: An Overview" (2019) SSRN: <https://ssrn.com/abstract=3411869> accessed 3 December 2019, p. 88.

[5] Amanda Levendowski, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem' (2018) Wash. L. Rev. 579, 590.

[6] Some argue that ML will cause "the end of code". See Jason Tanz, 'Soon We Won't Program Computers. We'll Train Them Like Dog' (*Wired.com,* 17/5/2016) <https://www.wired.com/2016/05/the-end-of-code/> accessed 3 December 2019. For a critical approach, see Andrew Vogan, 'Let's Explore Wired's Article about 'The End of Code', (*Art+Logic,* 17/5/2016) https://artandlogic.com/2016/05/software-developers-response-wireds-end-coding-article/ accessed 3 December 2019.

[7] Surden (n 4) p. 88.

[8] In fact, researchers acknowledged that it is easier to program a computer to learn to be intelligent rather than programming a computer to be intelligent, see Schönberger (n 3) p. 11.

[9] See below, para. 18.

[10] See Eleonora Rosati, 'Copyright as an Obstacle or Enabler? A European Perspective on Text and Data Mining and its Role in the Development of AI Creativity' (2019) SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3452376> accessed 3 December 2019 p. 1 ff. and references cited therein.

[11] On that topic, see among others Christophe Geiger & Giancarlo Frosio & Oleksandr Bulayenko, 'Crafting a Text and Data Mining Exception for Machine Learning and Big Data in the Digital Single Market' in Xavier Seuba & Christophe Geiger & Julien Pénin (eds.), *INTELLECTUAL PROPERTY AND DIGITAL TRADE IN THE AGE OF ARTIFICIAL INTELLIGENCE AND BIG DATA,* (2018) CEIPI/ICTSD publication series on "Global Perspectives and Challenges for the Intellectual Property System", Issue No. 5, Geneva/ Strasbourg, pp. 97-111 and, in particular, p. 97 and 109 and references cited.

[12] See: <https://en.wikipedia.org/wiki/Algorithmic_art> accessed 3 December 2019. This kind of art production is known as computer art or generative art. For the latter see <https://en.wikipedia.org/wiki/Generative_art> accessed 3 December 2019.

paintings[13], or even poems[14] and novels[15]. In these cases, AI systems are trained on data sets that consist of the type of works relevant to each project, that are (at least at the moment[16]) created by humans, such as texts, photographs, musical compositions and the like. These *"training works"* correspond to the data set used as training material. However, it is very likely[17] that many of these *training works are protected by copyright law*[18]. For example, for the "creation" of the "SKYGGE" pop album "Hello World"[19], the first pop album composed by AI, several copyrighted musical works have been used as training data ("inspirations") for the AI to generate novel output: "Ballads, Pop of the 60s, Brit Pop of the 2010s, Bossa novas of the 60s, Caribbean songs, Soul Music from the 80s, Musicals of the 60s, French Pop from the 80s, Purcell"[20], most of which are copyrighted material. Similarly, for the creation of the novel "1 The Road", the machine has been trained "with three different text corpora, each with about 20 million words one with poetry, one with science fiction, and one with "bleak" writing"[21]. Besides, copyrighted human works are used as training data in other AI applications, such as Natural Language Processing (NLP)[22].

4 **Copyright law concerns over Machine Learning workflow.**ML process, in analogy with the TDM methods, raises copyright law issues to the extent that the *use of works for ML purposes requires typically copying and/or adaptation of these works*[23]. Consequently, apart from output interrogations, regarding the proprietary status of the 'intelligent' artistic/creative output produced by the machine[24] (including the question of whether authors' rights over their works also extend to outputs produced by AI, after being trained on these works[25]), another

[13] See for instance the Edmond de Belamy portrait (2018), a painting printed on canvas and created by algorithm. For more information see: https://en.wikipedia.org/wiki/Edmond_de_Belamy accessed 3 December 2019. The painting in question was the first artwork created using Artificial Intelligence to be featured in a Christie's auction. See https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx accessed 3 December 2019.

[14] See the interesting website http://botpoet.com/ accessed 3 December 2019, which implements a Turing test for poetry and the user is called to guess whether the poem is written by a human or by a computer.

[15] See for instance the novel "1 The Road" (Jean Boîte Editions 2018), with "Writer of writer" Ross Goodwin. More information at: <https://en.wikipedia.org/wiki/1_the_Road> accessed 3 December 2019 and <https://jean-boite.fr/products/1-the-road-by-an-artificial-neural> accessed 3 December 2019.

[16] Things might turn more (or, under certain conditions, less) complicated in case that training works are the output of AI-driven creative process.

[17] Levendowski (n 5) p. 582.

[18] Schönberger (n 3) p. 1; Geiger *et al.*, Crafting (n 11) p. 109: "These artificial intelligence learning processes must use inputs possibly protected by IPRs to create wholly transformative outputs."Of course, there are also training material which either do not qualify for copyright protection (e.g. due to lack of originality or because they are simple facts or pure data) or their protection has ended (e.g. public domain works). In this paper we will not examine further the issue of copyrightability of training works and we will focus on copyright issues arising from the use of copyrighted works as training data in the course of ML workflow.

[19] The "Hello World" album started as a research project, namely the Flow-Machines project, conducted at Sony Computer Science Laboratories and University Paris 6, and funded by the ERC. See https://www.helloworldalbum.net/.

[20] See the album pitch at: <https://www.facebook.com/pg/flowSKYGGE/about/> accessed 3 December 2019. *Adde* the description for song "Daddy's Car", a song composed in the style of Beatles by Sony CSL Research Lab: "The researchers have developed FlowMachines, a system that learns music styles from a huge database of songs". *Cf.* Rosati 2019 (n 10) p. 3: "How could it be possible for AI to create a song in the style of The Beatles if it did not also have access to The Beatles repertoire?".

[21] See <https://en.wikipedia.org/wiki/1_the_Road>.

[22] For instance, researchers had used 11,038 novels for training a neural network to model a system that can create natural language sentences, see Schönberger (n 3) p. 12.

[23] Schönberger (n 3) p. 13: "ML hence often faces a fundamental problem since it may have as a condition precedent that one or even several copies are made of any work used as training data"; Rosati 2019 (n 10) p. 3: "[C]opyright law poses potential restrictions to the training of AI for the purpose of creative endeavours, even if the copies made of pre-existing content are only used internally and are instrument to the creation of something else."

[24] This question is outside the scope of this paper. On this topic see, among the abundant literature, Rosati 2019 (n 10) p. 2, footnote 5 and references cited therein.

