MACHINES INFRINGING ON COPYRIGHT
Liability and Justifications in Machine Learning

Peltoniemi Sara
Alustatalous ja EU:n digitaaliset sisämarkkinat
University of Turku Faculty of Law
July 2020

TURUN YLIOPISTO
Oikeustieteellinen tiedekunta
PELTONIEMI, SARA: Machines Infringing on Copyright: Liability and Justifications in Machine Learning
OTM-opinnäytetyö, 61 s.
Immateriaalioikeus
Heinäkuu 2020
Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin Originality Check-järjestelmällä.

Tekijänoikeusjärjestelmän ja teknologian välillä on jo pitkään ollut jännitteitä. Uudet teknologiset menetelmät haastavat tekijänoikeussääntelyä mukautumaan. Uusin näistä ilmiöistä on koneoppiminen. Vaikka koneoppiminen ei itsessään ole uusi ilmiö on sen tuomat haasteet tekijänoikeudelle tulleet pinnalle vasta viime aikoina.

Koneoppiminen on osatekoälyä ja sitä hyödynnetään monilla eri aloilla. Tutkimuksessani keskityn koneoppimisen hyödyntämiseen uuden musiikin luomisessa. Tekijänoikeus suojaa musiikkia monilla eri tavoilla sävellyksestä ja sanoituksesta äänitallenteeseen. Koneoppimisessa tärkeää on datan hyödyntäminen koneen opetusvaiheessa. Opetusdatan tulee olla samanlaista kuin halutun lopputuloksen. Näin ollen tässä tapauksessa opetusdatan tulee olla tekijänoikeudella suojattua materiaalia.

Ensimmäinen ongelma koneoppimisessa tekijänoikeuden näkökulmasta on suojattujen teosten käyttäminen opetusdatana. Toinen ongelma on vastuu koneen luomien teosten mahdollisesti rikkoessa toisen tekijänoikeutta. Tutkimukseni käsittelee mahdollisia tekijänoikeusrikkomis tilanteita sekä poikkeuksia tekijänoikeudessa joiden perusteella suojattujen teosten käyttö olisi sallittua koneoppimisessa.

Asiasanat: koneoppiminen, tekoäly, tekijänoikeus, tekijänoikeus rikkomukset, immateriaalioikeus

UNIVERSITY OF TURKU
Faculty of Law
PELTONIEMI, SARA: Machines Infringing on Copyright: Liability and Justifications in Machine Learning
Master Thesis, 61 s.
Intellectual property rights
July 2020
The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service.
– –

There have been tensions between the copyright system and technology for a long time. New technological methods challenge copyright regulation to adapt. The latest of these phenomena is machine learning. While machine learning is not in itself a new phenomenon, its challenges to copyright have only recently surfaced.

Machine learning is a part of artificial intelligence and is utilized in many different fields. In my research, I focus on utilizing machine learning to create new music. Copyright protects music in many different ways, from compositions and lyrics to sound recordings. In machine learning, the use of data in the learning phase is important. The teaching data should be similar to the desired outcome. Therefore, in this case, the teaching data must be copyrighted material.

The first problem with machine learning from a copyright perspective is the use of protected works as learning data. Another problem is liability for works created by machines if they infringe on someone's copyright. My research deals with possible situations of copyright infringement as well as exceptions in copyright on the basis of which the use of protected works would be allowed in machine learning.

Keywords: machine learning, artificial intelligence, copyright, copyright infringements, intellectual property rights.

TABLE OF CONTENTS

REFERENCES

Bibliography

Aarnio (1989) A Aarnio *Laintulkinnan teoria: yleisen oikeustieteen oppikirja* (WSOY 1989)

Alpaydin (2014) E Alpaydin *Introduction to machine learning* (Third edition The MIT Press 2014)

Atkinson and Fitzgerald (2014) B Atkinson and B Fitzgerald, *A Short History of Copyright: The Genie of Information* (2014)

Axhamn (2016) Johan Axhamn, *Databasskydd* (Stockholms universitet , 2016)

Ballardini, He and Roos (2019) Rosa Maria Ballardini, Kan He & Teemu Roos, 'AI-Generated Content: Authorship and Inventorship in the Age of Artificial Intelligence' in Taina Pihlajarinne, Juha Vesala and Olli Honkkila (eds), *Online Distribution of Content in the EU* (Edward Elgar Publishing 2019)

Coelho, Richert and Brucher (2018) L P Coelho, W Richert and M Brucher *Building machine learning systems with Python: explore machine learning and deep learning techniques for building intelligent systems using scikit-learn and TensorFlow* (3th edition, Packt Publishing 2018)

Edelman (1979) Bernard Edelman, *Ownership of the image: elements of the Marxist theory of law* (Routledge & Kegan Paul 1979)

Emrouznejad and Charles (2018) Ali Emrouznejad and Vincent Charles. *Big Data for the Greater Good*. (Vol. 42. Cham: Springer 2018)

Geiger, Frosio and Bulayenko (2018) C Geiger, G Frosio and O Bulayenko 'Crafting a Text and Data Mining Exception for Machine Learning and Big Data in the Digital Single Market' in X Seuba, C Geiger and J Pénin (eds), *Intellectual Property and Digital Trade in the Age of Artificial Intelligence and Big Data* (CEIPI/ICTSD publication series on "Global Perspectives and Challenges for the Intellectual Property System", Issue No. 5 2018).

Günther (2019) Petteri Günther, 'Chapter 4: Industrial Internet Solutions and Data Ownership Versus Control Over Data' in Rosa-Maria Ballardini, Olli Pitkänen, et al. (eds), *Regulating Industrial Internet through IPR, Data Protection and Competition Law* (Kluwer 2019)

Harenko, Niiranen and Tarkela (2016) Kristiina Harenko, Valtteri Niiranen and Pekka Tarkela, *Tekijänoikeus* (2nd edition, Talentum Media, 2016)

Hilty (2018) R Hilty 'Big Data: Ownership and Use in the Digital Age' in X Seuba, C Geiger and J Pénin (eds), *Intellectual Property and Digital Trade in the Age of Artificial Intelligence*

and Big Data (CEIPI/ICTSD publication series on "Global Perspectives and Challenges for the Intellectual Property System", Issue No. 5 2018).

Husa (2013) J Husa *Oikeusvertailu* (Lakimiesliiton kustannus 2013)

Hutter (2005) M Hutter, *Universal Artificial Intelligence – Sequential Decisions Based on Algorithmic Probability*, (Springer 2005)

Laddie and others (2000) Huhg Laddie and others *The Modern Law of Copyright and Designs* (3th edition, Butterworths, 2000)

Lovelace (1843) Ada Lovelace notes on Charles Babbage's Analytical Engine published in *Scientific Memoirs Selected from the Transactions of Foreign Academics of Science and Learned Societies* (1843), 691.

Mitchell (1997) T M. Mitchell, *Machine Learning* (McGraw-Hill 1997)

Nilsson (2010) N J. Nilsson *The quest for artificial intelligence – a history of ideas and achievements* (2010) Cambridge university press

Pihlajarinne and Ballardini (2019) Taina Pihlajarinne and Rosa-Maria Ballardini 'Chapter 9: Owning Data via Intellectual Property Rights: Reality or Chiemera?' Networks' in Rosa-Maria Ballardini, Olli Pitkänen, et al. (eds), *Regulating Industrial Internet through IPR, Data Protection and Competition Law* (Kluwer 2019)

Pila and Torremans (2016) J Pila and P Torremans, *European Intellectual Property Law* (Oxford University Press 2016)

Vesala and Ballardini (2019) Juha Vesala and Rosa-Maria Ballardini, 'Chapter 6: AI and IPR Infringement: A Case Study on Training and Using Neural Networks' in Rosa-Maria Ballardini, Olli Pitkänen, et al. (eds), *Regulating Industrial Internet through IPR, Data Protection and Competition Law* (Kluwer 2019)

Articles

Barron (2004) Anne Barron, 'The Legal Properties of Film' (2004) The Modern Law Review vol. 67(2), 177-208.

Boyd and Crawford (2011) Danah Boyd and Kate Crawford, 'Six Provocations for Big Data' (2011) A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.

Bridy (2012) AnneMarie Bridy, 'Coding Creativity: Copyright and the Artificially Intelligent Author' (2012) 2012 Stan Tech L Rev 5

Bridy (2016) AnneMarie Bridy, 'The evolution of Authorship: Work Made by Code' (2016) vol. 39 Columbia Journal of Law & the Arts 395

Carroll (2019) Michael W. Carroll, ’Copyright and the Progress of Science: Why Text and Data Mining Is Lawful’ (2019) UC Davis Law Review Vol. 53:893, pp.895-964

Cason and Müllensiefen (2012) Robert J S Cason and Daniel Müllensiefen, 'Singing from the Same Sheet: Computational Melodic Similarity Measurement and Copyright Law' (2012) vol. 26 International Review of Law, Computers & Technology 25-36.

Chavannes (2018) Remy Chavannes ‘IP protection of deep learning systems’ (2018) <https://blog.chavannes.net/2018/10/ip-protection-of-deep-learning-systems/> accessed 4.3.2020.

Chiou (2019) Theodoros Chiou ‘Copyright lessons on Machine Learning: what impact on algorithmic art?’ (2019) Jipitec 10(3) 398.

Cockfield (2004) Arthur Cockfield ’Towards a Law and Technology Theory’ (2004) Manitoba Law Journal vol. 30 (3) 383

Commission Staff Working Document – Impact Assessment on the modernization of EU copyright rules – Part 1.

Copeland (2000) Jack Copeland, ‘What is Artificial Intelligence?’ (2000) <http://www.alanturing.net/turing_archive/pages/reference%20articles/what%20is%20ai.html > accessed 7.3.2020.

Denicola (2016) Robert C. Denicola ‘Ex Machina: Copyright Protection for Computer Generated Works’ (2016) Rutgers University Law Review vol. 69, 251.

Fairhurst (2019) Oliver Fairhurst, ’When does AI infringe copyright?’ (2019) < http://ipkitten.blogspot.com/2019/03/when-does-ai-infringe-copyright.html> accessed 20 February 2020.

Gervais (2007) Daniel J Gervais, ‘The Protection of Databases’ (2007) 82 Chi-Kent L Rev 1109.

Gillick, Tang and Keller (2010) Jon Gillick, Kevin Tang and Robert M. Keller, ’Machine Learning of Jazz Grammars’ (2010) Computer Music Journal vol. 34, no. 3, pp. 56-66.

Grimmelmann (2016) James Grimmelmann, 'Copyright for Literate Robots' (2016) 101 Iowa L Rev 657.

Jütte (2017) BJ Jütte, ‘Reconstructing European Copyright Law for the Digital Single Market : Between Old Paradigms and Digital Challenges’ [2017] Luxemburger Juristische Studien – Luxembourg Legal Studies

Lae (2011) Elina Lae, ’Mashups – A Protected Form of Appropriation Art or a Blatant Copyright Infringement?’ [2011] available at SSRN < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2003854> accessed 27 January 2020.

Lambert (2015) Paul Lambert, ‘Computer Generated Works and Copyright: Selfies, Traps, Robots, AI and Machine Learning’ (2015) 34 European Intellectual Property Review

Land, Waibel and Hinton (1990) Kevin J. Land, Alex H. Waibel and Geoffrey E. Hinton, ’A Time-Delay Neural Network Architecture for Isolated Word Recognition’ (1990) Neural Networks Vol. 3. 23.

Lehr and Ohm (2017) David Lehr and Paul Ohm, 'Playing with the Data: What Legal Scholars Should Learn about Machine Learning' (2017) 51 UCD L Rev 653.

Leval (1990) Pierre N Leval, 'Toward a Fair Use Standard ' (1990) 103 Harv L Rev 1105.

Lukoseviciene (2019) Aurelija Lukoseviciene, ‘The contractual protection of authors in Copyright in the Digital Single Market Directive – does the reality live up to the expectations?’ (2019) NIR.

Manyika, Chui, et al. (2011) James Manyika, Michael Chui, et al. ‘Big Data: The next frontier for innovation, competition, and productivity’ (2011) Mckinsey Global Institute report.

Margoni (2018) Thomas Margoni, ‘Artificial Intelligence, Machine learning and EU copyright law: Who owns AI?’ (2018) CREATe Working Paper 2018/12.

Pila (2010) Justine Pila, ’Copyright and Its Categories of Original Works’ (2010) Oxford Journal of Legal Studies, Vol. 30, No. 2, pp.229-254

Reynolds (2009) Graham Reynolds, ’A Stroke of Genius or Copyright Infringement? Mashups, Copyright, and Moral Rights in Canada’ (IP OSGOODE, 24 August 2009) <https://www.iposgoode.ca/2009/08/a-stroke-of-genius-or-copyright-infringement-mashups-copyright-and-moral-rights-in-canada/> accessed 27 January 2020.

Sag (2009) Matthew Sag, 'Copyright and Copy-Reliant Technology' (2009) 103 Nw U L Rev 1607.

Sag (2020) Matthew Sag DOES COPYRIGHT REQUIRE AUTHORIZATION TO USE DATA “SUBSISTING IN COPYRIGHT WORKS?” (2020) <http://infojustice.org/archives/42016> accessed 30.6.2020.

Samuelson (1986) Pamela Samuelson, ‘Allocating Ownership Rights in Computer-Generated Works’ (1986) 47 University of Pittsburgh Law Review 1185.

Schlackman (2018) S Schlackman, 'Who holds the Copyright in AI created Art' (artrepreneur, art law journal, 22 April 2018) <https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html> accessed 23 January 2020

Schönberger (2018) Daniel Schönberger, 'Deep Copyright: Up- and Down-stream - Questions Related to Artificial Intelligence (AI) and Machine Learning (ML)' (2018) <https://ssrn.com/abstract=3098315> accessed 25.3.2020.

Smith (2013) S Smith, 'Iamus: Is this the 21st century's answer to Mozart?' (BBC News, 3 January 2013) <https://www.bbc.com/news/technology-20889644> accessed 21 January 2020

Sobel (2017) Benjamin L.W. Sobel, 'Artificial Intelligence's Fair Use Crisis' (2017) 41 Columbia Journal of Law & the Arts 45.

Timonen (1998) Pekka Timonen 'Oikeustaloustiede – mitä se on?' (1998) Lakimies 1/1998 100.

Internet

<folkrnn.org> accessed 21 January 2020

Billboard Top 100 Chart <https://www.billboard.com/charts/hot-100> accessed 23 February 2020.

Spotify Top 50 Global <https://open.spotify.com/playlist/37i9dQZEVXbMDoHDwVN2tF> accessed 23 February 2020.

Podcast by Tyler Renelle, 'Machine Learning Guide' <http://ocdevel.com/mlg> accessed 7.3.2020.

Theory of music <http://www2.siba.fi/muste1/index.php?id=1&la=fi> accessed 7.3.2020.

'The world's most valuable resource is no longer oil, but data' (The Economist, May 6th 2017 edition) <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data> accessed 14.3.2020

< https://www.billboard.com/p/billboard-charts-legend> accessed 11 March 2020.

Elements of AI, < https://course.elementsofai.com> accessed 12.3.2020.

WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI): Second Session Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence < https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf> accessed 28.3.2020.

WIPO Conversation on Intellectual Property Policy and Artificial Intelligence (AI): Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence < https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=499504> accessed 20.6.2020.
JISC (2012) The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. Available at: <http://bit.ly/jisc-textm> accessed 5.7.2020.
<https://www.eecis.udel.edu/~amer/Table-Kilo-Mega-Giga---YottaBytes.html> accessed 6.7.2020.

Legislation

Berne Convention for the Protection of Literary and Artistic Works: Texts. Geneva: World Intellectual Property Organization, 1982.
WIPO Copyright Treaty, 1996 **WCT.**

EU

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 [2002] OJ L 167/10 **Infosoc Directive**.
Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L 77/20 **Database directive**.
Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs [1991] OJ L 122/42, replaced by Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L 111/16 (codified version) **Software directive**.
Council Directive 93/98/EEC of 29 October 1993 harmonising the term of protection of copyright and certain related rights [1993] OJ L 290/9, replaced by 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 [2006] OJ L 372/12 (codified version) **Term Directive**.
Directive 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 Directive 96/9/EC and 2001/29/EC [2019] OJ L 130/92 **DSM-directive**.
The Treaty on the Functioning of the European Union [2012] OJ C326/47.
Treaty of Lisbon amending the Treaty on European Union and the Treaty establishing the European Community, OJ C306, 17.12.2007.

