Copyright and Machine Learning: from human-created input to computer-generated output

Vadym Kublik
508506
Law and Information Society
University of Turku
Faculty of Law
December 2018

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin OriginalityCheck service

UNIVERSITY OF TURKU
Faculty of Law
VADYM KUBLIK:

Copyright and Machine Learning: from human-created input to computer-generated output

Master's thesis, 72 p.
Law and Information Society
December 2018

This thesis examines the legality of unauthorized reproduction of in-copyright works for the purpose of being used in Machine Learning processes. It focuses primarily on US and EU copyright systems as environments for Artificial Intelligence technological developments.

Machine Learning uses of creative works differ from traditional ones: they do not involve human readers, they do not display protected expression to the public and they analyse works to extract information not protected by copyright. Hence they raise a question of whether they actually should fall within a reach of exclusive rights of copyright holders.

In respect of the US copyright system, this study addresses the fair use doctrine that under certain conditions allows unauthorized reproduction of works. The research makes an attempt to apply the doctrine to Machine Learning uses by drawing parallels with recent case law on other technological uses of copyrighted works.

As regards the EU copyright realities, this research discusses Machine Learning uses within the scope of newly proposed copyright exception for Text and Data Mining. It firstly analyses whether exempting these uses from a copyright reach would meet the three-step test requirements. After that, it critically assesses the scopes of the exception proposed and negotiated on the EU policymaking level.

Additionally, this study discusses relations between AI-generated works and original human-created works used during the training process. It touches upon a question of possible reproduction of protected expression from original works in secondary ones and copyright-related consequences of that.

Keywords: *machine learning, artificial intelligence, copyright, fair use, functionality test, text and data mining, TDM, copyright exception, human rights*

TURUN YLIOPISTO
Oikeustieteellinen tiedekunta
VADYM KUBLIK:

Copyright and Machine Learning: from human-created input to computer-generated output

Tutkielma, 72 s.
Law and Information Society
Joulukuu 2018

Tämä tutkimus käsittelee tekijänoikeuksien alaisten teosten luvatonta kopiointia koneoppimistarkoituksiin. Tutkimus keskittyy ensisijaisesti Yhdysvaltojen ja EU:n tekijänoikeusjärjestelmiin, jotka ovat tekoälyn kehittämisen ympäristöjä.

Teosten käyttö koneoppimisympäristössä poikkeaa tavanomaisesta käytöstä. Käyttöön ei liity ihmisiä, koneet eivät julkista suojattuja ilmaisuja ja niiden tuottama sekundäärinen tieto ei ole suojattu suojattu tekijänoikeudella. Täten syntyy kysymys, tulisiko niiden kuulua tekijänoikeuslainsäädönnön piiriin.

Yhdysvaltojen tekijänoikeusjärjestelmässä on oikeudenmukaisen käytön periaate, joka tietyin edellytyksin sallii teosten kopioinnin ilman lupaa. Tutkimus pyrkii soveltamaan tätä periaatessa koneoppimiseen ja esittää vertailukohtia viimeaikaisista lakijutuista liittyen tekijänoikeuksien alaisen materiaalin teknologiseen käyttöön.

EU:ssa vallitseviin tekijänoikeuskäytäntöihin liittyen tässä tutkimuksessa käsitellään koneoppimisen uusia mahdollisuuksia, johon liittyy ehdotettu tekijänoikeuspoikkeus tekstin- ja tiedonlouhinnassa. Ensin analysoidaan, voisiko tällaisten käyttötarkoitusten vapauttaminen tekijänoikeuden piiristä täyttää kolmivaiheisen testin vaatimukset. Tämän jälkeen arvioidaan ehdotetun ja neuvotellun poikkeuksen soveltamisalaa EU:n päätöksentekoprosessin tasolla.

Lisäksi tässä tutkimuksessa käsitellään tekoälyn luoman teoksen ja alkuperäisen, ihmisen luoman teoksen, suhdetta oppimisprosessin aikana. Tutkitaan mahdollisuutta tuottaa uudelleen alkuperäisen työn suojattu sisältö koneen luomasta versiosta ja tämän tekijänoikeuteen liittyvistä seurauksista.