[25] For this question, see among others Giovanni Sartor & Francesca Lagioia & Giuseppe Contissa, 'The use of copyrighted works by AI systems: Art works in the data mill' (2018) SSRN <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3264742> accessed 3 December 2019.

thorny process issue[26] related with *copyright law concerns* arises: May protected works be used for machine training purposes within ML context without copyright restraints? Or does the use of protected works for ML purposes require prior authorization from rightholders of reproduction rights over the training works?[27]The question is fundamental, if one considers the impact it may have in the development of the whole AI field**,** *including algorithmic art,* which the present paper focuses*.*

5 The question will be investigated in the context of EU copyright law, by assessing the manipulation of training works within ML workflow in terms of reproduction right **(2)** and by examining the applicability of mandatory exceptions thereto **(3)**, before making an overall assessment of the current EU regulatory framework for artistic ML projects, as it is shaped after the 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 (DSM Directive) **(4)**.

B. Assessing ML workflow in terms of the EU reproduction right

I. The reproduction right under EU copyright law: a reminder

6 The EU *acquis* on copyright law establishes a comprehensive exclusive right of reproduction. More precisely, according to art. 2 of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society (hereinafter: "Infosoc Directive"), the right of reproduction is defined as the "exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part."[28] This article introduces a broad definition of acts covered by the reproduction right[29]. In addition to that, the ECJ case law has adopted a broad interpretation of the concept of reproduction[30]. This means that in the digital environment, to which the AI sphere belongs, any digital copy of a work, temporary or permanent, direct or indirect, has the potential to infringe copyright, irrespective of how transient, short or irrelevant from an economic perspective it may be[31], provided that it reproduces the creative expression of the initial work, even in part[32].

7 Besides, the adaptation right, i.e. the right to create (original) derivative works from existing ones, has mainly remained untouched by the Infosoc Directive[33] and, thus, it basically remains unharmonized at EU level[34]. However, given the broad definition of art. 2 Infosoc, some transformative uses of works may be, in fact, also qualified as reproductions[35] and, thus, be covered by the reproduction right, to the extent that the alterations undertaken give rise to further (mere) reproductions of previous works (without creative additions or modifications) and not creative adaptation. In any event, all copies of works that may be considered as "genuine" adaptations under national law are (or imply) acts of reproduction covered by EU *acquis*[36].

[26] *Cf.* Sartor *et alii* (n 25) p. 8, distinguishing between process issues and outcome issues related with the use of pre-existing works within the ML process ("the data mill").

[27] *Cf.* for a similar research question, Schönberger (n 3) p. 13. The question is relevant equally for both copyright and related rights field. For simplicity reasons, we limit our analysis to copyright law interrogations.

[28] This definition is much more sophisticated than Article 9(1) of the Berne Convention, which also refers to an exclusive reproduction right in any manner or form.

[29] This is justified, according to the European legislator and Court of Justice, by the need to ensure legal certainty within the internal market *Cf.* recital 21 Infosoc Directive; ECJ Case C5/08 *Infopaq International A/S v Danske Dagblades Forening* [16 July 2009] ("Infopaq I"), para. 41.

[30] See Infopaq I, para. 43.

[31] Thomas Margoni, 'Artificial Intelligence, Machine Learning and EU Copyright Law: Who Owns AI?', (2018) *CREATe Working Paper 2018/12* ; SSRN <https://ssrn.com/abstract=3299523> or <http://dx.doi.org/10.2139/ssrn.3299523> accessed 3 December 2019, section IV.

[32] Infopaq I, para. 39.

[33] Margoni (n 31) section III.3.b.

[34] Indeed, according to the decision of the ECJ Case C-419/13 *Allposters International BV v. Stichting Pictoright* [22 January 2015], para. 26, there is no equivalent right of adaptation right in the InfoSoc Directive.

[35] See Silke von Lewinski & Michel Walter, 'Information Society Directive', *in* Michel Walter & Silke von Lewinski (eds.), *European Copyright Law: A Commentary* (Oxford University Press 2010) 967 and 968.

[36] Jérôme de Meeûs d'Argenteuil & Jean-Paul Triaille & Amélie de Francquen, Study on the legal framework of text and data mining (TDM) (2014) <https://op.europa.eu/en/publication-detail/-/publication/074ddf78-01e9-4a1d-9895-65290705e2a5/language-en> accessed 3 December 2019, p. 32.

II. The existence of copyright-significant reproductions within the ML workflow

8 Given the contours of the reproduction right in the EU *acquis* according to art. 2 (1) Infosoc Directive and the meaning of "copy" under EU copyright law, ML workflow[37] usually entails several copyright-significant reproductions[38]. More precisely, (digital) copying of works (multiple, sometimes) may take place in the beginning of the AI project and at the first stage of a ML workflow[39], namely the stage that refers to the identification and collection of appropriate preexisting works from one or various sources, according to their relevance for the AI project, in order to create a *corpus* of training examples for the machine (*corpus compilation stage*)[40]. Indeed, the detection and preselection of works as training examples implies (digital) copying or digitalization of these works, to the extent that these works will be not simply accessed but also extracted, aggregated and then stored as 'data' in one or more locations (e.g. digital copies of photographs, scans of paintings[41], texts relevant to the AI project saved in a server or other tangible medium(s) accessible to the programmers of the project).

9 In the same vein, the works included in the *corpus* may be subject to copying during the so-called *preprocessing stage*[42]. This is a common preparatory stage for the main training process of the machine[43]. During this stage, the aggregated training works will be transformed into a *machine readable and understandable version* (e.g. conversion of a PDF document in plain text format[44]) which fits operational needs of the project[45]. This process implies adaptive use of the works, given that it encompasses the creation of modified copies of the training works (which, however, would probably not qualify as adaptations in the legal sense of the term, due to the lack of originality[46]). These copies will typically be assembled in a database (collection or library), known as the *training dataset of the project*, which will eventually be stored in a remote location, implying again reproduction of the training works[47]. Besides, during this stage, the training works may (also) be subject to manual verification and annotation (labeling). This manual programmers' task[48] aims to enrich the dataset with

[37] The technicalities presented in this paper reflect a simplistic synthesis of stages that may occur in ML activities. They may differ according to the ML technique used.

[38] See, re: TDM, Geiger *et al.*, Crafting (n 11) p. 98: "TDM usually involves some copying, which even in case of limited excerpts might infringe the right of reproduction". *Cf.* Rosati 2019 (n 10) p. 10: "In any case, it is necessary to stress at the outset that not all TDM practices require necessarily the extraction and/or copying of content. This may be because, for instance, the TDM technique employed does not require undertaking such activities at the outset."