FINLAND

HE 177/2002 vp.

UK

The Copyright and Rights in Performances (Research, Education, Libraries, and Archives) Regulations 2014, SI 2014/1372.

US

United States Copyright Act U.S.C..

U.S. COPYRIGHT OFFICE, COMPENDIUM OF U.S. COPYRIGHT OFFICE PRACTICES § 101 (3d ed. 2017)

TABLE OF CASES

GERMANY

Case I ZR 290/02 *Hit Bilanz* BGH 21 July 2005.

EU

C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, [2009] EU:C:2009:465.

C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, [2009] EU:C:2009:465, Opinion of AG Trstenjak.

C-283/10 *Circul Globus Bucureşti (Circ & Variete Globus Bucureşti) v Uniunea Compozitorilor şi Muzicologilor din România – Asociaţia pentru Drepturi de Autor (UCMR-ADA)* [2011] ECR I-12031.

C-604/10 *Football Dataco Ltd and others v Yahoo! UK Ltd and others* [2012] EU:C:2012:115

C-263/18 *Nederlands Uitgeversverbond, Groep Algemene Uitgevers* v *Tom Kabinet*, [2019] EU:C:2019:1111.

C-128/11 *UsedSoft* [2012] EU:C:2012:407

C-145/10 *Eva-Maria Painer v. Standard VerlagsGmbH and Others* [2013] ECLI:EU:C:2013:138

C-429/08 *Karen Murphy v. Media Protection Services Ltd* [2011] ECLI:EU:C:2011:631

C-173/11 *Football Dataco Ltd and Others v. Sportradar GmbH and Sportradar AG* [2012] ECLI:EU:C:2012:642.

US

Newton v. Diamond, 388 F.3d 1189 (9th Cir. 2004).

Authors Guild, Inc. v. Google, Inc., 804 F.3d 202 (2d Cir. 2015)

Ticketmaster Corp. V. Tickets.com, Inc., No. CV 99–7654 (C.D. Cal. 2000)

Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884).

Sega Enters. V. Accolade, Inc., 977 F.2d 1511 (9th Circ. 1992).

White-Smith Music Publishing Co. v. Apollo Co., 209 U.S. 1 (1908).

Campbell v. Acuff-Rose Music, Inc., 510 U.S. (1994).

Feist Publications v. Rural Telephone Service Co, 499 U.S. 340 (1991).

UK

Ladbroke (Football) Ltd v. William Hill (Football) Ltd [1964] 1 WLR. 273.

Designer Guild Ltd v Russell Williams (Textiles) Ltd [2001] 1 WLR 2416.

Nicol v Barranger [1917-23] MCC 219.

Austin v Columbia Gramophone 1917.

Francis Day Hunter Ltd v Bron, [1963] Ch. 587

WTO Panel

United States – Section 110(5) of US Copyright Act, Report of the Panel, WT/DS160/R (15 June 2000).

ABBREVIATIONS

AI	Artificial Intelligence
CJEU	Court of Justice of the European Union
EU	European Union
IPR	Intellectual Property Right
ML	Machine Learning
TDM	Text and data mining
TEU	Treaty of the European Union
TFEU	Treaty on the Functioning of the European Union

FIGURES

1 INTRODUCTION

1.1. Background

The fraction between technology and copyright is not a new phenomenon. For consumers the big sifts have been the PC revolution in 1980's and even more the Internet revolution in the 1990s. The internet made accessing and distributing works easier and faster than ever before. The copyright protection of works became more challenging and this caused problems with the protection.[1]

Before the digital revolutions, enforcing copyright and monitoring infringements was in fact a lot easier. The world was smaller in a sense that works did not reach as many people and because the public was also smaller it was easier to monitor the use of the work. With computers and the internet enforcing the exclusive right requires a lot of resources and time and it is still impossible to seize all of the infringements that occur at the digital platforms. The problem with these technologies is both qualitative and quantitative. The qualitative problem is all the new ways copyright protected works can be used and the quantitative problem is the possibility to copy works in massive volumes.[2] One question, that surely concerns copyright owners, is the use of their works by new technological means such as training data in machine learning (ML).

The complication with machine learning and copyright is seen as a new problem, however there have been cases about machines and copyright already in the early 20th century. One example is the case *White-Smith Music Publishing Co. v. Apollo Co.*[3] where a player piano and a robot piano played original musical works with the use of perforated paper rolls that only the player piano could read. The court held that because no human could read the paper rolls and they were part of a machine, the use of original works in the player pianos did not constitute copyright infringement. This case resulted to a new copyright act in the US which extended the copyright protection to 'perform the copyrighted work publicly - - in - - any form of record in which the thought of an author may be recorded and from which it may be read or reproduced'.

[1] Bridy (2012) 2.
[2] Grimmelmann (2016), 661.
[3] White-Smith Music Publishing Co. v. Apollo Co., 209 U.S. 1 (1908).

Thus, in the past when there has been a new technological system it has first been the case law that has reacted to the changed legal situation.

Regarding machine learning there have been already some steps towards solving this new technological issue, with the new copyright directive (DSM-directive) including an article about text and data mining.[4] However, the legislation is still far away from solving the legal issues around machine learning.

In many instances artificial intelligence (AI) and machine learning are used as synonyms but for the purpose of this paper I am making a distinction between these two. AI is a broader concept and ML is one part of it. AI as a whole is much more than a machine learning algorithm. As explained later in chapter 2 in more detail, machine learning is a part of AI and that is why I am talking about the evolution of AI together with machine learning. These two are closely linked and it is important to understand the connection between the two. Artificial intelligence has been a research field from the 1950's. There were three meetings in 1955, 1956 and 1958 that kick started the AI research field.[5] Thus, even if it is portrayed in media that AI and machine learning are current phenomenon they have been developed and researched for a long time.

The question as old as computers themselves is, could they learn by themselves. Programming computers to learn would mean that they could improve their performance automatically with experience. There is a long way still to a complete learning with computers, but there are many computer programs already that are made to learn.[6] Machine learning can be used in many different ways. One is to create art and music.

For music generating ML algorithms there are already many websites that provide consumers with easy and ready to use algorithms to generate new music. One of these is called AIVA[7]. By subscribing to the website, you will get access to generate your own music with the help of the ML algorithm AIVA. Another composing algorithm is called Iamus from Universidad de

[4] Directive 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 Directive 96/9/EC and 2001/29/EC [2019] OJ L 130/92 articles 3 and 4.
[5] Nilsson (2010), 73.
[6] Mitchell (1997) 1.
[7] www.aiva.ai

Málaga. This algorithm generates classical music and Iamus' compositions have been performed by the London Symphony Orchestra.[8]

The Next Rembrandt was a joint project with data scientists, engineers and art historians. The goal was to create an algorithm that could produce a new Rembrandt painting. To achieve this the team examined closely Rembrandt's painting techniques, style and subject matter. The artwork itself was produced with 3D printing technology. The team was successful, but the question is: Who owns the copyright of the painting? The work itself was created by a machine, but there were also investors and the people making the computer program involved. The question needed to be answered from the copyright perspective is, whether it is even possible for an algorithm generated work to fulfil the requirements laid down in copyright law.[9] The deciding factor in determining the author and possible protection of the work is the amount a human has had the possibility to impact the algorithm's decisions.

There has been a shift to digital authorship. This means that it is not automatically a human making a work. The technological development has led to a situation where it is an algorithm that makes a work autonomously.[10] This raises a lot of questions about authorship and legal personality. The aim of this research paper is not to answer questions about the authorship of ML generated works. However, the question is touched upon throughout the paper and especially when talking about copyright infringements and the person responsible for the infringement.

Copyright questions about machine learning are not the only legal questions concerning it. This technological development has raised many legal problems including General Data Protection Regulation (GDPR)[11] and bias in the training data. However, it is not possible to discuss these questions in great detail in this research paper. That said I think it is important to get an idea how ML is affecting many parts of the society and there are a lot of questions to be resolved outside the technology.

[8] Smith (2013).
[9] Schlackman (2018).
[10] Bridy (2012) 3.
[11] Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC. GDPR regulates personal data in the European Union.

1.2 The Research Problem and Question

In the DSM-directive recital 3 the fast development of technology demands legislation to be 'future-proof'. This is also the goal of the European Union (EU) so that technological development is not restricted. However, even if this is the goal it is clear that legislation cannot answer all the questions concerning copyright in the field of technology. Questions that are left unanswered are: 1) Can machine learning algorithms infringe on copyright and 2) Who is liable of the possible infringement?

The research questions in this paper is: does the exclusive right provided by copyright prohibit the use of data for ML training and can the end result of the algorithm infringe on someone's copyright? If the ML algorithm infringes on someone's copyright, who is liable for this infringement?

The discussion about copyright and AI or machine learning has been greatly focused on the authorship of machine generated works. The questions have been about originality of the generated work and possible protection of it as well as to whom the work belongs to.[12] There has not been as much discussion about copyright infringements made in machine learning processes. This is why I think it is important to steer the conversation to cases where ML can infringe on copyright and the question about liability of the infringement.

1.3 Scope and structure

The problems with machine learning are not limited to copyright. There are similar questions being raised regarding other IPRs and trade secrets. However, in this paper the focus is only going to be on copyright and machine learning. In this research paper I am focusing on those machine learning algorithms in which the end results could be protected by copyright as if it was made by a human. I am outlining all the algorithms that work with other types of problems, e.g. risk evaluations for loans or consumer buying habits.[13] Especially the focus will be on

[12] *See* Denicola (2016); Margoni (2018); Ballardini, He and Roos (2019).
[13] *See* Alpaydin (2014) s. 2-3. for more machine learning uses.

machine learning algorithms that examine popular music songs and attempt to produce music that would be popular. The purpose of these types of algorithms can be non-commercial but I would argue that there is at least a strong commercial incentive with music generating algorithms.

This research paper will mainly focus on EU law, but because of the novelty and global nature of the problem, there will be a lot of opinions on the matter from the US and UK. These three legislative systems all have a bit different approach to copyright and machine learning.

This research paper first introduces machine learning in the second chapter. The goal is to give a good overview what machine learning is and what kind of different uses it has. When there is a clear overview of the technical side it is easier to understand the problems, it brings to copyright. The third chapter explains the basics of copyright from a ML relevant view. After this in the fourth chapter I will discuss about protection of data and databases. Data is the prerequisite for functioning machine learning system, so it is important to understand what kind of data is protected as well as the protection of databases. The main topic, copyright infringements, of this paper will be discussed in chapter five. This chapter will explain firstly what copyright infringements are and in which situations ML can possibly infringe. This chapter will also discuss the question about liability in infringement situations. The final chapter before conclusions will discuss justifications for copyright protected works in ML.

1.4 Method

The methodology used in this research paper is mixed. Legal dogmatic approach is used to understand the current legislative situation and how it can answer to the challenges arisen from machine learning. The legal dogmatic approach is used to systemise and interpret the current legal situation.[14] I am also using features of law & history to research how legislation has transformed with other technological innovations such as photography. I cannot refrain from using law & technology approach because of the mere topic of this paper. Technology affects many fields in our society. On the other hand, technological advances promote social interests but on the other hand it can also lead to negative impacts if there are unexpected outcomes from

[14] Aarnio (1989) 48.

the technology. Because of the impact technology has on society laws regulating technology have an impact on society as well. The goal of law and technology theory is to understand how technological developments can affect public policy.[15] From the law & economics view I am focusing on the legal and economic uncertainty that comes from the current state of things concerning copyright and machine learning and how this can affect the development of technological innovation. Law and economics method studies the economic effects different regulatory alternatives have.[16] Because of the need for examining more than one legislative system, I will use comparative law methods as well.[17] However, the focus is on examining this topic on an international level with emphasis on European legal structure.

The source material includes many scholarly articles from the EU, US and UK. I am also referring to directives and local legislation about copyright protection and how these could be applied to machine learning. For the basics in machine learning as well as in copyright the sources consist of elementary literary works of these subjects. I have also used some internet sources to illustrate the existing machine learning algorithms in the musical field.

The topic and questions discussed in this research paper have not been realised in case law yet. Therefore, the possible answers are based on scholarly opinions and copyright case law outside the machine learning subject matter. Even though the use of copyright protected works in machine learning have not realised in EU case law yet, it is likely that they will in the near future. There are no actual provisions regulating machine learning in EU law, nor in that matter in Finnish law, or case law. Therefore, when tackling this matter there is a need to draw similarities and conclusions from other laws and case law and how they would apply to this new technology.

[15] Cockfield (2004) 386.
[16] Timonen (1998) 100.
[17] *See* Husa (2013).

2 WHAT IS MACHINE LEARNING

2.1 Machine learning as part of Artificial Intelligence

To understand what machine learning is, we must understand Artificial intelligence. AI can be described as a construction of intelligent systems, with an input and output, and their analysis.[18] AI is something that requires intelligence from a human, but it is performed by a machine. Intelligence is the capability to modify one's actions to reflect changed circumstances.[19] However, defining AI is not simple task because it is developing all the time. Autonomy and adaptivity are key terms that can explain AI. These mean that AI has the ability to perform tasks autonomously and to learn from experience.[20]

Artificial intelligence consists of six fields: machine learning, reasoning and knowledge presentation, problem-solving, perception, natural language processing, and robotics.[21] Learning is the basis of intelligence and that is why machine learning is also the base for artificial intelligence. Because ML is such a significant part of AI these terms are used as synonyms by many scholars. However, I think it is important to make a distinction between them. AI is much more than just the ability to learn.

One interesting question regarding AI is if machines can truly be intelligent. This discussion is not new and actually one of the first persons to question the intelligence of a machine was Ada Lovelace, a 19th century mathematic. She was doubtful about Charles Babbage's Analytical Engine and its potential. She described the machine as follows:

> 'The Analytical Engine has no pretensions whatever to *originate* any thing. It can do whatever we *know how to order it* to perform. It can *follow* analysis; but it has no power *anticipating* any analytical relations or truths.'[22]

[18] Hutter (2005) 2.
[19] Copeland (2000).
[20] Ch. 1(I) *Elements of AI*, < https://course.elementsofai.com> accessed 12.3.2020.
[21] *See* Copeland (2000) and Podcast by Tyler Renelle, 'Machine Learning Guide' especially episode 002 'What is AI/ML' gives a good overview of what AI and ML are and what the relation between them is, <http://ocdevel.com/mlg> accessed 7.3.2020.
[22] Lovelace (1843) 722.

The opposite view is given by Alan Turing. Alan Turing developed the 'imitation game' in 1950, which is nowadays known as the Turing test. Machine intelligence can be tested with the 'imitation game'. The goal of the game is for an interrogator to conclude from two persons which one is a man and which is a woman. To do this the interrogator can ask questions from these two persons, but they cannot see or hear each other. The man's goal is to get the interrogator to make a false identification as the woman is hoping for a right one. When the man is replaced with a computer the test starts testing the machine's intelligence. If the machine can trick the interrogator to first believe the machine is actually a person and secondly to make the interrogator to make a false identification it would seem that the machine is intelligent. Thus, if a person cannot distinguish between an answer given by a machine or a real person the machine is seen as artificially intelligent.[23]

However, Ada Lovelace's argument would still apply to the Turing test because the machine has been programmed to try to deceive and get a false identification. That said I am not attempting to answer the question about 'truly' intelligent machines, but it is an interesting question to think about.

2.2 Machine learning

Machine learning can be defined as systems that are given a task and are able to improve their performance in that task with experience or data.[24] As stated above, machine learning is a part of AI, but machine learning has its own subfields as well. These are supervised, unsupervised and re-enforcement learning. Supervised learning can still be divided into deep and shallow learning. However, these classifications are not easy to identify, and many ML methods use a combination of these.[25] There are also to other parts to machine learning. The first part is the training part which requires training data and the second part is the utilization of the machine in its indented use.

[23] A. M. Turing, 'Computing Machinery and Intelligence' (1950) Vol. 59, No. 236 Mind, 433.
[24] Ch. 1(II) *Elements of AI*, < https://course.elementsofai.com> accessed 12.3.2020.
[25] Ch. 4(I) *Elements of AI*, < https://course.elementsofai.com> accessed 13.3.2020.