[39] Of course, it is also possible that ML workflow is based on preexisting collections of works that may be used as training examples. In this case, the corpus of training data itself may be protected as database, by *sui generis* right and/or copyright. In the present paper we will not further analyze this parameter.

[40] *Cf.* from a NLP approach, Margoni (n 31) section II.

[41] For instance, in the Next Rembrandt Project (<www.thenextrembrandt.com> accessed 3 December 2019), the machine has been trained to produce Rembrandt-style painting on 346 Rembrandt's paintings, that have been previously 3D scanned in high resolution, see Ralf T. Kreutzer & Marie Sirrenberg, *Understanding Artificial Intelligence. Fundamentals, Use Cases and Methods for a Corporate AI Journey* (Springer 2020) 219.

[42] Cf. Reto Hilty & Heiko Richter, *Position Statement of the Max Planck Institute for Innovation and Competition on the Proposed Modernisation of European Copyright Rules Part B Exceptions and Limitations (Art. 3 – Text and Data Mining)*, (2017) Max Planck Institute for Innovation & Competition Research Paper No. 17-02 <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2900110> accessed 3 December 2019, para. 14, p. 4.

[43] *Cf.* Christophe Geiger & Giancarlo Frosio & Oleksandr Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects* (March 2, 2018) Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2018-02 ; SSRN: <https://ssrn.com/abstract=3160586> or <http://dx.doi.org/10.2139/ssrn.3160586> accessed 3 December 2019, p. 5 (referring to TDM): "copying substantial quantities of materials which encompasses: a. preprocessing materials by turning them into a machine readable format and analyzed directly from their source [...]".

[44] See e.g. Margoni (n 31) section II.

[45] For an example of preprocessed musical compositions, see Gaëtan Hadjeres & François Pachet, 'Deep Bach: A steerable model for Bach chorales generation' (3 December 2016) <https://arxiv.org/pdf/1612.01010v1.pdf> accessed 3 December 2019, pp. 4-5.

[46] *Cf.* Geiger *et al.*, Crafting (n 11) p. 98: "[...] pre-processing to standardize materials into machine-readable formats might trigger infringement of the right of reproduction."

[47] *Cf.* Geiger *et al.*, Crafting (n 11) p. 98.

[48] See e.g. Surden (n 4) p. 91, footnote 20: "In many cases, machine learning algorithms are trained through carefully validated training sets of data in which the data gas been carefully screened and categorized by people." ;

labels relevant to targeted patterns, styles etc. and constitutes a feature of the so-called supervised (machine) learning[49]. In this scenario, a similar (and eventual more genuine) adaptive use of the works would take place, deriving from the alterations made by the programmers on the training works (i.e. manual additions of labels and annotations within a text, a painting etc.). Following this intervention, the training dataset will now consist of labeled/annotated copies of training works.

10 The main training stage of the ML workflow, namely the computational and statistical analytical processing / "mining" of the dataset, equally involves copying of the training works. In general, during this stage the machine "reads the works" (a process also called "machine or robot reading") and implements the ML algorithm in order to recognize and extract from the (labeled or unlabeled) training dataset empirical observations, *such as patterns, styles or other micro-elements*[50]. As far as it concerns the algorithmic art field in particular, the algorithmic pattern analysis is oriented in allowing the machine to detect ('learn') *technical and esthetic elements* or other creative aspects[51] (in other words, *ideas*[52]) embodied in these training works[53] and/or predict patterns or features attached to a certain label within the training works[54]. Independently of the ML technique and type of algorithm used, the copying of training works is generally indispensable and unavoidable within this information-acquisition stage[55], given that these data files need to be copied

ibid. p. 93: machine learning often (but not exclusively) involves learning from a set of verified examples of some phenomenon."

[49] For a concise presentation on that topic, see <https://en.wikipedia.org/wiki/Supervised_learning> accessed 3 December 2019; Surden (n 4) p. 93. However, it should be noted that ML may be implemented in the framework of AI-generated art projects with limited or no human guidance, i.e. without verified or labeled data (this method refers to the so-called unsupervised learning and *deep learning*, based on multi-layered artificial neural networks). See on that topic among others, Levendowski (n 5) p. 13: "Alternately, researchers can set an AI system loose on training data with limited human guidance and leave it to the system to determine which features comprise the concept of a cat, a technique called "unsupervised learning."; Andres Guadamuz, 'Do Androids Dream of Electric Copyright? Comparative Analysis of Originality in Artificial Intelligence Generated Works' (2017) SSRN <https://ssrn.com/abstract=2981304> accessed 3 December 2019, p. 3: "Deep Dream transforms a pre-existing image using machine learning mathematical methods that resemble biological neural networks, in other words, the machine mimics human thinking and makes a decision as to how to transform the input based on pre-determined algorithm. What is novel about Deep Dream, and other similar applications of neural networks, is that the program decides what to amplify in the image modification, so the result is unpredictable, but also it is a direct result of a decision made by the algorithm."

[50] In this case, the training data will correspond to the experience needed for the machine to be turned into knowledge. See Shai Shalev-Shwartz & Shai Ben-David, Understanding Machine Learning: From Theory to Algorithms (Cambridge University Press 2014) <http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning> accessed 3 December 2019, p. 19.

[51] *Cf.* Guadamuz (n 49) p. 1, referring to the "Next Rembrandt Project", a Project that led to the creation of a Rembrandt-styled painting, created using deep learning algorithms and facial recognition techniques (<www.thenextrembrandt.com>): "The machine used something called "machine learning" to analyse technical and aesthetic elements in Rembrandt's works, including lighting, colouration, brushstrokes, and geometric patterns. The result is a painting where algorithms have produced a portrait based on the styles and motifs found in Rembrandt's art."; Schönberger (n 3) p. 12-13: "According to the study, the training data allowed the researcher to "explicitly model holistic properties of sentences such as style, topic and high-level syntactic features".

[52] Indeed, from a copyright law view, technical and esthetic patterns usually fall under the sphere of ideas, according to the traditional idea/expression dichotomy. See e.g. Daniel Gervais, 'The Machine As Author', (2019) Iowa Law Review, Vol. 105; SSRN <https://ssrn.com/abstract=3359524> accessed 3 December 2019, p. 24: "TDM is looking, if anything, for ideas embedded in copyright works."

[53] In that case, the machine, through repeated training and practice becomes able to label the patterns, features and characteristics within the dataset by itself. These training algorithms are known as discriminative algorithms. *Cf.* Surden (n 4) p. 91: "After analyzing several such examples, the algorithm may detect a pattern and infer a general "rule". [...] In general, machine learning algorithms are able to automatically build such heuristics by inferring information through pattern detection in data."