Machine learning searches potential hypotheses from a vast amount of information or data.[26] It constructs computer programs that use experience in a task to improve their performance. A big part of this learning is searching for different hypotheses.[27] Learning in ML is presented in the following model:

> "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."[28]

To implement this definition of learning to music algorithms; T=the creation of a new song, P=the number of streams of the song and/or the placing on top lists, and E=comparing the similarity of the song to other high streamed songs.

The ML algorithm can use example data or past experience for learning.[29] The training of a computer program can be done in different ways. The methods used are *direct* or *indirect* training. When making a ML algorithm it is not enough to recognize the task, performance and experience. One must choose the training experience which means what kind of training – direct or indirect – is used. This leads to the problem of credit assignment. The training experience should correspond to the performance that the algorithm must perform eventually. After this it must be determined what is the exact type of knowledge to be learned and how this target knowledge is represented. The final decision is choosing a learning mechanism. There are four of these systems; the Performance System, the Critic, the Generalizer and the Experiment Generator. The Performance System functions by using a new problem and to solve the problem it uses past performances. It improves by learning from past mistakes and successes. The Critic mechanism uses the history of the tasks to produce new training examples. The Generalizer uses hypothesis based on past training tasks. The Experiment Generator generates new problems from the current hypothesis. This leads to getting the highest learning rate.[30]

Machine learning can be done in many different ways. Concept learning uses positive and negative examples in the training data to determine a solution from potential hypotheses. As

[26] Mitchell (1997) 14.
[27] Mitchell (1997) 17-18.
[28] Mitchell (1997) 2.
[29] Alpaydin (2014) 3.
[30] Mitchell (1997) 5-13.

Mitchell has expressed it concept learning is '[i]nferring a boolean-valued function from training examples of its input and output.'[31] The most used learning method is called decision tree learning. The decision tree shows learning outcomes from estimating target functions.[32] This method can be presented as below for easier understanding. The goal for the algorithm is to produce popular music. It goes through different hypothesis to learn which combinations give the desired outcome.

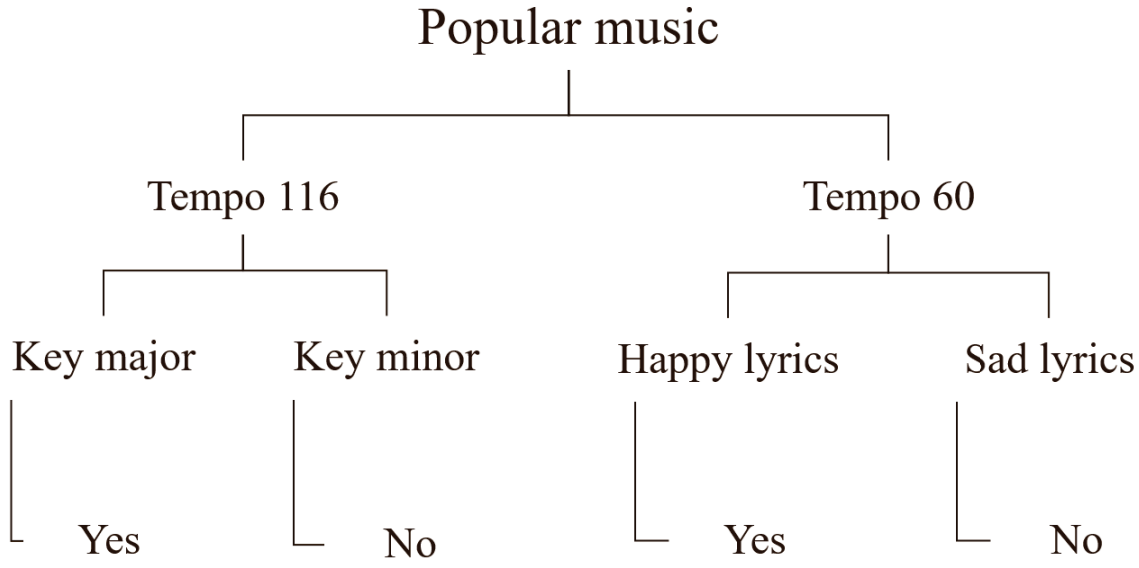


Figure 1 Example of decision tree learning.

Artificial neural networks are the third ML method. Its strength is the ability to interpret real-world data.[33] One of the uses of artificial neural networks is speech recognition.[34] Bayesian learning uses a quantitative method and probability to find optimal solutions. It counts the probability of different hypotheses by using quantitative data.[35] Instance-based learning stores the training data. When a new problem is represented the algorithm uses the stored data in its memory to solve the new problem. This can be used for example in legal reasoning in cases by

[31] Mitchell (1997) 20-21.
[32] Mitchell (1997) 52.
[33] Mitchell (1997) 81.
[34] *See* Land, Waibel and Hinton (1990).
[35] Mitchell (1997) 154.

using previous cases as references.[36] Biological evolution is used as an underlying idea in genetic algorithms.[37] The problem with all these above mentioned methods, which are inductive methods, is that they need enough data to work properly. Reinforcement learning means that when a desired action is performed a reward follows and when an unwanted action happens a penalty follows.[38] Explanation-based learning or analytical learning can be used to explain the training examples by using prior knowledge. This way the method is not solely based on the training data. It can differentiate between relevant and irrelevant features in the data.[39] However, analytical learning has its shortcomings as well as it relies on the prior knowledge so much and if the prior knowledge is not correct the algorithm is misled or if there is no prior knowledge the algorithm has nothing to work with. By combining inductive and analytical learning it is possible to get the benefits from both without the shortcomings.[40]

2.3 Training data

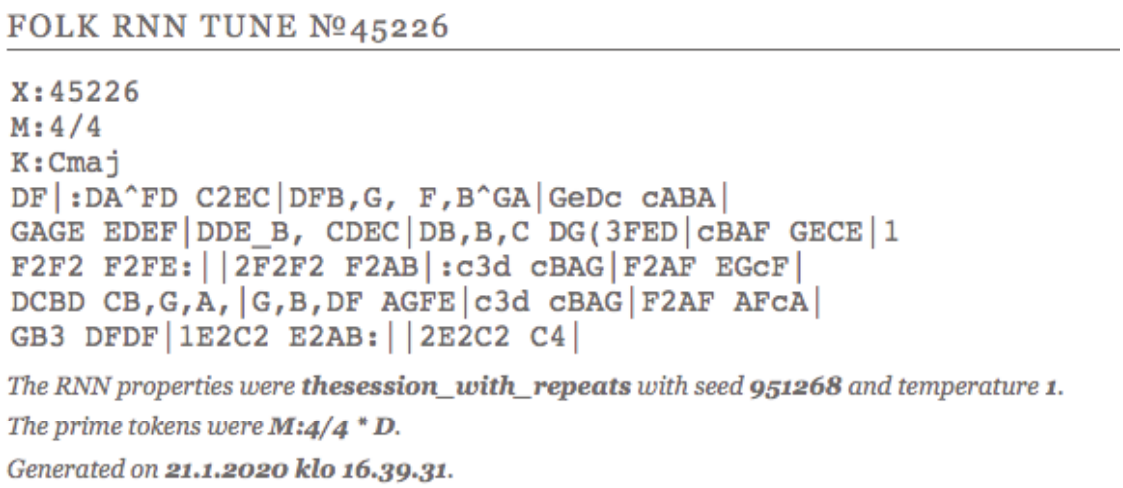
As mentioned above, to have a good algorithm it must use inductive and analytical learning which means that it needs training data as well as prior knowledge. Training data is essential for machine learning. Without it there would not be any information from which the machine could learn. For example, ML algorithms that try to mimic human authorship by making musical works need to use copyright protected works as training data. At the moment countless of copyright protected works are used as training data without authorisation from the copyright owners.[41]

It is not enough that there is a lot of data for training, the data must be of good quality as well. Imagine trying to teach a person, who's never heard music, to compose a song for a talent competition with only couple of example songs with poor quality and questionable popularity. The end result will not win the talent show. The same goes for ML training data, it must be of good quality, similar to the type of data that the end result is aspired to be and there must be ample amount of data.

[36] Mitchell (1997) 230-231.
[37] Mitchell (1997) 249.
[38] Mitchell (1997) 367.
[39] Mitchell (1997) 307-308.
[40] Mitchell (1997) 334-335.
[41] Sobel (2017) 48.

2.4 Machine Learning Algorithms Generating Music

There are already many ML algorithms that can generate music. As an example, I am using *folk-rnn.* Recurrent neural network (RNN) is used to create folk music. Anyone can access and use the algorithm from their website.[42] For the algorithm to generate the music shown in figures 1 and 2 it took it under a minute.


FOLK RNN TUNE №45226

```
X:45226
M:4/4
K:Cmaj
DF|:DA^FD C2EC|DFB,G, F,B^GA|GeDc cABA|
GAGE EDEF|DDE_B, CDEC|DB,B,C DG(3FED|cBAF GECE|1
F2F2 F2FE:||2F2F2 F2AB|:c3d cBAG|F2AF EGcF|
DCBD CB,G,A,|G,B,DF AGFE|c3d cBAG|F2AF AFcA|
GB3 DFDF|1E2C2 E2AB:||2E2C2 C4|
```

The RNN properties were ***thesession_with_repeats*** *with seed* ***951268*** *and temperature* ***1****.*
The prime tokens were ***M:4/4 * D****.*
Generated on ***21.1.2020 klo 16.39.31****.*

Figure 2: representation of the music song as the computer generates it.

[42] <folkrnn.org> accessed 21 January 2020

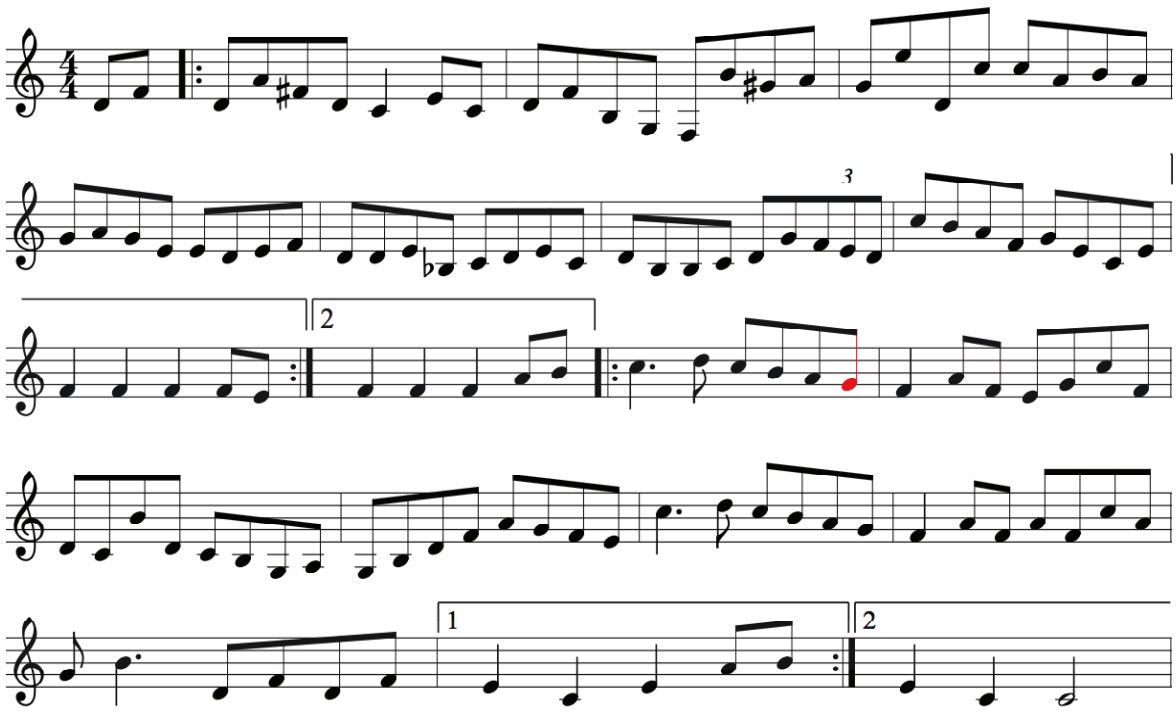


Figure 3: representation of the song in note form.

Another example is a jazz generating algorithm. Gillick, Tang and Keller created an algorithm that could generate jazz solos that are normally performed as improvised.[43] The challenge was to create an algorithm that could generate jazz music that would seem improvised. Originality within the structural and harmonic guidelines in jazz was the goal of this project.[44] The ML method used in this was grammatical inference. This means that the algorithm tries to find rules by analysing the training data fed to it. The training data could have accepted positive samples and not accepted negative samples. In the case of producing jazz melodies the algorithm learns from the data and generates new melodies which are similar to the ones fed to it.[45]

The algorithm must analyse many things from the data given to it. These include the metre, the key and the tempo of the songs analysed. If using the song generated by *folk-rnn* and seen in picture 2, the metre is 4/4 which means that there are four beats in a metre. The key is C major because there are no markings after the metre marking. Tempo is not clear from the notes, however. Tempo can be marked to the notes by a number that expresses the bpm which stands

[43] Gillick, Tang and Keller (2010).
[44] Gillick, Tang and Keller (2010) 57.
[45] Gillick, Tang and Keller (2010) 59.

for beats bet minute or it can be expressed with Italian words such as lento (slowly), allegro (happily) and presto (quickly).[46]

2.5 Text and Data Mining

Text and data mining (TDM) is used to analyse vast amounts of data. TDM algorithms are capable of identifying patterns from this data.[47] The use of a ML method for identifying these patterns from a large amount of data is called data mining. However, machine learning is more than just able to find patterns and connections from data, it has the ability to learn.[48] The DSM-directive[49] article 2(2) defines text and data mining 'any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations'.

TDM can be divided into four phases. All the phases must be done so that information can be mined from the texts and data. The below figure 4 demonstrates these four phases.

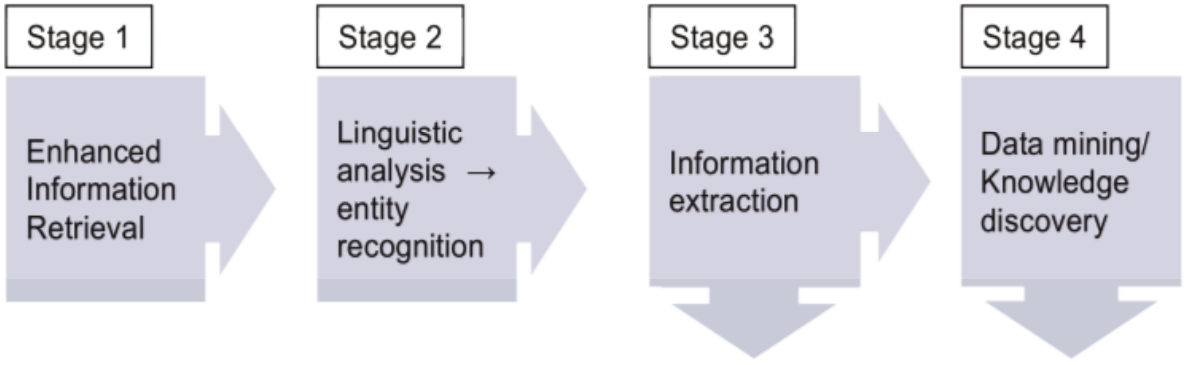


Figure 4 TDM process.[50]

The first stage is to recognize the relevant sources. This is done by sophisticated keyword searches. The second stage is to alter the found documents and their unstructured text and data

[46] *See* < http://www2.siba.fi/muste1/index.php?id=1&la=fi> accessed 7.3.2020, for detailed information about the theory of music.
[47] Carroll (2019) 902. *See also* Copyright directive 2019 recital 8.
[48] Alpaydin (2014) 2-3.
[49] Directive 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 Directive 96/9/EC and 2001/29/EC [2019] OJ L 130/92 **DSM-directive**
[50] JISC (2012) The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. Available at: <http://bit.ly/jisc-textm> 13.

to a form that a computer can obtain structured data from them. In the third stage information is extracted and identified. The fourth and final stage is where the actual mining happens. New knowledge and meaningful patterns can be found from the identified information by mining. Because of the amount of data the computer has to go through in this process, it would be impossible for humans to find the same information and patterns.[51]

TDM can use only the mere facts or data, or it can be done without reproduction of protected works, or the reproduction is only temporary. In these cases, there is no copyright infringement.[52] This will be discussed in more detail later on in this paper. With regard of machines making musical works it is important for the machine to be able to analyse the songs used in the training and identify patterns in songs that have become popular.