[54] In this case, training algorithms are known as Generative Algorithms or Generative Adversarial Networks. See among others Ian Goodfellow & Jean Pouget-Abadie & Mehdi Mirza & Bing Xu & David Warde-Farley & Sherjil Ozair & Aaron Courville & Yoshua Bengio, 'Generative Adversarial Nets', (2014) QC H3C 3J7 Département d'informatique et de recherche opérationnelle, Université de Montreal <https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf> accessed 3 December 2019.

[55] See Schönberger (n 3) p. 16: "Copying the works is indispensable to the training process"; Triaille *et al.* (n 36) p. 29: "technically speaking, it is often considered that data analysis involves, at some stage (particularly in steps 2 and 4 mentioned above), the copying of all or part of the data

in the memory of the machine and/or by computers of a network that is eventually used for the analytical processing of the works (e.g. in case of an analysis implemented through the use of an ML cloud server). However, copying in this case would possibly be temporary and incidental, as these copies do not need to be retained once they are run through the AI system[56].

11 Finally, the ML process may lead to the creation of a robust set of rules that has been abstracted and inferred from the analytical processing of the works *(internal "mental" model*[57]*)*. This is a knowledge-acquisition stage for the machine *(creative) knowledge discovery*[58]*)*. The model will be used by the machine in order to make automated (intelligent) decisions (machine output) regarding new and unknown future input[59], and in particular, in order to proceed with creative "choices" that will lead to the creation of machine-generated art[60]. This set of abstract rules may be eventually saved in a permanent file as the ML training output[61]. This stage would imply adaptive uses or partial reproduction of training works, as long as these works or (some of their protected elements) could be identifiable in their initial or in an altered form within the file of the training output[62]

12 According to the above presentation, ML workflow may involve several copies of training works that could be summarized under two categories: simple reproductions; and copies and adaptive uses of the training works, which, however, might qualify as simple reproductions, as they will not necessarily allow the free and creative choices of the programmer who controls the ML workflow[63]. All the above copies would in principle qualify as acts of reproductions according to art. 2 (1) Infosoc Directive, even if they are not the main objective of the project[64] and, as a consequence, might trigger copyright infringement[65], unless they are rendered lawful (by means of an exception or by contract[66]).

under investigation."

[56] See Schönberger (n 3) p. 16: "[T]he copies do not need to be retained once they are run through the neural network".

[57] Surden (n 4) p. 92: "the rule sets that form the internal model are inferred by examining and detecting patterns within data".

[58] See Eleonora Rosati, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects* (Briefing requested by the JURI Commission of the European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs PE 604.942, 2018) <https://publications.europa.eu/en/publication-detail/-/publication/fdb4ecaa-20f1-11e8-ac73-01aa75ed71a1/language-en/format-PDF/source-search> accessed 3 December 2019, p. 6.

[59] *Cf.* Levendowski (n 5) p. 590: "Most AI systems are trained using vast amounts of data and, over time, hone the ability to suss out patterns that can help humans identify anomalies or make predictions. Well-designed AI systems can automatically tweak their analyses of patterns in response to new data, which is why these systems are particularly useful for tasks that rely on principles that are difficult to explain."

[60] It should be noted that in case of deep learning systems the machine input may involve autonomous creative decisions which may be unpredictable, as machines will be able to mix and combine multiple sources and end up to novel output through its "algorithmic brain paths". Within this context, any human contribution to the output is secondary. This fact raises the fundamental question of the proprietary status of this creative output, which is extensively discussed by legal scholars (see above ftnote 24), but falls outside the scope of this paper.

[61] *Cf.* Margoni (n 31) section II.

[62] *Cf.* Triaille *et al.* (n 36) p. 49 (referring to TDM output): "Normally, the output does not contain any of the original works that were mined, the works have been analysed and only some information were kept."; Geiger *et al.*, Crafting (n 11) p. 99: "[...] the TDM output should not infringe any exclusive rights, as it merely reports on the results of the TDM quantitative analysis, typically not including parts or extracts of the mined materials." A different question arises as to whether the creative output of the machine (e.g. the algorithmic creation) might be qualified as a work deriving from one or multiple works

[63] See, from a NLP perspective, Margoni (n 31) section III.3.c.

[64] *Cf.* Rosati 2019 (n 10) p. 3.

[65] Christophe Geiger & Giancarlo Frosio & Oleksandr Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects* (March 2, 2018), p. 8; Geiger *et al.*, Crafting (n 11) p. 98: "[...] any reproductions resulting in the creation of a copy of a protected work along the chain of TDM activities might trigger copyright infringement." *Cf.* Triaille *et al.* (n 36) regarding data analysis, p. 31.

[66] Alternatively, one could consider that the use of (lawfully accessed) works for ML purposes is simply a normal use of works which falls outside the copyright monopoly by default. However, this is not the approach adopted by the EU legislator. See on that approach, Hilty & Richter (n 42) para. 13, p. 4. *Cf.* also Theodoros Chiou, 'Copyright law and algorithmic creativity: Monopolizing inspiration?' (2019) paper presented at REDA CONFERENCE 2019, University of Cyprus/European University of Cyprus, Nicosia, 21-22 November 2019.

C. Applicability of exceptions and limitations

13 Given the exclusive character of the reproduction right, the above described acts of reproduction that may take place throughout the ML workflow would be lawfully undertaken in the EU territory only after the grant of a (contractual) authorization by rightholders, since they would fall, *a priori*, under the scope of art. 2(1) Infosoc Directive. Naturally, prior authorization would not be necessary only if a (mandatory) exception and limitation of the reproduction right contained in the EU *acquis* could be applicable and cover the acts in question. Although there is no explicit exception and limitation covering the reproductions of copyrighted works for ML purposes, there are, however, at least two existing *mandatory*[67] exceptions, whose application could possibly be relevant. These are:

- the exception for temporary acts of reproductions (art. 5(1) Infosoc Directive)
- and the exception(s)for Text and Data Mining (TDM) (art. 3 and 4 of the DSM Directive).

I. Exceptions for temporary acts of reproduction

14 The exception of temporary acts of reproduction has not been conceived for ML but, basically, for web browsing and caching[68], i.e. the technological advances of the late 90's. However, given its limited[69] but horizontal scope and technological neutrality, it may also be invoked in the ML context[70], insofar its requirements are cumulatively met[71] in accordance with its restrictive interpretation[72]. Temporary acts of reproduction, according to art. 5(1) Infosoc Directive[73], are transient *(ephemeral)* or incidental to an integral and essential part of a technological process and should not present independent economic significance. In addition, this process should enable lawful use of works (i.e. authorized by the rightholder or not restricted by law)[74]. Moreover, according to the ECJ[75], a reproduction act is transient only if its duration is limited to what is necessary for the proper completion of the technological process in question, it being understood that the process *must be automated so that it deletes that act automatically, without human intervention*[76]. Notwithstanding the fact that all the above-mentioned reproductions within the ML workflow are carried out in the context of the implementation of an integral and essential part of a technological process, namely ML, not all of these reproductions would be eligible for this exception.