[51] JISC (2012) The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. Available at: <http://bit.ly/jisc-textm> 13-14.
[52] DSM-directive recital 9.

3 COPYRIGHT

3.1 Copyright system

In 1710 England enacted the first real copyright law, the Statute of Anne. This does not mean that copyright protection was invented only then. Protection of works has its roots in Greek and Roman civilizations.[53] However, there are two different approaches to copyright: the common law and civil law systems. Even though the end results in both systems are generally the same the emphasis is different. In the civil law system the emphasis is on the author and the creative factor as in the common law system the focus is on the economic factor.[54]

Authorial works are mainly protected by copyright and related rights. They give the owner an exclusive right for a limited term. Protected works vary from songs and paintings to databases and computer programs.[55] The exclusive right starts automatically from the second the work is created. The way copyright differs from other intellectual property rights (IPR) is that there is no need for applications. There are however requirements to be fulfilled for the exclusive right to exist. These requirements will be discussed in the following chapters.

Copyrights are regulated both on international level (Berne Convention and TRIPS-agreement) as well as on EU and national level. In the Berne Convention for the Protection of Literary and Artistic Works article 2 musical compositions with or without words are included in the meaning of literary and artistic works.[56] Copyright protection is harmonized in the EU with multiple directives to ensure the free movement of protected works.[57]

---

[53] Atkinson and Fitzgerald (2014) 3.
[54] Jütte (2017) 42.
[55] Pila (2016) 243.
[56] A number of other legislations refer to the Berne Convention for definition for what is covered by copyright. For example, TRIPS Agreement article 9 lays down the obligation to comply with the Berne Convention and the Finnish Copyright Act defines the subject matter in article 1 and an artistic work includes a musical work.
[57] Directive 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 Directive 96/9/EC and 2001/29/EC [2019] OJ L 130/92 **DSM-directive,** 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, [2011], OJ L 167/10 **Infosoc directive**, Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs, [1991], OJ L 122/42; republished as Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs, [2009] OJ L/16 **Software directive**, Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, [1996] OJ L 77/20 **Database directive**, and 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, [2006], OJ L 372/12; originally

Because musical works are explicitly included in national and international copyright legislation there is little uncertainty of musical works getting copyright protection when made by a human author. Music can have not only one but three types of copyright protection. Firstly, the lyrics and the melody are given copyright protection. Secondly, the recorded sounds are protected separately. And lastly, the musical arrangement can get protection.[58]

The goal for the copyright system is to promote knowledge and creativity.[59] For authors to be willing to publish their works there must be incentives. This is achieved by giving exclusive rights to the authors. The beneficiary from the copyright protected works is the public when original works are communicated to them.[60] The aim of the EU harmonization of copyright legislation with the DSM-directive is to make digital uses of copyright protected works possible when it comes to education, research and preservation of cultural heritage. This aim is seen to be achieved by having exceptions and limitations that make digital and cross-border uses possible while at the same time preserving a high level of protection of exclusive rights.[61]

The exclusive right consists of economic and moral rights given to the author. Economic rights include reproduction in many forms and making available to the public. Moral rights include the requirement to state the authors name correctly and the work must not be altered in a way that could harm the authors reputation.[62]

3.2 Originality

There is no mention of originality in international copyright treaties such as the Berne Convention. However, it is seen that the Convention includes an indirect definition of

published as Council Directive 93/98/EEC of 29 October 1993 harmonizing the term of protection of copyright and certain related rights, [1993] OJ L 290/9 **Term directive**.

[58] Laddie and others (2000) para 3.51.
[59] The Infosoc directive recital 4 sets out the goal of investing in creativity and innovation through intellectual property protection. The United States Constitution Art. I, §8, cl. 8 states that the progress of science and useful arts must be promoted.
[60] *See* case Authors Guild v. Google Inc., 804 F.3d 202, 212 (2d Cir. 2015).
[61] Commission Staff Working Document – Impact Assessment on the modernization of EU copyright rules – Part 1, 82.
[62] *See* Finnish Copyright Act section 2 and 3; and the Berne Convention article 5 and 6bis.

originality.[63] In article 2(5) of the Berne Convention there is a requirement of intellectual creativity that can be seen as a requirement for originality.

The EU originality criteria has been developed through case law. In the *Infopaq*[64] case the question before the court concerned 'reproduction in part' and originality. The reproduction in part will be discussed later in the paper, but now the focus is on the originality criteria. It was affirmed in the *Infopaq* case that for a work to be protected it must be original. The decisive factor is the author's own intellectual creativity.[65] There is a two-stage test that the Court of Justice of the European Union (CJEU) has developed to determine if a work is original and protected or not. The first step is to determine if the creation has left space for free and creative choices and the second step is how much the author used their freedom and creativity and does the work have the authors 'personal mark'.[66]

The originality requirement is not bound to absolute novelty as in the patent system novelty is one of the requirements.[67] Therefore, if two authors make an exact same musical work both of them can have copyright protection of their works. Nevertheless, these works have to be made autonomously and without knowledge of the other work.

Under the UK Copyright Act s 1(1)(a) protected works are original literary, dramatic, musical and artistic works. The originality requirement is included in the first section of the Act. The originality requirements mean that the new work cannot be copied straight from an existing work. For a work to be original it must be independent which requires skill, knowledge, mental labour, taste or judgment.[68] However, the originality requirement when it comes to copyright protection is quite moderate. Thus, with only a small amount of independent skill one can get copyright protection for one's work. The situation is different when it is a question about a copyright infringement, but that will be discussed in chapter 6.

[63] Gervais (2007) 1113.
[64] Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465.
[65] Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465, para. 45.
[66] Pila & Torremans (2016) 271-272.
[67] *See* European Patent Convention article 52 and 54(1) for patentability requirements.
[68] Laddie and others (2000) para 3.57.

The difference between the EU and UK view on originality is that the EU requires a higher standard of originality. In the EU law originality requires author's own intellectual creativity as in the UK the use of skill and labour is the decisive factor.

3.3 Balance of Rights

Many people, such as scientists and consumers, argue that access to information should be without hindrances. All obstacles on information hinder the development and knowledge from spreading.[69] Some original works need more investments from the author than others. For example, taking a photograph can be done without any substantial investments from the authors part. On the contrary composing a musical work or making a database can require a great deal of time, money and intellectual creativity from the author. Therefore, authors of these demanding works have an interest to protect their works to get compensations from them.

Giving incentives for authors can lead to situations where the original purpose of the copyright system of promoting innovation and creativity is hindered. However, the primary objective of the copyright legislation is to promote creativity in all of its forms. By giving the author exclusive rights to their work it gives an incentive to them. It is argued that this also promotes the production and dissemination of creations and innovations.[70] The DSM-directive aims to provide exceptions and limitations to the exclusive right to accomplish a fair balance between the authors and the users of the protected works.[71]

There are fundamental rights that contradict each other when it comes to the copyright system. These fundamental rights can be found in the Charter of Fundamental Rights of the European Union (CFR). The first right is the right to property. CFR article 17(2) includes intellectual property and states that it must be protected. The right most conflicting with the right to intellectual property is the right of expression and information from article 11 CFR. Also, the right to education is important because with a limitless copyright protection education would be hindered. The right to privacy must also be taken into account (article 7 CFR).

[69] Gervais (2007) 1110.
[70] The primary objective of the copyright legislation: HE 177/2002 vp.
[71] DSM-directive 2019 recital 6.

Protecting authors investments and making sure they get compensation for their works is one of the goals of EU copyright legislation. When authors are compensated fairly for their works it gives them incentives to create more and the financial compensation they are getting makes it possible for them to continue the creation of new works.[72] By harmonising the legal protection at the EU level the investments authors put in to their works is being protected and compensated accordingly.

The relationship between machine learning and copyright raises specific questions about the balance of interests. When training the machine, works that can be protected by copyright are used and the copyright owners have an interest to prohibit the use or get compensation from the use of their works. If they do not have control of the use of their works the incentive for creating new works can be hindered and this is contrary to the aim of promoting the creation of art. On the other hand if the use of data in training is made hard or impossible it hinders technological development.

3.4 Author

3.4.1 EU perspective

For a work to be desired by the public it usually has to have an author the audience know. When a work is known by the public the possibility of infringements becomes larger. People are willing to get their hands on the work even if it is not by lawful means.[73] So the more popular an author is the more likely the work is infringed on. However, if the author of a work is a computer algorithm are people that likely to infringe upon the work?

As a main rule the term author is not defined in any EU or international legal instruments. Because of the lack of definition, a common definition has been formed. According to this established definition author of a work is the person whose intellectual creation the work expresses.[74] In other words the person who made free and creative choices and the work bears this person's personal mark is seen as the author.[75] However, there are three directives with a definition of author. These can be found in the Database directive, Software directive and Term

[72] Council Directive 2006/115/EC of 12 December 2006 on Rental and Lending Rights [2006] OJ L 376, recital 5.
[73] Jütte (2017) 57-58.
[74] Pila & Torremans (2016) 292.
[75] Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465 para. 45.

directive[76]. In the Database directive author is defined as the natural person or the group of natural persons who created the database. The definition in the Software directive is the same as in the Database directive. However, it broadens the definition in article 3 so that the scope of protection is the same as for literary works. The term directive defines author for cinematographic and audiovisual works. Article 2(1) states that the author is the principal director of a cinematographic or audiovisual work.

There is an exception to an author actually having the exclusive economic right to the work. This is realized if a computer program is created by an employee. Stated in article 3 of the Software directive is that all the economic rights concerning the computer program belong to the employer if the program was created by an employee in the execution of his duties or following the instructions given by the employer if it is not otherwise provided in the employment contract. In other directives there are no mentions about copyright for employers. However, in the Database directive recital 29 the discretion about authorship concerning employee created databases is left for the Member States. In many civil law states the ownership of a work made during employment is defined in a contract between the employee and employer. This is the case in France and Finland. However, in Germany the exclusive right can never be transferred from the employee.[77]

The copyright system is tightly relying on the author. The author is the owner of the exclusive right and they have the power to decide on the use of the work. Also, the term of the protection is determined by the author's life.[78] The minimum protection time is stated in the Berne convention and it is 50 years from the authors death. The idea behind this particular term was that the work would be protected during the authors life as well as during the following two descendants' lives.[79] In the EU the Term directive regulates to protection times. However, these terms apply only to economic rights and not moral rights of the author.[80] According to the Term directive the protection time in the EU is 70 years from the death of the author. If the author is

[76] Council Directive 93/98/EEC of 29 October 1993 harmonising the term of protection of copyright and certain related rights [1993] OJ L 290/9, replaced by 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 [2006] OJ L 372/12 (codified version).
[77] Pila & Torremans (2016) 293.
[78] Aurelija Lukoseviciene, 'The contractual protection of authors in Copyright in the Digital Single Market Directive – does the reality live up to the expectations?' (2019) NIR.
[79] Term directive recital 6.
[80] Term directive recital 20.

not known the term of protection cannot be tied to the death of an author. Therefore, in cases of anonymous works the term is 70 years from the publication of the work.

3.4.2 Common law perspective

According to the United Kingdom's (UK) Copyright, Designs and Patents Act 1988 copyrightable works are only the ones listed in the Act. These are original literary, dramatic, musical and artistic works (LDMA); sound recordings, films and broadcasts; and the typographical arrangement of published editions. However, the act leaves a lot to interpretation which leads to legal uncertainty.[81] According to the same Copyright Act section 9 author is defined as a person who creates the work. It is then defined further concerning different types of works. For literary, dramatic, musical and artistic works which are computer-generated the author is the person who made the necessary arrangements to make the work happen. Thus, computer generated works are protected by copyright, and the owner or author is the person who makes the arrangements necessary for the creation of the work, i.e. the arrangements for computer generated work.

The question then follows who is the author if the arrangements do not include a person? Legislation do not have straight answers to questions with new technology and it cannot have taken into account every possible future development. This leaves a large space for interpretation. The meaning of arrangements and necessary arrangements have a central role in the interpretation. Arrangements could be seen as not including the technology or the algorithm producing the work. If the arrangements were seen as the building up the ML algorithm they are far from the arrangements actually producing the work in the end. The same algorithm can produce many different works so the coding an algorithm and the algorithm producing a work can be quite distant from each other. The link might be seen as too distant for the arrangements used to make the algorithm to be considered as the arrangements in the meaning of the Copyright Act.[82]

There is a section in the Irish Copyright and related Rights Act that defines computer-generated works. According to the section 2(1) computer-generated works are works that are generated

[81] Pila (2010) 230.
[82] Lambert (2015) 8.

by a computer in circumstances where the author of the work is not an individual. According to the Act author for a computer-generated work is a person who creates the work. This definition includes the person by whom the arrangements necessary for the creation of the work are made.

3.5 Adaptations and free associations

According to the Berne convention article 2(3) altered works are protected works and they include translations, adaptations, arrangements of music and other alterations of a literary or artistic work and they shall be protected as original works without prejudice to the copyright in the original work.

The Finnish Copyright Act section 4 distinguish between adaptations and free associations. Adaptations are translations or adaptations, or conversions to other art forms. The altered work does get copyright protection but only insofar that it does not infringe on the original work. Free associations however are new independent works and they are not dependent on previous original works.

According to the Finnish Copyright Act section 2(1) the author has the exclusive right over copies in altered form as well. Therefore, for altered works there must be authorisation from the author of the original work if it is a question about more than just correcting obvious mistakes on the original work. Alterations that need the authorisation from the author are adding, extracting or converting the style of the original work. As a principle rule, alterations that intervene with the independency and originality of the original work need authorisation.[83] An example of an alteration could be the making a new arrangement of a musical work. If a third person would want to exploit the altered work in a copyright protected manner, they would need permission for it from the author of the altered work as well as the author of the original work.[84]

Free associations obtain inspiration from the original works, but they are new original works. When using the information, principles and ideas from protected works it is allowed without

[83] Harenko, Niiranen and Tarkela (2016) 76.
[84] Harenko, Niiranen and Tarkela (2016) 77.

authorisation and the works created like this are given copyright protection the same way as the works they got their inspiration from.[85] However, it is not possible to give a straight answer what is seen as an adaptation and what is a free association. The matter must be resolved case-by-case.[86]

3.6 Rights provided by copyright and related rights

3.6.1 The reproduction right

The rights provided by copyright can be found in section 2 in the Finnish Copyright Act. The author of the work has the right to determine if and how their work is being used. They can renounce their rights entirely or give access rights.[87] The reproduction right is one of the most important rights provided by copyright.

The Berne convention article 9 defines the right to reproduction as reproducing a work 'in any manner or form'. At the EU level the basic definition of the reproduction right can be found in Information society directive (Infosoc)[88]. The Software directive[89] and Database directive[90] specify this right regarding original computer programs and databases. According to the Infosoc directive article 2 the author of the work has the exclusive right to the reproduction of the work. The right includes the authorisation and prohibition of any kind of reproduction of the work. Thus, even an indirect as well as a temporary and partly reproduction of a protected work is at the consideration of the author.

The difference between temporary and permanent is the process in which the reproduction can be destroyed. In a temporary reproduction the destroying of the work is automatic but for a permanent reproduction to be destroyed it must be done by a human explicitly.[91] As a main rule

[85] Harenko, Niiranen and Tarkela (2016) 78-9.
[86] Harenko, Niiranen and Tarkela (2016) 79.
[87] Harenko, Niiranen and Tarkela (2016) 26.
[88] 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 [2002] OJ L 167/10.
[89] Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs [1991] OJ L 122/42, replaced by Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L 111/16 (codified version).
[90] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L 77/20.
[91] Pila & Torremans (2016) 300.

both types of reproduction, temporary and permanent, are at the discretion of the owner of the exclusive right.