15 To begin with, beyond some acts of reproductions of training works that are temporary and incidental, such as the copies of works that are likely to be made during the phase of analytical processing of works, there are other several acts of reproduction that are not covered by this exception *ab initio*. In

[67] Non-mandatory exceptions could also be applicable, such as private copying (art. 5(2)(b) of the Infosoc Directive), however they remain unharmonized at the EU level.

[68] See recital 33 Infosoc Directive.

[69] Ch. Geiger, G. Frosio & O.Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects* (March 2, 2018) (n 65) p. 11.

[70] *Cf.* Recital 9 DSM Directive: "acts of reproduction provided for in Article 5(1) of Directive 2001/29/EC, which should continue to apply to text and data mining techniques".

[71] See Infopaq I, para. 55; Order of the Court, in Case C302/10, *Infopaq International A/S v Danske Dagblades Forening* [17 January 2012] ("Infopaq II"), para. 26; Schönberger (n 3) p. 16.

[72] See Infopaq I, para. 56.

[73] See Article 5(1) Infosoc Directive: 1. Temporary acts of reproduction referred to in Article 2, which are transient or incidental [to] an integral and essential part of a technological process and whose sole purpose is to enable: (a) a transmission in a network between third parties by an intermediary, or (b) a lawful use of a work or other subject-matter to be made, and which have no independent economic significance, shall be exempted from the reproduction right provided for in Article 2.

[74] See also recital 33 Infosoc Directive: The exclusive right of reproduction should be subject to an exception to allow certain acts of temporary reproduction, which are transient or incidental reproductions, forming an integral and essential part of a technological process and carried out for the sole purpose of enabling either efficient transmission in a network between third parties by an intermediary, or a lawful use of a work or other subject-matter to be made. The acts of reproduction concerned should have no separate economic value on their own. To the extent that they meet these conditions, this exception should include acts which enable browsing as well as acts of caching to take place, including those which enable transmission systems to function efficiently, provided that the intermediary does not modify the information and does not interfere with the lawful use of technology, widely recognized and used by industry, to obtain data on the use of the information. A use should be considered lawful where it is authorized by the rightholder or not restricted by law.

[75] Infopaq I, para. 64.

[76] See also Margoni (n 31) section IV.2.: "[...] and are automatically destroyed at the end of the process."

fact, several acts of reproductions made within the ML workflow would probably *not be transient*[77]. This would be essentially the case of the reproductions of works that are likely to take place during the *corpus* compilation phase or the reproductions made during the preprocessing/annotation stage of the training material or the abstraction of the internal model. In fact, the deletion of copies in these stages is dependent on the will of the responsible for ML workflow and the AI project[78]. Besides, it is not at all certain that they will wish to dispose these reproductions, which means that there is a risk that the copies will remain in existence for a longer period, according to their needs (e.g. for further development of the AI project or even for trade of these copies)[79]. For the same reasons, these copies *would not be incidental* with regard to the main purpose of use of the work; i.e. The implementation of the learning algorithm and the training of the machine, to the extent that these copies are not temporary[80].

16 Besides, the independent economic significance of acts of reproductions undertaken within the ML workflow cannot be excluded. For instance, corpus compilation might have separable and independent economic significance (if traded in the form of a database), which is distinct to the economic significance of the ML process and output[81]. In fact, the use of works as training material and, in particular, their inclusion in datasets intended for ML projects is already the object of licensing agreements[82].

17 As a consequence, the exception of temporary acts of reproduction does not offer a stable framework for indistinctively manipulating training works within the ML workflow without prior authorization from the rightholders[83], since several acts of reproduction that are likely to take place within the ML workflow will not be covered by this exception[84]. Alternatively, the responsible for ML activity shall be in the position to support the fulfillment of the strict and

[77] *Cf.* Triaille *et al.* (n 36) p. 46 (referring to data mining): "[...] is further unlikely that a temporary copy used to mine data is transient, the work mostly being available for a certain period of time to be transformed, loaded and/or analyzed."

More favorable in exception coverage, Schönberger (n 3) p. 16, stating that "the copies do not need to be retained once they are run through the neural network."

[78] *Cf.* Ch. Geiger, G. Frosio & O. Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects* (March 2, 2018) (n 65) p. 11 re: the application of the temporary reproduction exception to TDM process.

[79] *Cf.* Infopaq I, para. 69-70.

[80] *Cf.* Infopaq II, para. 22, referring to the Infopaq I ruling, on the absence of transient or incidental character of copies made within a data capture process.

[81] *Cf.* Margoni (n 31) section IV.2.: "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."; Triaille *et al.* (n 36) p. 47 (referring to data mining): "It seems that every acts involved in the data mining process can have a great economic value. Potentially, we can imagine that the first extraction can have an independent/separate economic significance, but it depends on what the "miner"/"copy-maker" does with the result of the first extraction (e.g. if he sells or licenses the results of the extraction). It is thus a question of fact."

[82] See for instance the licensing terms of AIVA, a service that allows algorithmic creation of musical compositions, <https://www.aiva.ai/legal/1> accessed 3 December 2019: "Licensee is not permitted to use the Audio and/or MIDI Composition as part of a training dataset for any Machine Learning, Deep Learning or statistical algorithm. If the Licensee wishes to use the Audio and/or MIDI Composition as part of a training dataset, this use case would be ruled by a separate Licensing Agreement, to be negotiated and signed between the parties."*Cf.* Hilty and Ricther (n 42) para. 26, p. 7: "the provision of normalized data solely for the purpose of TDM is a business model".

[83] Margoni (n 31) section IV.2. *Cf.* Triaille *et al.* (n 36) p. 50: "It means that this exception will not provide much relief (or really rarely) for data analysis activities." From a TDM perspective, Ch. Geiger, G. Frosio & O. Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects* (March 2, 2018) (n 65) p. 11: "The mandatory exception for temporary acts of reproduction might apply to limited TDM techniques. Recital 10 of the DSM Draft Directive itself clarifies that this exception still applies but its application would be limited to TDM techniques which involve only the making of temporary reproductions transient or incidental to an integral and essential part of a technological process which enables a lawful use with no independent economic significance. Doubts have been repeatedly casted on whether all these requirements are fulfilled by reproductions done for TDM purposes especially whether these reproductions are transient and have no economic relevance."