3.6.2 Communication to the public and making available to the public

In the Database Directive communication to the public means any communication to the public. However, in the Infosoc directive, that applies to all other protected works, the scope of communication to public is slightly narrower. The Infosoc directive provides protection to only transmissions, i.e. communicating the work so that the public can have access to it from a place and time chosen by the public. The Infosoc directive recital 23 defines communication to the public as a mean of transition 'to the public not present at the place where the communication originates'. This definition was affirmed in the case Football Association Premier League and Others.[92] According to the Berne Convention article 11 the exclusive right of authors includes the public performance of the works as well as any communication to the public.

3.6.3 Distribution right

The distribution right means the right to control the distribution of the work to the public in any form. For databases this right is regulated in its respective directive.[93] The article 5(c) states that 'any form of distribution to the public of the database or of copies thereof' is included in the database authors exclusive right. Other works are regulated in the Infosoc directive article 4(1). It states that the authors of original works have the 'right to authorise or prohibit any form of distribution' of their original works or copies of them.

The Infosoc directive article 4(2) however gives the right to prohibit distribution only until the first sale or other transfer of ownership has been made in the EU by the owner or with the owner's consent. The same exhaustion clause can be found in the Database directive. However, in the Software directive the distribution right is different when it comes to exhaustion. The right holder holds the right to prohibit distribution of further rental or copying even after the first sale of a copy.

---

[92] C-283/10 *Circul Globus Bucureşti (Circ & Variete Globus Bucureşti) v Uniunea Compozitorilor şi Muzicologilor din România – Asociaţia pentru Drepturi de Autor (UCMR-ADA)* [2011] ECR I-12031, para. 200.
[93] Database directive article 5(c) and Software directive article 4(1)(c).

The main rule is that the distribution right covers only tangible works. The right is exhausted when the work is given to circulation in the EU. This means that if a copyright owner to a CD sells the CD inside the EU the byer can distribute the CD to a third party without the copyright owner's consent. This rule covers copyrightable works that are regulated by the Infosoc directive.[94] However, the Software directive exhaustion clause does not differentiate between tangible and intangible works. Therefore, for computer programs the distribution right is not exhausted when the copy is first circulated.[95]

In the case *Nederlands Uitgeversverbond, Groep Algemene Uitgevers v Tom Kabinet*[96] the question before the court was about second-hand e-books and if it was considered communication to the public or distribution to the public under the Infosoc directive. An e-book is not a computer program and the economical and functional view of a material book and an e-book are not the same therefore it was not a question about distribution to the public but communication to the public.[97]

3.7 Related rights

Addition to the copyright protected works there are similar subject matters that are protected as related rights. Related rights protect non-authorial subject matter as copyright protects authorial and original works.[98] Related rights are harmonized on EU level and the main provisions can be found in the Infosoc directive. Other directives amend the provisions set out in the Infosoc directive such as the Term directive. The Rome convention and TRIPS Agreements have provisions concerning protection for performances, broadcasts and recordings of sounds. According to the Finnish copyright act chapter 5 related rights include performances, sound and video recordings, radio and television transmissions, catalogues and databases, photographs, and press reports.

Many works and subject matters can hold protection from copyright as well as related rights. This does not mean that they are just overlapping. It does however mean that different interests

[94] *See* Infosoc directive recital 28.
[95] C-128/11 *UsedSoft* [2012] EU:C:2012:407 para. 55.
[96] C-263/18 *Nederlands Uitgeversverbond, Groep Algemene Uitgevers* v *Tom Kabinet*, [2019] EU:C:2019:1111.
[97] C-263/18 *Nederlands Uitgeversverbond, Groep Algemene Uitgevers* v *Tom Kabinet*, [2019] EU:C:2019:1111 paras. 54, 58 and 72.
[98] Pila & Torremans (2016) 285.

are protected, and protection can belong to different persons.[99] For example, different interests are protected when it comes to a musical work. The work itself is protected as an authorial work by copyright. The recorder of the work gets protection under the related right for phonogram producers for the recording. When the work is being performed the performer is awarded the related right to the subject matter of the performance.[100]

The musical works used as teaching material for machines can be covered with not only copyright but related rights as well. These rights can belong to many different people.

3.8 Exceptions and limitations to the exclusive right

Even though copyright gives many rights to the owner it does not mean that there are no limitations or exceptions to the right. In this chapter I am introducing the exceptions to the exclusive right, but a more thorough examination of the relevant exceptions for the ML topic will be discussed later. The reason for exceptions and limitations is the need to take into account other conflicting rights and interests as discussed in chapter 3.3 about balance of rights. They make it possible to use protected works in specific situations and to accomplish public policy objectives.[101] Without any exceptions or limitations to the exclusive right it would greatly hinder peoples' possibility to express themselves and it would also hinder education and innovation.

In the EU legal framework exceptions and limitations mean that the owner of the exclusive right does not have the right to authorise or prohibit the use of a work when an exception or limitation applies. The individual or institution that benefits from the exception or limitation are called beneficiaries. There is no need for permission from the copyright owner because the law gives authority to the beneficiary[102]

The 'three step test' can be found in the article 10 and article 13 of the TRIPS Agreement. It gives the possibility to the parties of the Treaty to enact limitations or exceptions to the

[99] Pila & Torremans (2016) 288.
[100] *See* Pila & Torremans (2016) 286, performances are given copyright protection in the United States.
[101] Commission Staff Working Document – Impact Assessment on the modernization of EU copyright rules – Part 1, 80.
[102] Commission Staff Working Document – Impact Assessment on the modernization of EU copyright rules – Part 1, 80.

exclusive right provided for authors. However, this possibility only applies when three conditions are fulfilled. These conditions are that: limitations or exceptions apply to 'certain special cases', the normal exploitation of the work is not interfered, and the legitimate interests of an author is not interfered. All of the conditions have to be met for a limitation or exception to be accepted. Therefore, an author's exclusive right to a work is the main rule but to safeguard other interests it is possible to have exceptions and limitations. The same principle can be found in the Infosoc directive article 5(5).[103]

The principle was put to the test in the WTO Panel decision from 2000[104]. The EU (then European Communities) contested the US Copyright Act Section 110(5) as amended by the Fairness in Music Licensing Act. The Section 110(5) allows transmission of musical works without remuneration in certain cases. The prerequisites for this exception of the exclusive right are: the use of devices regularly used privately; the size of the place of transmission; the number of loudspeakers or displays; entry without a fee; no further transmission; and the copyright protected work transmitted is licensed. The question was if the exception in the US Copyright Act fulfilled the conditions laid down in article 13 of the TRIPS Agreement. The conditions were examined individually by the Panel. The first condition is 'certain special cases'. The limitations or exceptions in the light of the first condition should be clearly defined and the scope and reach of the limitation or exception should be narrow. However, the legitimacy of a special purpose must not be transparent in legislation even though it is beneficial in case of needing proof of the purpose of the limitation or exception.[105] Because of the high number of cases[106] where the US granted an exception for transmission of copyright protected music the Panel concluded that the exception does not comply with the first condition 'certain special cases'.[107] As the case shows the copyright protection is highly valued and it is not easy to make exceptions from it.

In the EU the protection of copyright and related rights is seen as one of the most important ways to ensure that European culture has the necessary resources and that the authors and

[103] See also Berne Convention art. 9(2).
[104] United States – Section 110(5) of US Copyright Act, Report of the Panel, WT/DS160/R (15 June 2000).
[105] United States – Section 110(5) of US Copyright Act, Report of the Panel, WT/DS160/R (15 June 2000) paragraph 6.112.
[106] *See* United States – Section 110(5) of US Copyright Act, Report of the Panel, WT/DS160/R (15 June 2000) paragraph 6.122. The D&B database in 1998 contained information that approximately 70 per cent of all drinking and eating establishments, and 45 per cent of all retail establishments are included in the Section 110(5) exception.
[107] United States – Section 110(5) of US Copyright Act, Report of the Panel, WT/DS160/R (15 June 2000) paragraph 6.133.

performers can act independently and in a dignified way.[108] Therefore, even though it is possible to have exceptions and limitations to copyright having them is not self-evident. The exclusive right has a strong standing. This can also be seen in the DSM-directive where there is a clear goal to have a fair balance between different interests, but the authors' exclusive right is still the main protectable interest.[109]

[108] Infosoc directive recital 11.
[109] DSM-directive recital 6.

4 PROTECTION OF DATA AND DATABASES

4.1 Big data

Data has become irreplaceable and has now more value than oil. The most valuable listed companies today are all in the data business. These are Amazon, Apple, Google, Facebook and Microsoft.[110] McKinsey Global Institute report on Big data estimated that 7 exabytes of new data of companies will be stored in 2010 and 6 exabytes by consumers.[111] To put this into perspective one exabyte is 1 000 000 000 gigabytes. Five exabytes is the same amount than all the words people have ever spoken.[112] There is more data in the world than words spoken so the term big data is spot on. It is clear that data is valuable, but who owns it and how is it protected? The EU does not provide explicit protection to data as such.[113] There are three dimensions in the ongoing industrial revolution. The newest dimension is data being in the focus. The two earlier dimensions have been the PC revolution and internet revolution. As the amount of data has been growing exponentially the focus has shifted to it in the revolution.[114]

Big data is a term used for big amounts of data where connections and patterns can be found.[115] However, one of the problems with Big Data is the ability to access it. A lot of data is gathered by different companies, for example social media companies. The data is valuable for them and they understandably restrict access to it from others. Thus, usually the data is completely protected, or it is shared only for remuneration or in small amounts.[116]

The growth of data has been exponential, and it has been happening globally in all business sectors.[117] In the music industry companies like Spotify and Apple music have a large amount of data of users' musical preferences. For ML algorithms discussed in this research paper, data from Spotify or Apple music would be extremely valuable. The algorithm could analyse what

---

110 <https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data> accessed 14.3.2020.
111 Manyika, Chui, et al. (2011) 15.
112 < https://www.eecis.udel.edu/~amer/Table-Kilo-Mega-Giga---YottaBytes.html> accessed 6.7.2020.
113 Günther (2019) 63.
114 Hilty (2018) 87. The Big data era has also been called the fourth industrial revolution, *see* Geiger, Frosio and Bulayenko (2018) 97.
115 Boyd and Crawford (2011) 2.
116 Boyd and Crawford (2011) 12.
117 Manyika, Chui, et al. (2011) 18.

types of songs become popular in different person groups (e.g. age groups) and then analyse these songs to learn to make its own song.

Data is defined as 'reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing'. Humans and machines can both process data.[118] However, in the age of big data it is impossible for a human to have enough capacity to go through all the relevant data and find patterns from it. This is why machines are irreplaceable because they have to capacity needed.

Big data and collection of data is not a new thing. Data has always existed, and it has been produced for a long time, however the technical advances in the 1990s and early 2000s made it possible to store the data. Data has been stored for 20-30 years already but this does not mean it has been utilized for the same amount of time. Often the data that was stored was just stored and there was actually no information and value gained from the data. However, in recent years technological advances and the drastic increase in amount of data has made it possible to actually utilize the data in many ways.[119]

Big data and analytics techniques such as machine learning and text and data mining are inseparable. Without large amounts of data these techniques do not function and give reliable results and on the other hand without these techniques the amount of data is too great for a human or even less advance machine to understand and make use of.[120]

4.2 What type of data is protected and how is it protected?

There are two types of data personal data and non-personal data.[121] In this work I am only discussing non-personal data as that is the data ML algorithms producing music are using. The ML algorithms need two types of data to function, the actual popular songs and the databases containing the ranking of the popular songs. I am first discussing the protection of data and how it can be used and then the database protection. The main problems concerning data is the

[118] ISO/IEC2382-1, revised by ISO/IEC2382:2015–Information technology–Vocabulary (2015).
[119] Emrouznejad and Charles (2018) 5.
[120] Emrouznejad and Charles (2018) 3.
[121] Pihlajarinne and Ballardini (2019) 115.

protection of it, accessibility to it and processing it. These unanswered questions might have a hindering effect on the digital single market in the EU.[122]

Data can be protected in different ways. Patents protect data in inventions and trade secrets protect data in companies.[123] Trade secrets as well as contractual mechanisms and technical protection measures are the commonly used protection means for data. The ownership, usage and access to data can and is agreed between parties by contracts. The lack of consistent regulation for non-personal data protection and the fragmented practises causes uncertainty.[124] Complications as databases are protected under copyright or sui generis -right but the data in these databases might not be protected. The ownership of data is not legislated in any Member States or on the Union level. This does not mean that the copyright legislation does not affect the ownership of data in anyway.[125]

There is no prohibition of using data as such. Thus, the use of data is allowed provided that the use does not infringe on exclusive rights, e.g. reproduction. Therefore, the pure information of a work is not protected. This means for example the length of a song.[126] Copyright protected works include data and facts, and these are not necessarily protected and therefore can be freely used.[127]

UsedSoft GmbH v. Oracle International Corp case concerned exhaustion of right to distribution for software.[128] In the case the downloading a computer program was free and the user got the right to permanently store the program on their server, however the prerequisite for the download was a licensing agreement between the software owner and the user.[129] The court stated that the authorization for downloading a copy of a software by the copyright owner constitutes a transfer of the right of ownership.[130] This transfer of the right of ownership is seen as sale in the meaning of the Software directive and with the transfer the right to distribution is

[122] Pihlajarinne and Ballardini (2019) 115-116.
[123] WIPO Conversation on Intellectual Property Policy and Artificial Intelligence (AI): Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence paragraph 32 <https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=499504> accessed 20.6.2020.
[124] Pihlajarinne and Ballardini (2019) 123-4.
[125] Pihlajarinne and Ballardini (2019) 119.
[126] Vesala and Ballardini (2019) 102-103.
[127] Sag (2020).
[128] C-128/11 *UsedSoft GmbH v. Oracle International Corp*. [2012] ECLI:EU:C:2012:407.
[129] C-128/11 *UsedSoft GmbH v. Oracle International Corp*. [2012] ECLI:EU:C:2012:407 paras. 21-23.
[130] C-128/11 *UsedSoft GmbH v. Oracle International Corp*. [2012] ECLI:EU:C:2012:407 para. 46.

exhausted. Therefore, downloading computer programs are based on licensing as well as sale of goods.[131] It was concluded that resale of a license including a downloaded software is allowed, and the subsequent acquirer can rely on the exhaustion of the distribution right. However, the license has to be granted by the rightholder originally for an unlimited period and the rightholder has gotten remuneration for the work.[132] This decision insinuates that intangible goods have specific ownership rights.[133]

Text and data mining is promoting innovation by making it possible to process large amounts of information and discovering new knowledge and patterns from this information. However, there has been uncertainty of copyright protection for the works that are subject to mining.[134] This is why the DSM-directive addresses text and data mining and has the exception for it in the new directive. The TDM exceptions can be found under title II in the directive. Therefore, the mining of data is not free from copyright legislation. It can be concluded that even if data *per se* is not protected by copyright the usage of it can infringe on copyright.

Even though data is not protected by copyright there have been discussions of the possible need to start protecting it by new IP rights.[135] As discussed in the previous chapter about big data it is clear that data has become increasingly important and holds a lot of value. Data can be generated by humans or by machines. The IPR systems are based on the assumption that works are created by human creativity and innovation. However, machines generate data in an even larger scale than humans can and the ownership of machine generated works is scattered and unclear. When discussing the possible new IP protection for data these differences between machine generated and human generated data should be taken into account.[136]

The Database directive's purpose is not to protect the information or materials presented in the database. The creation of data and the intellectual effort and skill put into it is not significant when evaluating database protection under the Database directive.[137] Also, as the case law in

[131] C-128/11 *UsedSoft GmbH v. Oracle International Corp*. [2012] ECLI:EU:C:2012:407 para. 48.
[132] C-128/11 *UsedSoft GmbH v. Oracle International Corp*. [2012] ECLI:EU:C:2012:407 para. 88.
[133] Pihlajarinne and Ballardini (2919) 119.
[134] DSM-directive recital 8.
[135] WIPO Conversation on Intellectual Property Policy and Artificial Intelligence (AI): Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence, issue 11 <https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=499504> accessed 20.6.2020.
[136] Pihlajarinne and Ballardini (2019) 118.
[137] C-604/10 *Football Dataco Ltd and others v Yahoo! UK Ltd and others* [2012] EU:C:2012:115 paras. 33-34. *See* Database directive recital 9,10 and 12 for the purpose of the directive.

the EU has shown, for copyright protection the work must be original, the author's own intellectual creation and it must have the author's personal touch or the free and creative choices made expresses the authors personality.[138] There are two main reasons why it would be hard to get copyright protection for data. The first one is that data typically is not creative enough that it would meet the conditions the case law has laid for protection. The second reason is that a lot of data is being produced by machines and the copyright system is based on the idea that authors must be human.[139]

The data in copyright protected works is important for a ML algorithm in the training process. Without the data and facts, the algorithm could not learn from the works given to it. Popular songs are somewhere between two and a half minutes to four minutes long. If the algorithm were to make a song that was 10 seconds or 20 minutes, it is likely that it would not become popular.

4.3 Database protection

To have a functioning ML algorithm only using songs for training is not enough when the goal is to make them popular. This is why there is a need for databases that contain information about the popular songs, e.g. top lists as in Spotify Top 50 or Billboard Hot 100 Chart[140]. Without the information about what kind of songs are popular in a given time it is left for the subjective view of the people choosing the training data to decide which songs are popular and which are not. The Billboard Hot 100 Chart bases their ranking of songs on their sales, airplay and streaming data weekly.[141] This information is crucial for machines so that they are able to analyse what songs are popular and for how long and are there some underlying patterns that make songs perform well on the list.

Collections are protected works under the Berne Convention Article 2(5) which states that 'collections of literary and artistic works - - by reason of the selection and arrangement of their

[138] *See* C-5/08 *Infopaq International A/S v. Danske Dagblades Forening* [2009] ECLI:EU:C:2009:465; C-429/08 *Karen Murphy v. Media Protection Services Ltd* [2011] ECLI:EU:C:2011:631; C-145/10 *Eva-Maria Painer v. Standard VerlagsGmbH and Others* [2013] ECLI:EU:C:2013:138; and C-173/11 *Football Dataco Ltd and Others v. Sportradar GmbH and Sportradar AG* [2012] ECLI:EU:C:2012:642.
[139] Pihlajarinne and Ballardini (2019) 120.
[140] Spotify Top 50 Global < https://open.spotify.com/playlist/37i9dQZEVXbMDoHDwVN2tF> and Billboard Top 100 Chart < https://www.billboard.com/charts/hot-100> accessed 23 February 2020.
[141] < https://www.billboard.com/p/billboard-charts-legend> accessed 11 March 2020.

contents, constitute intellectual creations' that are protected under copyright. There is no clear reference to database protection in the Berne Convention. However, the Convention sets out obligations on the Member States that are the minimum protection they must satisfy. It is concluded from the Berne Convention articles 2(1) and 2(5) as well as from the work made by the drafters of the articles that all intellectual selections and arrangements are protectable. This means that regardless of the IP status of the works included in the collections, the collection in itself can be protected by copyright.[142] It is thus concluded that collection of information or other non-protected substances are protected under the Berne Convention if their selections and dispositions are intellectually created.[143]

In the TRIPS agreement[144] compilations of data are explicitly protected as such in article 10(2). The article 9 of the agreement states that the members of the TRIPS agreement have to follow the Berne Convention articles 1 to 21. Therefore, the TRIPS protection of databases must follow the requirement of intellectual creation set out in the Berne Convention.[145]

On EU level databases are protected by copyright and by a *sui generis* database right in the Database directive. It was seen that the protection level for databases was not enough to protect the interests of the database owners. A Supreme Court case in the US, *Feist Publications v. Rural Telephone Service Co*[146], was the reason behind the Database directive. Databases are given specific protection because the making of a database requires a lot of work in investing in 'human, technical and financial resources'. The exploitation of databases, as in reproduction and other uses, are available easily and with considerably less effort than it took to make one.[147]

Database is defined in the directive as 'a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means.'[148] In article 3(1) it is further defined that copyright protection is given only to databases that are intellectually created by the author. The author of the database, when there is a qualitative and/or a quantitative substantial investment of the contents of the database, is given *sui generis* -right. The decisive is the investment in regards of acquiring, authentication

[142] Gervais (2007) 1112.
[143] Axhamn (2016) 61.
[144] Agreement on Trade-related Aspects of Intellectual Property Rights.
[145] Axhamn (2016) 62-3.
[146] Feist Publications v. Rural Telephone Service Co, 499 U.S. 340 (1991).
[147] Database directive recital 7.
[148] Database directive article 1(2).

or arrangement of the contents to prevent the extraction and/or re-utilisation of the whole or substantial part of the database's contents.[149] Extraction and re-utilisation are defined explicitly in the directive article 7(2). Extraction is the copying of the contents of the database permanently or temporarily as a whole or a substantial part of them. The re-utilisation concerns the communication to public in regards of the contents of a database. In the ML environment extraction is happening when databases are given as training data.

Copyright protection of databases and the *sui generis* right to them are not depended on each other. If a database does not get protection from the *sui generis* right, it can still get protection under copyright and the other way around. This can be concluded from the Database directive article 3(1) and article 7(1) as well as from the recitals of the directive.[150]

The case *Football Dataco* concerned the intellectual property rights of English and Scottish football league fixture lists.[151] Database protection requires author's own intellectual creation. This means that the work must be original. The author must have made free and creative choices of data in the database and the data must have been selected and arranged in a way that the author's creation is original.[152] By making these choices the author stamps their 'personal touch'.[153]

The case *Hit Bilanz*[154] was about a hit list for German top songs collected weekly. The BGH (der Bundesgerichtshof, German supreme court) held that a top list of songs is a protected database under the Database Directive. Thus, according to the directive and case law musical top lists are protected databases. The possible infringements in ML happen when machines are fed top list databases as training data.

[149] Database directive article 7(1).
[150] *See* C-604/10 *Football Dataco Ltd and others v Yahoo! UK Ltd and others* [2012] EU:C:2012:115 para. 27
[151] C-604/10 *Football Dataco Ltd and others v Yahoo! UK Ltd and others* [2012] EU:C:2012:115 para. 2.
[152] C-604/10 *Football Dataco Ltd and others v Yahoo! UK Ltd and others* [2012] EU:C:2012:115 paras. 37-38.
[153] C-604/10 *Football Dataco Ltd and others v Yahoo! UK Ltd and others* [2012] EU:C:2012:115 para. 38 and C-145/10 *Painer* [2011] ECR I-12533 para. 92.
[154] Case I ZR 290/02 *Hit Bilanz* BGH 21 July 2005.

5 COPYRIGHT INFRINGEMENTS

5.1 What constitutes as an infringement

According to the UK Copyright, Designs and Patent Act 1988 section 16(2) an infringement on a copyrighted work happens when a person does any of the acts restricted by the exclusive right without a licence or an authorisation from the copyright owner. Similarly, the Finnish Copyright Act section 56a defines a copyright infringement as an act of wilful or gross negligence affecting the authors economic or moral rights. The Finnish Act has made a distinction between copyright offences and copyright violations. The offences are regulated in the Criminal code but most of the copyright infringements are penalised as violations under the Copyright Act section 56a.[155]

Even though copyright is referred as an exclusive right and gives the author an exclusive right to use one's work, the more important part of copyright is the negative right. Copyright prevents others from using a protected work. It prohibits the reproduction and communication to the public from others than the copyright owner. And if someone has used a protected work in ways that is protected the person is liable to pay compensation.[156]

In the UK law there is a distinction between primary and secondary infringements. Primary infringements are acts that are restricted by copyright such as reproduction and communication to the public. Secondary infringements on the other hand are subsidiary to primary infringements and strict liability is not enforced on the secondary infringers.[157]

When determining if an act is infringing on a protected work there must be an examination of the similarity between the two works and the originality of the reproduction as well as the temporary nature of the copying. In reproduction infringements there are two types of infringing acts: copying a protected work as is and copying with altering the original work. In the former situation in the UK it has to be examined if the part copied is substantial or not. In the latter the substantiality has to be examined as well but regards to the altered parts.[158]

---

[155] Harenko, Niiranen and Tarkela (2016) 578.
[156] Laddie and others (2000) para 3.121 and *Nicol v Barranger* [1917-23] MCC 219, 231 (Lord Strendale MR).
[157] Laddie and others (2000) para 3.122.
[158] *See* Laddie and others (2000) paras 3.121-3.3.134 about copying as is and 3.135-3.139 about altered copying.

The question about infringements in machine learning is twofold: using protected works as training data and the machine-made works being similar to existing ones. As explained in chapter 2.2 about machine learning in general, machine learning is divided to the training part and the using of the finished algorithm. Infringements concerning the training of the machine are happening when the machine is programmed. On the other hand, the infringements caused by the use of the machine can be caused by different persons than the programmer. In the following chapters I am discussing the two different infringement situations as well as the liability of these infringements. For the first one the question that has to be answered is if training a ML algorithm is something that falls within the scope of the exclusive right provided by copyright regulations.[159]

The WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI): Second Session raised questions regarding possible infringement situations in AI and machine learning. The main questions raised were if using copyright protected data in machine learning training is considered an infringement and if not, should there be an explicit exception in legislation for this? If there is a need for an exception should it only cover specific cases such as non-commercial use? If the use of training data is seen as an infringement how does this affect the use of existing exception for text and data mining?[160]

5.2 The possible infringement situations in ML

5.2.1 Exploiting copyright protected works

The author of a work is granted an exclusive right over it. This right includes the right for reproduction; communication to the public and making available to the public; distribution; and rental and lending. The author of a copyright protected work has an exclusive right over the use of the work. The main rule is that all actions which the exclusive right covers are seen as infringements if not especially allowed. Reproduction is the most obvious exclusive right

[159] *See* Vesala and Ballardini (2019) 99.
[160] WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI): Second Session Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence, Issue 7 <https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf> accessed 28.3.2020.

machine learning might infringe on.[161] Therefore, I am focusing on infringement situations caused by reproduction.

The possible infringement situations when gathering training data for an ML algorithm could include reproduction of a protected work. When training an algorithm, it is clear that there will be copies made in most cases.[162] Copies are made if the learning data must be converted to a machine readable from. In the case of music generating algorithms the WAV format of the learning data is better than the usual mp3 that music is generally stored.[163] There are generally two types of reproductions made in machine learning: simple reproductions and copies and adaptive uses. These copies and adaptive uses might be simple reproductions because the programmer of the ML algorithm might not have free and creative choices over them.[164] Any reproduction can constitute an infringement under the Infosoc directive.[165] Therefore, even if it is a question about simple reproduction or copying and adaptations there is a high risk for copyright infringement when using machine learning. However, the copies of the training data are only available to the machine and they are stored electronically. The question therefore, is if this copying falls under an exception from the exclusive right. Furthermore, it is unclear if the technological process in machine learning falls under the exclusive right of copyright owners.[166]

For text and data mining the processes can involve reproducing copyright protected works. This can happen when mined works must be downloaded for the process.[167] In the DSM-directive article 4 reproduction of works for text and data mining purposes is allowed. This exception allows the copy made to be stored as long as it is necessary for TDM according to article 4(2) of the directive. Before the TDM exception reproduction without the authorisation of the copyright holder was possible only when the reproduction was temporary. The Infosoc directive article 5(1) requires the copying to be temporary but also it must be an integral and essential part of a technological process. Furthermore, the reproduction must be transient or incidental and for a lawful use with no independent economic significance. As machine learning does not

[161] Vesala and Ballardini (2019) 102.
[162] *See* Chiou (2019) 402.
[163] Coelho, Richert and Brucher (2018) 268.
[164] Chiou (2019) 404.
[165] Geiger, Frosio and Bulayenko (2018) 98.
[166] Vesala and Ballardini (2019) 101.
[167] Commission Staff Working Document – Impact Assessment on the modernization of EU copyright rules – Part 1, 104-105.

directly fall under the TDM exception the reproduction happening in machine learning should fulfil the requirements in Infosoc directive article 5. However, machine learning algorithms producing music cannot in my opinion be seen as having no independent economic significance so even solely because of this the exception for reproduction is not fulfilled if the algorithm requires the protected work to be reproduced in the learning process even if the algorithm only uses the data in the work.

5.2.2 The end result is similar to an existing protected work

5.2.2.1 Similarity

Infringement of copyright in music is not a question of note for note comparison, but of whether the substance of the original copyright work is taken or not. It falls to be determined by the ear as well as by the eye.[168]

According to the UK Copyright Design and Patents Act 1988 (CDPA) section 16(2) infringement happens when the copyright protected work is used in a way that is protected by the exclusive right and the author of the work has not given a license or authorization for the usage. Restrictions include using the work in a whole or any substantial part of it according to section 16(3)(a). Section 17 states that copying of a copyright protected work is an infringement. Reproducing a musical work in any material form is considered copying and it includes storing the work electronically. The meaning of 'any material form' is that it includes copying that is not identical to the protected work. In other words, the exclusive right covers the identical reproduction as well as non-identical.[169] Thus, for an infringement to take place there must be some level of similarity between the two works and the similarity must constitute for at least a substantial part of the infringed work.

Similarity is examined case-by-case and only when the works are non-identical.[170] With wholly identical works the infringement is fairly clear. This is why I am focusing on the examination of non-identical works and their similarities as well as the examination of what constitutes as original and the author's own intellectual creation. However, it is good to keep in mind that

[168] *Austin v Columbia Gramophone* 1917**,** at 415 – 409.
[169] Cason and Müllensiefen (2012) 26.
[170] Cason and Müllensiefen (2012) 26-27.

works that are made independently or with inspiration from another independent work, are not infringing even if they are completely similar.[171]

Similarity can be caused by five different situations: the latter work is derived from the former; the converse; the alleged infringed and alleged infringing work were derived from the same origin; pure coincident that two independent works are identical; and the contents of the works are the kind that it is inevitable for the works not to be similar or identical.[172] In the case of machine learning the possible situations would be the derivation of the former work, the derivation of the same origin or pure coincident. Pure coincident however would be quite improbable.

5.2.2.2 Substantial part and author's own intellectual creation

The question about the substantial part of work is considered as a reference to the protected work. It does not matter if the similarity is not a substantial part of the alleged infringing work but if it is a substantial part of the infringed work.[173] In the UK examination is on the substantial part of a work as in the EU the focus is on the originality and the author's own intellectual creation.

When examining the substantial part, essential is that the quality of the part copied and not the quantity.[174] In the UK copying of a substantial part happens when the whole work is copied and when it is a question about altered copying. Altered copying happens when a protected work is not copied as is, but it is copied with modifications. This was the case in *Designer Guild v Russell Williams*.[175] If a work is not 'independently originated' it constitutes an infringement. However, it is complicated to distinct between copying a substantial part and using the idea behind the work. The copying of an idea is allowed because ideas are not protected by copyright.[176] It is required to examine how a copyright work is original when it comes to altered copying.[177]

---

[171] Laddie and others (2000) para 3.125.
[172] Laddie and others (2000) para 3.127.
[173] Cason and Müllensiefen (2012) 29.
[174] *Ladbroke (Football) Ltd v. William Hill (Football) Ltd* [1964] 1 W.L.R. 273, Lord Reid p. 276.
[175] *Designer Guild Ltd v Russell Williams (Textiles) Ltd* [2001] 1 WLR 2416.
[176] *Designer Guild Ltd v Russell Williams (Textiles) Ltd* [2001] 1 WLR 2416, 1.
[177] Laddie and others (2000) para 3.139.

On EU level the case *Infopaq*[178]considered the EU law equivalent to the substantial part requirement. The case concerns the reproduction of a newspaper articles by scanning and converting them into a text file as a part of a data capture process and storing and printing the reproduction. The legal question in the case was if this reproduction is seen as an infringement or does it fall to the exception of transient acts in article 5(1) of the Infosoc directive. For the subject of this research paper the *Infopaq* case is interesting in the view of what is considered the author's own intellectual creation of the parts of a protected work and what amount of the work can be seen as original when taken out of the whole work.