[84] *Cf.* Hilty and Richter (n 42) para. 5, p. 2: "In fact, TDM usually requires a not merely temporary reproduction, for which Article 5(1)(a) InfoSoc Directive would not apply."

cumulative requirements of the said exception[85] which derogates the general principle established by Infosoc Directive, namely the requirement that the rightholder authorizes any reproduction of a protected work[86]. This becomes a complicated and precarious task, given that the exception in question did not anticipate the features of ML workflow[87].

II. TDM exceptions within the DSM Directive

18 ML workflow, as seen above, implies computational and statistical analysis of works used as training material. In fact, the analytical processing of training works is a form of data mining, to the extent that it consists in the automated processing of digital materials, which may include texts, data, sounds, images or other elements, or a combination of these, in order to uncover new knowledge or insights[88]. As a consequence, a relationship of intersection might be seen between ML and TDM[89], to the extent that TDM is an essential[90] tool used within the ML workflow, in order to navigate through the training material and produce the necessary derivative data that will train the ML algorithm[91]. Accordingly, the legal regime applying to TDM will also cover TDM activities undertaken within ML context[92]. Thus, the assessment of the applicability of mandatory TDM exceptions introduced by DSM Directive on articles 3 and 4 on the ML workflow seems pertinent.

1. TDM exception introduced by Article 3 DSM Directive

19 Article 3[93]of the DSM Directive introduces a new mandatory exception on the reproduction right

[85] *Cf.* Geiger *et al.*, Crafting (n 11) p. 100, referring to the application of this exception for TDM purposes, mentioning that "application of temporary reproduction exception remains limited to residual cases for the large number of specific requirement that must be fulfilled, apparently in a cumulative manner according to the CJEU."

[86] Infopaq II, para. 27.

[87] Schönberger (n 3) p. 16: "It is quite obvious that the legislator did not have ML in mind when drafting the said provision. Hence some legal uncertainty remains and the related jurisprudence of the CJEU is not without ambiguity." *Cf.* however rec. 9 of the DSM Directive, which explicitly refers to the application of this exception in the context of TDM.

[88] Definition of TDM in Triaille *et al.* (n 36) p. 17.

[89] *Cf.* Schönberger (n 3) p. 17-18: "[A] relationship might be seen between ML and text and data mining (TDM) although ML is much further down the line than TDM, which ultimately aims at the expressive elements of a work creating output derived from such elements".

[90] For the importance of TDM within ML context see e.g. C. Holder, M. Iglesias, J.-P. Triaille, J.-M. Van Gysegnem (eds.), *Legal and regulatory implications of Artificial Intelligence. The case of autonomous vehicles, m-health and data mining*, (Publication Office, Luxembourg 2019) < https://op.europa.eu/en/publication-detail/-/publication/f962b17b-5c04-11e9-9c52-01aa75ed71a1/language-en/format-PDF> accessed 3 December 2019, p. 27: "TDM is an essential component of many AI projects"; Open letter to the Commission, 'Maximising the benefits of Artificial Intelligence through future-proof rules on Text and Data Mining' (9 April 2018) <http://eare.eu/assets/uploads/2018/03/OpenLetter-to-European-Commission-on-AI-and-TDM_9April2018.pdf> accessed 3 December 2019: "foundational role that Text and Data Mining plays in AI"; "a building block for both machine and deep learning"; Geiger *et al.*, Crafting (n 11) p. 97: "Text and data mining (TDM) thus serves as an essential tool to navigate the endless sea of online information [...]". *Adde* 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 3 December 2019.

[91] Geiger *et al.*, Crafting (n 11) p. 109: "TDM has been a fundamental technique to make machine learning possible by copying or crawling massive datasets and empowering artificial intelligence autonomous decision –making and creativity." *Cf.* Rosati 2019 (n 10) p. 2: "Although classical TDM and machine learning have different utility, it should not be overlooked that both use the same key algorithms to discover patterns in data."

[92] *Cf.* Holder *et. al.* (n 90) p. 27: "the legal regime applying to TDM can have an impact on the future development of AI [...]. The development of AI leads to a growing relevance of TDM regime and of its possible weaknesses".

[93] Article 3. Text and data mining for the purposes of scientific research.

1. Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organizations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.

2. 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.

of rightholders for TDM purposes. In particular, according to art. 3(1) of the DSM Directive, reproductions and extractions of works made in order to carry out text and data mining of these works made could be undertaken without prior authorization from the rightholder by non-profit research organizations and cultural heritage institutions[94] for the purposes of scientific research, under the condition that they have lawful access to the works in question and that the copies of works may be stored in a secure environment and no longer than necessary for the purposes of scientific research, including for the verification of research results (art. 3(2) DSM Directive).

20 The wording of the exception is broad in the sense that it covers any reproduction or extraction of work made for TDM purposes, including non-temporary reproductions and it is important that it cannot be overridden by contract. Thus, in the ML context, it would cover reproductions that are necessary both for the (lawful) access to works, their retention and their mining and for a duration that is necessary for the purposes undertaken, which, however, shall be exclusively purposes of scientific research. Moreover, the above exception covers the TDM activities undertaken within the ML context carried out by a specific category of beneficiaries[95], i.e. research organizations and cultural heritage institutions. The TDM exception of Art. 3 could accommodate copies of training works that are connected to their analytical processing made within ML workflow, insofar as they are undertaken by the small circle of beneficiaries of that exception and that their analytical processing aims at purposes of scientific research. Due to this narrow approach regarding the beneficiaries and purposes of TDM activity, the exception could be invoked regarding very specific ML projects and certainly not by startups and other businesses of the private sector (even if they engage in analytical processing of works within ML context for scientific purposes).

2. TDM exception introduced by article 4 DSM Directive

21 Article 4[96] of the DSM Directive TDM introduces a more inclusive exception than the one of Article

3. Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.

4. Member States shall encourage rightholders, research organizations 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.

[94] On that point, see Ch. Geiger, G. Frosio & O. Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects* (March 2, 2018) (n 65) p. 26: "much discussion regarding this proposal does concern whether the TDM exception's beneficiaries should not be limited to research organizations. To qualify for the exception, research organisations must operate on a not-for-profit basis or by reinvesting all the profits in their scientific research, or pursuant to a public interest mission."

[95] Critical on this narrow approach, already re: the DSM Directive Proposal, Ch. Geiger, G. Frosio & O. Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects* (March 2, 2018) (n 65) p. 32: "The TDM exception should not be limited to research organisations but extended to all those enjoying lawful access to underlying mined materials – as the right to read should be the right to mine- especially in order not to cripple research from start-ups and independent researches." ; European Copyright Society, General Opinion on the EU Copyright Reform Package, (24 January 2017), available at: <https://europeancopyrightsocietydotorg.files.wordpress.com/2015/12/ecs-opinion-on-eu-copyright-reform-def.pdf>, part 2, p. 5: "we therefore regret the fact that the Directive proposes to limit the benefits of the exception to "research organisations" as narrowly defined in the Directive. In our view, data mining should be permitted for non-commercial research purposes, for research conducted in a commercial context, for purposes of journalism and for any other purpose."; Rosati (n 10) p. 9: "Its scope, however, should not be unduly narrow and such as to stifle innovation coming from different sectors, whether research organizations or businesses. In this sense, the EU legislature should carefully consider who the beneficiaries of the resulting exception should be, as well as the uses allowed of works or other subject-matter for TDM purposes."

[96] Article 4. Exception or limitation for text and data mining. 1. Member States shall provide for an exception or limitation to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, Article 4(1)(a) and (b) of Directive 2009/24/EC and Article 15(1) of this Directive for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining.

2. 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.

3. The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine readable means in the case of content made publicly available online.

3. In particular, all reproductions and extractions of works and other subject matter made for the purposes of text and data mining are exempted from the rightholder's monopoly, insofar as the works are lawfully accessible and the reproductions and extractions are retained for as long as is necessary for the purposes of text and data mining.

22 This exception could be invoked, *a priori*, within the framework of any ML project, in order to cover all reproductions and extractions connected with the analytical processing of protected training works, as it does not contain a *ratione personae* or purpose limitation. Nonetheless, according to art. 4 (3) of the DSM Directive[97], the application of this exception may be *opted-out* in an appropriate manner by the rightholders. This opt-out may be exercised either by use of technical measures, such as machine-readable means and metadata[98], or contractual agreements[99] (such as terms and conditions of a website or a service[100]), or even unilateral declarations such as disclaimers[101], by which the rightholder would reserve the right to make reproductions and extractions for data analysis purposes under their exclusive control.

23 Notwithstanding its general character, this TDM exception still fails to offer a stable ground for using (reproducing) protected works for ML purposes. In fact, the lawful analytical processing would require prior legal assessment regarding the exercise of the opt-out mechanism provided in art. 4(3) of the DSM Directive. This raises significant obstacles in undertaking ML activities in the EU territory, even for works that are lawfully available online. True, the main source of training data for ML projects is the Web itself and the information generally available therein[102] and this could also apply in the field of algorithmic art to some extent. However, the access to freely and lawfully available works online does not necessarily mean lawful access for TDM purposes[103], since the rightholder would be in position to reserve his rights on data analysis of their works by use of appropriate means, as described above.

24 In sum, according to the current TDM exception regime, rightholders generally *remain able to license* and, consequently, to forbid, the uses and reproductions of their works for data analysis purposes, including analytical processing in the ML context[104], except for reproductions and extractions made by research organizations and cultural heritage institutions for the purposes of scientific research, according to art. 3 DSM Directive. Therefore, possibly most ML projects could not simply rely on the above TDM exceptions for freely using training works within the ML workflow they implement. Due to the opt-out mechanism introduced by art. 4(3) DSM, the use and reproductions of training works for their analytical processing within the ML context implies confirmation as to whether it could be undertaken without prior authorization from the rightholder. This, however, unavoidably involves time consumption, costs and, in some cases, uncertainty while it jeopardizes the application of the exception in practice[105].

4. This Article shall not affect the application of Article 3 of this Directive.

[97] Article 4(3). The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online.

[98] E.g. by adding robot.txt type metadata to their content online, see Hugenholtz (n 90).

[99] *Cf. a contrario* art. 7 para. 1 DSM Directive.

[100] Holder *et. al.* (n 90) p. 28: "[...] on a website, the terms and conditions could still validly prohibit TDM being made of the contents of the website."

[101] Rec. 18 DSM Directive.

[102] See Holder *et. al.* (n 90) p. 29.

[103] See Rosati (n 10) p. 4: "freedom of access does not necessarily entail that the content (text and data) is also free of legal restrictions."; *ibid.*, p. 5: "Lawful access to content – whether because such *content is freely accessible* or access has been obtained through a *licence* – does not necessarily entitle one to undertake TDM in respect of such content (text or data)." See also Rec. 18 DSM Directive.

[104] *Cf.* Hilty & Richter (n 42) para. 7, p. 3, referring to the draft proposal of DSM Directive:"The proposed limitation would allow for the conclusion *e contrario* that TDM is a separable type of use."

[105] See on that point, Rosati 2019 (n 10) p. 21. *Cf.* Daniel Gervais, 'Exploring the Interfaces between Big Data and Intellectual Property Law', (2019) JIPITEC 10 (1) <https://www.jipitec.eu/issues/jipitec-10-1-2019/4875/#ftn.N10113> accessed 3 December 2019, para. 46: "first, it is not always clear to a *human* user whether a source is legal or not; the situation may be even less clear for a machine. Second, and relatedly, if the source is foreign, a determination of its legality may require an analysis of the law of the country of origin, as copyright infringement is determined based on the *lex loci delicti*—and this presupposes a determination of its origin (and foreignness) to begin with. Perhaps a requirement targeting sources that the user *knows or would have been grossly negligent in not knowing* were illegal might be more appropriate."

D. Conclusions

25 In light of the preceding analysis, some conclusions may be formulated.

Firstly, in the era of the 4th industrial revolution and Web 4.0, works will not perceived merely as digital content but rather as (big) data that are used as "training material" in order to "teach" machines how to make 'intelligent decisions', including the production of algorithmic creations. In addition, *works are also turned into (meta)-data*[106], especially through their analytical processing, which allows the recognition and extraction of patterns, styles and other features to be read and understood by machines.

26 Secondly, Machine Learning is a so-called copy-reliant technology[107]. As such, given the broad definition of reproduction right in art. 2 of the Infosoc Directive and the broad interpretation made by the ECJ, it is in principle subject to the realm of copyright in the EU.