The data capture process in *Infopaq* uses a search word and includes the five previous words and five following words from the search word. This amounts to eleven words together which are reproduced.[179] The question is if these eleven words can be seen as original and contribute to the originality of the whole work. The individual words themselves are not the intellectual creation of an author, but the choosing, sequencing and combining of the words can result in originality and therefore copyright protection.[180] The same goes for musical works. The individual notes are not original but the combination of them can be. In the *Infopaq* case the determination of originality was left for the national court to decide. However, no matter how small a part is from a protected work, if it expresses the intellectual creation of an author it is protected by copyright.[181] This means that the reproduction of an original part is prohibited as it is for the whole work.

The Advocate-General has defined 'reproduction' and 'reproduction in part' in their opinion. Reproduction is seen as saving a copyright protected work to an information medium and from this follows that reproduction in part is saving parts of the protected work to an information medium. It is however important to interpret the reproduction in part with a balance between technological necessities and the purpose of the copyright protection.[182] The balance between technological necessities and the purpose of the copyright protection is something that should be considered with machine learning as well. From a technical viewpoint reproduction is a

[178] C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465.
[179] C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465, para. 20.
[180] C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465, para. 45.
[181] C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465, para. 51.
[182] C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465, Opinion of AG Trstenjak, paras. 56-58.

necessity for machine learning. The question is if the purpose of copyright protection is to prohibit reproduction of training data which is visible and used by the machine and not humans?

5.2.2.2 Mashups

Mashups are new songs made by mashing up two or more existing songs or pieces of them with the help of a sound editing software. A classic mashup 'A vs. B' has the instrumentals from one song and the lyrics from another. In a way ML algorithms that learn from music and generate music after the learning process are doing mashups. However, the algorithms analyses the data in the works and learn from them and not simply mash together parts of different works.

According to Graham Reynolds most mashups in Canada are made without the copyright owners' approval.[183] I would argue the same applies globally and most mashups are made without any permissions from the copyright owner. The same will likely apply to using protected works as training data for machine learning.

According to the US Copyright Act section 114(b) 'the exclusive right of the owner of the copyright in a sound recording is limited to the right to prepare a derivative work in which the actual sounds fixed in the sound recording are rearranged, remixed, or otherwise altered in sequence or quality.' According to the same section 114(b) 'the exclusive rights of the owner of the copyright in a sound recording do not extend to the making or duplication of another sound recording that consists entirely of an independent fixation of other sounds, even though such sounds imitate or simulate those in the copyrighted sound recording.' The second part of the section refers to cover-songs. It would seem that in the US a mashup author could only use the fair use defence and not *de minimis*.[184] Both of these are defences for copyright infringements and will be discussed in more detail on chapter 6.

[183] Reynolds (2009).
[184] Lae (2011) 7-8.

5.3 Who is liable for the infringement?

5.3.1 General

To be able to determine who is liable for the infringement, the author or a person in charge of the infringing work must be identified. It is clear that it is not possible to charge the ML algorithm with copyright infringement. Is it the one that has collected the data for the training, the one who has coded the algorithm, or the one who has made changes in the algorithm? Assuming that the used training data has been obtained with the permission of the copyright owners, the humans involved in the making of the algorithm have not indented for the machine to produce an infringing work. Should they have been better at coding the machine so that it would not produce infringing songs? Would this even be possible because the examination of similarity in infringement suits lie on subjective opinions of humans? It could be that the machine, instructed not to do infringing works, does not detect similarity but a subjective opinion could.

The consensus in the scholarly field is that authorship cannot be given to machines.[185] One argument for this is that machines do not need incentives to create works. The intellectual property system has as the main purpose to have incentives for creating and inventing by granting exclusive rights.[186]

5.3.2 Author of the infringing work

5.3.2.1 Can a machine be an infringer?

When the camera and photographs were invented the copyright protection for photographs was unclear. It was argued that photographs did not require creativity from the photographer. The human input was only to enable the camera to take a photograph.[187] The authorship of photographs was contested in the US in the case *Burrow-Giles Lithographic Co. v. Sarony*.[188] The case concerned a photograph taken of Oscar Wilde by Napoleon Sarony. The Burrow-Giles

[185] *See* Grimmelmann (2016) 671.
[186] Samuelson (1986).
[187] *See* Edelman (1979) and Barron (2004) 177-208.
[188] Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884).

Lithography used lithographs of the photograph without Napoleon Sarony's permission. The court had to decide of the scope of the copyright protection. A board meaning for originality was given. Importance was given to the authors creativity and author was defined as the originator of the work.[189] The causation was also a justification for giving the exclusive right to the photographer. Without a human with an idea of a photograph there would not be a work of art.[190] In the case of a photograph the machine is only a means to create the work, but the creativity is still left to humans.

The same discussion is happening now: a machine is producing the work but what is the human input in the process? The situation differs from the one in the 19th century in the way that from a technical viewpoint arguably the goal is to produce works without any, or at least minimal, human input with autonomous machine learning algorithms. In the case of algorithms producing musical works the human input consist of the coding of the algorithm and determining the goal; here producing a popular musical work. The part that would normally require creativity from a human, making a musical work, is in a machine learning situation made by an algorithm.

One of the problems raised in the discussion of machine authorship is the copyright protection's term. Machines do not die so what the term of the protection should be?[191] It is not my objective to try to answer this question, but I think it is important to raise these questions to show how complex machine learning is from a copyright viewpoint.

According to the UK Copyright Act s 16(2) if a person authorises another without a license to commit acts that infringe upon copyright the authoriser is the infringer. Thus, if a machine infringes copyright could the person or persons who coded the machine be seen as the authorisers? However, is it authorisation if the plan was to make a musical works generating machine that would not infringe on copyright?

In scholarly discussions there have been six different possible answers to the question who the author for works made by machine learning is. These are the ML algorithm itself[192], the

[189] Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884) at. 57-58.
[190] Burrow-Giles Lithographic Co. v. Sarony, 111 U.S. 53 (1884) at. 61.
[191] *See* Chavannes (2018).
[192] *See* Bridy (2016) and Samuelson (1986) 1192-1199.

programmer of the algorithm[193], the user[194], the owner of the data used[195], joint authorship[196], or no one[197]. It is argued that even in the most complex deep learning algorithms there are human input. According to this type of arguing a machine does not create anything or do anything independently.[198]

In civil law systems there is a strong connection between human creativity and the produced work. For example, only a human can be an author according to the German 'principle of creativity' (*Schöpferprinzip)*.[199] In the common law system UK legislation gives authorship to the person who made the arrangements for the creation if it is a computer generated work.[200] In the US copyright protection is given to 'original works of authorship' according to the Copyright Act §102. The Copyright Office in the US has further defined that only works that have been created by a human being can be protected.[201] It is clear that under the current legislations a machine cannot be an author. An entirely other question is should it be, but that is outside of the scope of this paper. In my opinion if a machine cannot be an author it cannot be an infringer either. Merely because the author of an infringed work cannot demand damages from a machine, it would be an intolerable situation for rightsholders.

5.3.2.2 Human input in making and using the machine

Volitional conduct doctrine analyses which of many likely infringers should be seen as the direct infringer. However, computers cannot have volition.[202]

The programmer has the power to decide what the goal of the output is and what kind of data is given to the machine to achieve this goal. Without a programmer there would not be a

[193] Samuelson (1986) 1205-1220.
[194] Samuelson (1986) 1200-1204 and Samantha Fink Hedrick, 'I Think, Therefore I Create: Claiming Copyright in the Outputs of Algorithms' (2019) 8 NYU J Intell Prop & Ent L 324, 334-336.
[195] Samantha Fink Hedrick, 'I Think, Therefore I Create: Claiming Copyright in the Outputs of Algorithms' (2019) 8 NYU J Intell Prop & Ent L 324, 348-349.
[196] Samuelson (1986) 1221-1223.
[197] Samuelson (1986) 1223-1228 and Schönberger (2018).
[198] *See* Samantha Fink Hedrick, 'I Think, Therefore I Create: Claiming Copyright in the Outputs of Algorithms' (2019) 8 NYU J Intell Prop & Ent L 324, 332.
[199] *See* Schönberger (2018) and Urheberrechtsgesetz vom 9. September 1965 (BGBl. I S. 1273), das zuletzt durch Artikel 1 des Gesetzes vom 28. November 2018 (BGBl. I S. 2014) geändert worden ist.
[200] Copyright, Designs and Patents Act, 1988 § 9(3).
[201] U.S. COPYRIGHT OFFICE, COMPENDIUM OF U.S. COPYRIGHT OFFICE PRACTICES § 101 (3d ed. 2017), para. 306.
[202] Grimmelmann (2016) 671.

machine able to generate musical works. The programmer must make many decisions about the machine and what the goal is, therefore, the programmer has a big responsibility of the training part of the machine.[203] The programmer is the one who decides what training data is used and even thought they might not be the one who does the explicit reproduction of protected works they are the one who give the order to do it.

The user of the finished algorithm might have input on generated works. This is the case with *folk-rnn* that was introduced in chapter 2.4. The user has the ability to have input on six different factors to the generated musical work. It can be argued that the user has had an opportunity to substantially influence to generated work. The user might also have the opportunity to decide on the communication to the public.[204] However, for a user to be seen as an infringer the programmer should have left a certain amount of freedom for the user. If the programmer has restricted or limited the input the user has over the generated work the user cannot be seen as an infringer in my view. The amount of input the user has over the end result should be decisive.

The general rule in copyright is that the author of the work also owns the exclusive right to it. However, both the Software directive and the Database directive make it possible to give ownership of copyright to an employer when made by an employee.[205] Even though other directives have not explicitly given or left room for Member States to decide for ownership for employers, it is seen acceptable given many national laws do have provisions giving employers this right.[206] For example, the UK Copyright Act 11(2) gives the employer the ownership of any work made during employment.

Machine learning algorithms falls under the Software directive and therefore it is clear that if it is made under an employment contract the ownership belongs to the employer. The question is if there is an infringement when training the machine under employment who is liable for the infringement, the employee programming or the employer who has ownership?

---

[203] *See* Lehr and Ohm (2017) 657 about the human involvement in the training of a machine.
[204] *See* Samantha Fink Hedrick, 'I Think, Therefore I Create: Claiming Copyright in the Outputs of Algorithms' (2019) 8 NYU J Intell Prop & Ent L 324, 334-336.
[205] The Software directive article 2(3) explicitly states that a computer program made by an employee as part of their work duties is owned by the employer regarding copyright. The Database directive recital 29 gives the Member States the right to decide on this matter but it allows giving the right to the employer.
[206] Pila & Torremans (2016) 293.

To give no one the authorship over a machine generated work could work but from an infringement point of view I do not think that liability should not fall to anyone.

6 JUSTIFICATIONS FOR UNAUTHORISED USE

6.1 Introduction

Machine learning mimics learning the way humans learn. It is undisputed that when a human learns they make a copy of what they have learnt in their brain and this copying is allowed and is not included in the scope of copyright protection. However, when similar learning and inevitable copying happens in ML the scope of the copyright system becomes trickier.[207] The main rule is that all digital copying is forbitten without authorization.[208]

This is not the first-time copyright legislation has had to give exceptions to the reproduction right for a technological system to function. Not so long time ago web browsers were facing a similar issue. It would be impossible to use the internet without the exception in Infosoc directive article 5(1). Article 5(1) states that temporary copies that are transient or incidental and an integral and essential part of a technological process and whose sole purpose is to enable transmissions between third parties by an intermediary, or lawful use of a work with no independent economical value are allowed. This exception makes it possible to transmit network between third parties through an intermediary by allowing the technologically necessary copying.

In my view an exception for using protected works in machine learning training should be enacted or at least the legal uncertainty of the situation should be solved. Machine learning is part of our society and it will be developed even further. The legal uncertainty will hinder the development of this technology and it is clear that this is a problem that has to be solved. The problem about the use of data has been noticed on the international level.[209] However, in the following chapters I am discussing the possible justifications for unauthorized use within the current legislations.

---

[207] Margoni (2018) 2.
[208] *See* Infosoc direktive article 2.
[209] *See* WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI): Second Session Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence, Issue 7 <https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf> accessed 28.3.2020.

6.2 Dissimilarity between original work and infringing work

In a case of an infringement suit raised on the grounds of similarity and unauthorized altered work there must be an assessment of similarity. When doing the assessment, one must answer the question about what kind of similarities are relevant in copyright.[210] For example, most of the songs use the same musical notes but they are not similar because of the arrangement of the notes. If a single note was protected or even two consecutive notes, it would be impossible to create new music.

The justifications for similarity infringement suits are twofold. On the one hand one can argue that the similarity is only on parts that are not protected by copyright such as ideas. Or on the other hand the new work is original and individual on its own and should be protected as a new work separate from the original.

Sampling is a technique (generally an electronical technique) where a sample is taken from an existing phonogram and the sample is used to make a new work.[211] A recent German case concerning sampling was before the ECJ. One of the questions raised in the case was if taking a short sample of a work falls under the exclusive right of the phonogram producer. The court came to the conclusion that sampling falls under the exclusive right to reproduce and distribute regardless of the length of the sample. Therefore, even a short sample would constitute copyright infringement under the Infosoc directive and the Charter of Fundamental Rights of the European Union. However, the court noted that if the sample taken to create a new work is modified in a way that the modified and original sample are not recognizable to the ear this modified sample is not infringing the original sample's copyright.[212] In this case the sample was not modified. The defendants used a sample of approximately two seconds of the claimants' earlier work and looped it continuously in the new work.[213] Because the sample was not modified, and it was recognizable to the ear there was an infringement of the claimants' exclusive right. In this case the court did not go to lengthy considerations what then constitutes

[210] Harenko, Niiranen and Tarkela (2016) 77.
[211] C-476/17 *Pelham GmbH and Others* v *Ralf Hütter and Florian Schneider- Esleben* [2019] EU:C:2019:624, para. 35.
[212] C-476/17 *Pelham GmbH and Others* v *Ralf Hütter and Florian Schneider- Esleben* [2019] EU:C:2019:624, para. 39.
[213] C-476/17 *Pelham GmbH and Others* v *Ralf Hütter and Florian Schneider- Esleben* [2019] EU:C:2019:624, para. 16.

a modified sample to unrecognizable to the ear. However, it can be concluded that if a phonogram is modified in a way that it is not recognizable then there is no infringement.

If there is no direct or indirect link between two works, it is not an infringement as there is no reproduction in the sense of the copyright legislation. In such a case there is only a resemblance between the works.[214] In the following chapter I am discussing the case *Francis Day Hunter Ltd v Bron*,[215] and incidental similarity.

6.3 Incidental similarity

As discussed in chapter 3.5 there is a difference between adaptations and free associations. In the chapter 5.2.2 the derivative use is explained. One defence against an infringement suit is that if the machine has not used the alleged infringed work in the training data, the similarity or identicality is purely incidental and therefore it does not constitute as an infringement. Then it could be argued that the machine has originated the work independently and therefore no infringement has occurred. As explained in chapter 3.2 about originality, copyright protection does not require absolute novelty. This means that if two works are created individually and even if they are similar, there is no copyright infringement from the latter work.

In the UK case *Francis Day Hunter Ltd v Bron*,[216] the plaintiffs argued that the defendants had copied eight bars of their song 'In a Little Spanish Town' to the defendants' song 'Why'. The defendants' defense relied on unconscious similarity between the two works. The composer Peter de Angelis argued that he had not coconsciously reproduced the plaintiffs' song and if he had ever heard the song 'In a Little Spanish Town' it has been when he was young, and it has been unconsciously. The court stated that there was a degree of similarity between the two works.[217] However, this did not clearly mean that there was an infringement. The court had to consider the differences between conscious and unconscious reproducing. It was stated that in infringement situations conscious and unconscious reproduction are both considered infringements. However, for conscious or unconscious copying to happen the infringer must

[214] Laddie and others (2000) para 3.128.
[215] *Francis Day Hunter Ltd v Bron*, [1963] Ch. 587
[216] *Francis Day Hunter Ltd v Bron*, [1963] Ch. 587
[217] *Francis Day Hunter Ltd v Bron*, [1963] Ch. 587, 610.

have been familiar with the allegedly infringed work.[218] The court came to the conclusion that there was no conscious or unconscious copying on the defendants' part and dismissed the plaintiffs' case.[219]

From the case *Francis Day Hunter Ltd v Bron* it is clear that not all copying is considered an infringement even if there has been reproduction of an substantial part. This case gives strong proof that if a work was not used as training data for a machine learning algorithm and the machine generated work is similar to a previous work there has not been conscious or unconscious copying and therefore no copyright infringement.