27 As to the possible copyright exceptions, the coverage of the entire ML workflow and all acts of reproductions undertaken therein by sole or combined application of the mandatory exceptions that are relevant within the ML workflow (exception for temporary reproductions and TDM exceptions) is not straightforward and, in any event, should be examined on a case by case basis, given the variety of techniques and methods employed[108]. In addition, the formulation and limited scope[109] of the above-mentioned mandatory exceptions and their restrictive interpretation by the ECJ give rightholders the possibility to still *veto* the use of their works in many ML projects, including, the use of works as machine reading material[110], within the ML workflow. As a consequence, the current EU copyright law framework seems more favorable for rightholders' interests (especially since TDM and its employment for ML purposes, among others, is an activity subject to copyright restraints) and does not offer a stable and enabling legal framework for engaging in several ML activities that rely on copyrighted training works[111], including algorithmic art. This situation leads to legal uncertainty as to which acts of reproduction may be undertaken without prior authorization of rightholders[112]. Accordingly, the lawful use of preexisting works as training material would require prior assessment of their legal status of protection and eventual prior clearance of rights (most probably on a work by work basis)[113].

28 A no-risk approach towards use of works for ML purposes in the EU would be satisfied by the use of copyrighted training works on the grounds of a license agreement[114]or the use of non-copyrighted works as training material. Under these conditions, the utility of the use of open content as training

[106] Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning' (2019) Journal of the Copyright Society of the USA, Vol 66; SSRN <https://ssrn.com/abstract=3331606> or <http://dx.doi.org/10.2139/ssrn.3331606> accessed 3 December 2019, p. 59 ff.

[107] For that concept see Matthew Sag, 'Copyright and Copy-Reliant Technology' (2009) Northwestern University Law Review Vol. 103; The DePaul University College of Law, Technology, Law & Culture Research Series, Paper No. 09-001; SSRN: <https://ssrn.com/abstract=1257086> accessed 3 December 2019; Schönberger (n 3) p. 14.

[108] It should also be noted that the applicability of existing exceptions does not thwart moral rights questions (such as the paternity or integrity right) that may arise by the use of works within the ML process and especially their transformative manipulation.

[109] See among others, Geiger *et al.*, Crafting (n 11) p. 110: "It's narrow scope, however, will limit these substantive positive externalities to a comparatively small number of research institutions, while the DSM at large will still lag behind other jurisdictions, allowing a larger cluster of market players to engage legally in TDM activities."

[110] It is argued by commentators that machine reading should be exempted from copyright law realm. See for instance James Grimmelmann, 'Copyright for Literate Robots' (2016) 101(2) Iowa Law Review, 657: "[...] copyright law has concluded that it is for humans only: reading performed by computers doesn't count as infringement."

[111] *Cf.* Sag 2019 (n 106) p. 38: "there are very few places where the law is as clear and/or as favorable as in the United States".

[112] *Cf.* Hilty and Richter (n 42) para. 2, p. 1: "A clear legal framework avoids the complicated rights clearance between the parties involved and reduces investment risks."

[113] Of course, on a practical note, proving the use of a work as training material is not always easy for rightholders, since the creative output may be sufficiently differentiated from all training works. *Cf.* Triaille *et al.* (n 36) p. 87.

[114] Indeed, there seems to be an emergent derivative market of use of works for TDM purposes, which might extend to ML. However, the use of works for ML purposes as an object of licensing contracts should be further investigated to the extent that it could be qualified as a new (unknown) form of exploitation, which might raise additional implications in some jurisdictions. *Cf.* for a similar question regarding cloud computing from a Greek Law perspective, Th. Chiou, Music Licensing in the Cloud: The Greek Experience, (2014) GRUR Int., 3/2014, p. 228 ff.

material is important[115], as these works would often be fit for ML purposes, without the need to invoke the applicability of exceptions[116]. It might not be accidental that some emblematic AI projects in the EU are based on works of the public domain[117].

29 In sum, it seems that the new DSM Directive follows an approach that fits better to the digital era than to the new era of the 4th industrial revolution which features the penetration of AI systems in the field of creativity[118]. This means that the DSM Directive is a missed opportunity for *true* modernization of the European Copyright Law in the digital single market[119], to the extent that it does not take into account and, *a fortiori*, does not enhance the development of innovative machine art projects[120]. Nor does it improve the Union's competitive position, compared to other jurisdictions, as a prominent area in development of ML techniques, especially in the field of computer art[121]. Most importantly, the approach adopted in regulating the reproductions of protected works within the ML context might turn into an "own goal" in the age of algorithmic creations, if the new paradigm of creativity process is subject to copyright constraints[122].

[115] Triaille *et al.* (n 36) p. 25: "it goes beyond the scope of this Study to analyze the overall impact which the Open Access movement will have on TDM but it seems undeniable that it will facilitate TDM."

[116] *Cf.* Rec. 9 DSM Directive: "Text and data mining can also be carried out in relation to mere facts or data that are not protected by copyright, and in such instances no authorisation is required under copyright law."; Ch. Geiger, G. Frosio, O. Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects* (March 2, 2018) (n 65) p. 7: "works and other subject matter not protected by copyright or sui generis rights can be freely mined."; Cf. Sag 2019 (n 106) p. 49: "For example, Wikipedia includes a cornucopia of over 5 million Creative Commons licensed works in a fully machine readable format. This has made Wikipedia a key source of training data for nearly every modern AI system dealing with facts."

[117] See for instance the Next Rembrandt Project, where the machine has been trained to produce Rembrandt-style painting by using as training data 346 known paintings by Rembrandt (d. 1669), that are on the public domain. See https://www.nextrembrandt.com/.

[118] It should be noted that the terms "machine learning" and "artificial intelligence" are absent from the official texts, including the COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT on the modernisation of EU copyright rules Accompanying the document Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market and Proposal for a Regulation of the European Parliament and of the Council laying down rules on the exercise of copyright and related rights applicable to certain online transmissions of broadcasting organisations and retransmissions of television and radio programmes, Brussels, 14.9.2016 SWD(2016) 301 final PART 1/3 {COM(2016) 593} {COM(2016) 594} {SWD(2016) 302}.

[119] Although "[...] the objective of this Directive [is] the modernisation of certain aspects of the Union copyright framework [in order] to take account of technological developments and new channels of distribution of protected content in the internal market [...]", according to Recital 83 of the DSM Directive.

[120] Although "relevant legislation needs to be future-proof so as not to restrict technological development", according to Recital 3 of the DSM Directive.

[121] See Geiger *et al.*, Crafting (n 11) p. 110: "This might result in a critical weakness for the DSM while racing to reach a dominant position in the market for artificial intelligence technology. Being unable to make full use of the immense riches made available by big data streams in digital networks for artificial intelligence, machine learning, and neural network applications will put Europe in a disadvantaged position from which it might be hard to recover in the future."; Hugenholtz, (n 90); Rosati 2019 (n 10) p. 23, making reference also to the stage of national transposition of art. 4 DSM Directive.

[122] *Cf.* Rosati 2019 (n 10) p. 21: "In practice, this might have a negative impact on the (unlicensed) development of AI creativity."