6.5 Exceptions to the usage of copyright protected data

6.5.1 Proportionate use in the EU

As explained in chapter 4.2 the use of data *per se* is not protected by copyright and therefore an author cannot prohibit the use of data in their works. However, ML uses can still include actions that are protected by copyright. That is why there has to be a closer look to the exceptions of copyright protection.

The objective of the new Copyright directive is to have fair balance between the rightsholders and the utilisation of these works for technological innovations amongst other things.[220] Below I am discussing how this balance has been executed in the EU and how it differs from the method adopted in the US.

For ML it is crucial to be able to access data. Without training data there would not be any successful ML algorithms. We live in a world of Big Data as explained earlier in this paper. However, the access to it is not straight forward. Most of the data that is needed in music generating ML algorithms is protected by copyright. One possibility is to ask permission to use

[218] *Francis Day Hunter Ltd v Bron*, [1963] Ch. 587, 614.
[219] *Francis Day Hunter Ltd v Bron*, [1963] Ch. 587, 628.
[220] Directive 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 Directive 96/9/EC and 2001/29/EC [2019] OJ L 130/92, recital 2.

the data from each copyright owner. This however is not realistic due to the amount of data needed. Is there any other way to get access to this data?

In the Infosoc Directive article 5(1) temporary copies of protected works is allowed in some cases. The requirements for making temporary copies are that the copies must be transient or incidental, integral and essential part of a technological process and whose sole purpose is to enable a lawful use. Also, the use of the copies must be non-commercial.

In the *Infopaq*[221] case it was examined what is considered a temporary copy in the meaning of the Directive. The requirements for temporary copies without the permission of the copyright owner are cumulative which means that all of the requirements must be fulfilled for the exception to apply.[222] The copying can be seen as temporary only if the existence of the reproduction is limited to what is necessary for a technical process and the process is automated so that the copy is deleted immediately, without the need for human action, after when there is no longer a necessary purpose for the copy in the technical process.[223]

The DSM-directive made the use of Text and Data Mining (TDM) possible throughout the EU. However, the change started in the United Kingdom.[224] The UK adopted a regulation[225] in 2014 which allowed TDM research if it was for non-commercial uses. This means that contrary to the exclusive right the copyright author has over reproduction of the work, making copies for TDM purposes does not infringe on copyright. The person making the copy must have lawful access to the data and the copy must be made for computational analysis for the sole purpose of research and it must have a non-commercial purpose, according to the regulation article 3(2).

The DSM-directive article 3 allows TDM in certain situations for research organisations and cultural heritage institutions if the purpose is scientific research and the organisations and institutions have lawful access to the protected works. However, article 4 of the same directive gives the right for reproductions and extractions for all who have lawful access to protected works and it is for the purpose of TDM. The copies can be stored as long as it is necessary for

[221] C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465.
[222] C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465, para. 55.
[223] C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, EU:C:2009:465, para. 64.
[224] Carroll (2019) 896.
[225] The Copyright and Rights in Performances (Research, Education, Libraries, and Archives) Regulations 2014, SI 2014/1372.

TDM purposes. However, rightsholders can explicitly deny the use of their works if it is made in an appropriate manner.

The article 4 of the new copyright directive could be used as a defence for ML algorithms using protected works as training data. It is however unclear how the TDM exception actually applies to machine learning because machine learning techniques are not all text and data mining.

6.5.2 Fair use in the US

In the US the possibility of using copyright protected works is higher. The fair use doctrine is more lenient to the usage of protected works as the European counter parts are. The fair use doctrine can be found in section 107 of the US Copyright Act[226]. It states that regardless of the exclusive right of reproduction under fair use reproduction is not an infringement. However, there are some requirements for this. If the purpose of the reproduction is 'criticism, comment, news reporting, teaching [--], scholarship, or research' it is allowed under the fair use doctrine. Also, the commercial nature, the nature of the copied work, the portion of the work used, and the effect on the potential market value are all things taken into consideration when determining fair use. According to the Copyright Act section 106(1) the reproduction of works includes only copies and phonorecords. This means that some copies are not seen as infringing on copyright.[227] Section 101 of the Act defines copies as fixed and fixed is defined in the same section as a copy or a phonorecord which duration is longer than just temporary. However, it has been shown in case law that even a short and temporary copying is still copying in the meaning of the Act.[228] Fair use is, however, a safer argument and has succeeded in multiple cases.[229] The Google books case[230] was one of the big cases where the fair use doctrine was used.

[226] United States Copyright Act U.S.C..
[227] Carroll (2019) 922.
[228] *See* Ticketmaster Corp. V. Tickets.com, Inc., No. CV 99–7654 (C.D. Cal. 2000).
[229] *See* Ticketmaster Corp. V. Tickets.com, Inc., No. CV 99–7654 (C.D. Cal. 2000), Sega Enterprises, Ltd v. Accolade, Inc., 977 F.2d 1510 (9th Cir. 1993) and Sony Comput. Entm't, Inc. V. Connectix Corp., 203 F.3d 596 (9th Cir. 2000).
[230] Authors Guild, Inc. v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).

In the US the view has been that copyright infringements are possible to only humans. Thus, computers cannot infringe on copyright and it falls under the fair use doctrine.[231] In the US case *Sega v. Accolade*[232] it was a question about reverse engineering a video game to produce another one by a competitor. As the term reverse engineering implies it is a question about examining a technological invention piece by piece to figure out how it is made. In the *Sega v. Accolade* case the reverse engineering of a computer game, i.e. software, included copying the protected software even though the new game produced from this process had only a few similarities with the original work.[233] Because the produced new game was different from the original there was no infringement in that regard. However, the question was if the copying for the reverse engineering purposes constitutes an infringement. The court stated that reverse engineering is protected by the fair use doctrine because it is 'intermediate copying'.[234] Also, the way Sega's game was examined it was not in a way of using the game for the copyright protected content. The copying focused on the idea of the game.[235]

In the US it is concluded from the copyright doctrines that only copying that happens together with communicating the original work to the public is seen as a copyright infringement. Thus, if a protected work is copied but the original work is not expressed to the public it is not seen as an infringement. This non-expressive use is outside of the scope of the exclusive right the same way as facts and ideas included in works.[236] Could this argument be used to justify the use of copyright protected works as training data in machine learning? If a musical work would be played in a forest for no one to hear it is not communicated to the public.[237] Thus, using protected musical works as training data it is only the machine that is analysing the songs. No human is listening to the songs during the training of the machine.

The uses for the doctrine are variable and nowadays the possible transformative use, as the purpose and character, is decisive.[238] To fulfil the fair use doctrine and particularly the transformative use, it must be productive and the approach or the purpose must deviate from the original. The purpose of the doctrine is to promote innovations and creativity in society.

[231] Grimmelmann (2016) 658.
[232] Sega Enters. V. Accolade, Inc., 977 F.2d 1511 (9th Circ. 1992).
[233] Sega Enters. V. Accolade, Inc., 977 F.2d 1511 (9th Circ. 1992), at 1515-1516.
[234] Sega Enters. V. Accolade, Inc., 977 F.2d 1511 (9th Circ. 1992), at 1521-28.
[235] Grimmelmann (2016) 662.
[236] Sag (2009) 1639.
[237] *See* Grimmelmann (2016) 667.
[238] Sobel (2017) 50.

This is achieved when a derived use uses the original work as a base and adds something new to it. Therefore, these types of uses are in general accepted under the fair use doctrine.[239]

The fair use doctrine has made it possible to use many technologies and there are arguments for and against the doctrine enabling machine learning as well. One argument for accepting machine learning under the doctrine is that if machines cannot be authors they cannot engage, and value works either.[240] Transformative fair use is the justification used when machines commit actions that normally would fall under copyright protected actions. These actions include reproducing, storing and analysing protected works without the copyright owner's permission.[241]

Another example of transformative fair use was in the case Authors Guild v. Google Inc.[242] In this case the reproducing and distribution of copyright protected works fell under the fair use doctrine because information about the works distributed was about the works and not the expression of them.[243] Thus, there are two principles where reproduction by machines without authorization are allowed. These are the presumption that machines cannot utilize the expressions of the work independently, and the presumption that there is no commercial effect from the unauthorized uses.[244]

De minimis defence in US can be seen in the case *Newton* in the Ninth Circuit Court of Appeals.[245] The case concerned a copyright infringement suit from a jazz flutist James W. Newton against the Beastie Boys' song Pass the Mic.

6.6 Non-commercial uses

One of the deciding factors when examining the requirements for exceptions for copyright is the commercial factor of the use. The commercial use is in the centre of the exclusive right of

[239] Leval (1990). In the Supreme court case Campbell v. Acuff-Rose Music, Inc., 510 U.S. (1994) Leval's view on transformative use was affirmed and adopted four years later.
[240] *See* Grimmelmann (2016) 661.
[241] Sobel (2017) 52.
[242] Authors Guild v. Google Inc., 804 F.3d 202 (2d Cir. 2015).
[243] Authors Guild v. Google Inc., 804 F.3d 202, 215, 217 (2d Cir. 2015).
[244] Sobel (2017) 57.
[245] Newton v. Diamond, 388 F.3d 1189 (9th Cir. 2004).

authors. If the exclusions and limitations would give others the possibility to use protected works for commercial purposes the whole copyright system would lose its standing. Others could profit from works that they did not use their skills on and produce them.

For the Infosoc directive's reproduction exception to apply to ML training data the purpose of the ML algorithm must be non-commercial. It is likely that when the purpose of the use of the data is purely for research it would be seen as non-commercial. However, when the end result could have a commercial value it is a higher probability that the actions are seen as commercial. The same would be the case with end results that would compete with the original works.[246] This would be the case with the algorithms generating musical works. They would compete with the original protected works for popularity.

What about a case where the programmer of the machine has not had any commercial incentives and it has been purely for educational purposes, but the machine is available for others to use and a user makes a song with the machine and communicates it to the public for commercial value? Does the TDM exception only cover the training of the machine and the original purpose is what matters? In the case of having the algorithm for the public to use the question is no longer only about reproduction but about communicating to the public as well which is also covered by the author's exclusive right.[247]

6.7 Use of data

As explained in chapter 4 data itself contained in copyright protected works is not covered by the exclusive right. Machine learning uses data to learn. There is a strong argument for machine learning not infringing copyright protected works. Machine learning algorithms evolve their behaviour build on empirical data. Automatically learning to identify complicated structures and making intelligent decisions is one of the main issues in machine learning.[248]

[246] Fairhurst (2019).
[247] Infosoc directive article 3.Geiger, Frosio and Bulayenko (2018) 99: Communication to the public can be triggered in even non-commercial research uses of machine learning when the learning data must be shared with other researches for peer review.
[248] Manyika, Chui, et al. (2011) 29.

WIPO has started a conversation of the need to legislate machine learning in the copyright field with a draft issues paper and revised issues paper.[249] One of the questions raised is the use of data subsisting in copyright protected works in machine learning.[250] The phrasing of this question is problematic in some scholars' view. Matthew Sag has highlighted several problems with the phrasing which I have to agree with. The problems raised by him are the use of words 'used' and 'subsist'. The 'use' of data does not necessarily infringe on copyright and the 'use' is not happening by a human. Also, the data is not merely waiting to be removed from a copyright work. "[T]he data is derived by making *an external observation about the work*."[251] The facts in a work are not protected by copyright and therefore are freely used by anyone. Hence, the number of notes in a song or the length of a song are mere facts and not owned by the author. Machine learning algorithms can deduce relevant data from works. This would not be possible for a human to do and the data in the works is not just there for the picking.

I would argue that the mere use of data is not infringing copyright because it is not protected by it. Because machine learning uses data to learn this function is not prohibited. This basic idea in copyright that facts are not copyrightable is easy to understand when talking about conservative copyright uses. For example, a newspaper article about comparing different products and their prices is protected by copyright. However, if I use the facts in the article to learn about them and then choose a product, my use of the facts is not protected by the right. I would argue that this use of data would not raise any questions about copyright infringing. By analogy when a machine learning algorithm uses copyrighted works to extract data and learn it should not be considered a copyright infringement.

Text and data mining have an exception in the DSM-directive. Legal certainty was the main reason the EU decided that the exception was needed.[252] It was seen that the only way to have legal certainty in the copyright field with new technological measures was to have a directive

[249] WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI): Second Session Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence < https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_2_ge_20/wipo_ip_ai_2_ge_20_1.pdf> and WIPO Conversation on Intellectual Property Policy and Artificial Intelligence (AI): Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence < https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=499504> accessed 20.6.2020.
[250] WIPO Conversation on Intellectual Property Policy and Artificial Intelligence (AI): Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence Issue 8 < https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=499504> accessed 20.6.2020.
[251] Sag (2020).
[252] DSM-directive recital 8.

concerned with this. This way harmonisation across the Member States is ensured because all States have to implement the new directive in their legislation.[253] However, the TDM exception has not eliminated all the legal uncertainty the technological developments bring in question of data and copyright.

[253] Commission Staff Working Document – Impact Assessment on the modernization of EU copyright rules – Part 1, 81.

7 CONCLUSIONS

If it is made impossible to use copyright protected works as training data, it will hinder technological development as well as increase the unauthorized use of the protected works. Machine learning has been developed for decades and we are getting closer and closer to independent artificial intelligence. It is irrational to ignore the challenges these technological developments cause for the intellectual property field. One of the big challenges lies with copyright protection. The world is greatly different from the one where the original copyright system was created. However, there has been challenges with other developments such as the invention of the camera and the internet. The copyright legislation has been able to tackle these questions and now it would be a silly idea that a photograph would not be covered with copyright.

For machine learning and many other technologies to work they need data. As discussed in chapter 4 there is a lot of data and it is been produced more all the time. Data is also greatly valuable and desirable. However, it is not easy to get access to the data. Data can be protected by many ways such as patents and trade secrets. By copyright data in itself is not protected but accessing it from protected works is challenging. The data that machine learning uses is not itself protected by copyright. However, when feeding the data to the machine there are copies made from the protected works.

Even though the copyright protection differs between countries it is clear that there are open questions about the copyright scope in machine learning in all of them. ML uses might benefit from the more lenient view in the US as in the EU. However, because ML is and will be used globally it would be beneficial to have as close as possible legislative opinions on how to treat copyright protection to ML made works as well as infringements caused by ML. The data collected for ML training and the different end results from them will be cross-border.

ML can be used in many ways to make our lives easier and more efficient. The example of musical works generating algorithms was a deliberate choice for this paper. In my opinion understanding copyright protection not to mention machine learning is a complicated task and my goal was to give an understandable view of them both. There is little to no uncertainty about musical works being protected by copyright. Thus, the big questions are about combining

copyright legislation with an extremely technical and still fairly recent field of machine learning that is not yet regulated.

New technologies put the existing laws to the test and in many cases, they require new legislative measures all together. However, legislative processes do not move at the same speed as new technologies are invented. Therefore, it is necessary to try to interpret the existing law in a way that fits the technological innovations. This is the case with machine learning. Even though the technology in itself is not new the multiple uses and popularity are. In the case of machine learning teaching data, the best way to solve the question about copyright protected works used might not to be new legislation. However, there is a need to clarify the situation and have understandable guidelines for machine learning applications. As noted before, in this work the TDM exception was made to the DSM-directive to solve the legal uncertainty the technological process caused in copyright legislation.[254]

The legal uncertainty has been one of the reasons researches have given to the slow development of text and data mining in the EU.[255] For the EU to be competitive globally legislation cannot hinder the development of technologies. As machine learning technologies can be used in TDM and they use similar data the legal uncertainty regarding machine learning could affect the development in that field as well. However, machine learning can be used in many different fields and in some the innovation benefits might trump the rights of copyright owners. Making an exception to cover all machine learning applications is not the way to go to ensure balance between possible beneficiaries and copyright owners.

The question about using copyright protected works in machine learning is not the only big question around copyright in machine learning. The other side and problem is the infringement situations with machine generated works. It is unavoidable not to infringe copyright when using machine learning. However, the question about who the infringer is, is a complex question.

[254] DSM-diretive recital 8.
[255] Commission Staff Working Document – Impact Assessment on the modernization of EU copyright rules – Part 1, 104.