Avainsanat: *koneoppiminen, tekoäly, tekijänoikeus, reilu käyttö, funktionaalinen testi, tekstin ja tiedonlouhinta, TDM, tekijänoikeus poikkeus, ihmisoikeudet*

Table of Contents

Bibliography

Articles

Bowman, S.R. et al. 2016, May 12, "Generating Sentences from a Continuous Space", *arXiv 1*. Available: <www.arxiv.org/abs/1511.06349>.

Cabay, J. & Lambrecht, M. 2015, "Remix prohibited: how rigid EU copyright laws inhibit creativity", *Journal of Intellectual Property Law & Practice*, vol. 10, no. 5.

Geiger, C. et al. 2018, "The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects" (March 2, 2018). Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2018-02. Available at SSRN: <www.ssrn.com/abstract=3160586> [2018, 25.11] or <www.dx.doi.org/10.2139/ssrn.3160586>.

Grace et al. 2017, May 30, "When Will AI Exceed Human Performance? Evidence from AI Experts", *arXiv*. Available: <www.arxiv.org/abs/1705.08807>.

Grimmelmann, J. 2016, "Copyright for literate robots", *Iowa Law Review*, vol. 101, no. 2, p. 657.

Heess, N. et al. 2017, "Emergence of Locomotion Behaviours in Rich Environments", *arXiv*. Available: <www.arxiv.org/abs/1707.02286>.

Leval, P. N. 1990, "Toward a Fair Use Standard", *Harvard Law Review*, vol. 103, no. 5, pp. 1105-1136.

Levendowski, A. 2018, "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem". 93 *Wash. L. Rev.* 579 (2018), p. 6. Available at SSRN: <www.ssrn.com/abstract=3024938>.

Liu, B. et al. 2018, "Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training", *arXiv*. Available: <www.arxiv.org/abs/1804.08473>.

Matulionyte, R. 2016, "10 years for Google Books and Europeana: copyright law lessons that the EU could learn from the USA", *International Journal of Law and Information Technology*, vol. 24, no. 1, pp. 44-71.

Matzen, K. et al. 2017, “StreetStyle: Exploring world-wide clothing styles from millions of photos”, *arXiv*. Available: <www.arxiv.org/abs/1706.01869>.

Mezei, P. 2017, “De Minimis and Artistic Freedom: Sampling on the Right Track?”, *Jagiellonian University Intellectual Property Law Review*, vol. 139, no. 1/2018, pp. 56-67.

Mezei, P. 2013, “The Role of Technology and Consumers’ Needs in the Evolution of Copyright Law – From Gutenberg to the Filesharers”. Éva Jakab (Ed.): Geistiges Eigentum und Urheberrecht aus der historischen Perspektive, Lectiones Iuridicae 10, Pólay Elemér Alapítvány, Szeged, 2014, p. 71-79. Available at SSRN: <www.ssrn.com/abstract=2199352> or <www.dx.doi.org/10.2139/ssrn.2199352>.

Reese, A. 2005, “The Problems of Judging Young Technologies: A Comment on Sony, Tort Doctrines, and the Puzzle of Peer-to-Peer”, *Case Western Reserve Law Review* 55 (2005).

Sag, M. 2009, “Copyright and Copy-Reliant Technology”. *Northwestern University Law Review*, Vol. 103, 2009; The DePaul University College of Law, Technology, Law & Culture Research Series Paper No. 09-001. Available at SSRN: <www.ssrn.com/abstract=1257086>.

Samuel, A. L. 1959, “Some Studies in Machine Learning Using the Game of Checkers”, *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210-229.

Samuelson, P. 2010, “Google Book Search and the Future of Books in Cyberspace”, (2010) 94(5) *Minnesota Law Review*.

Sartor, G. et al. 2018, “The Use of Copyrighted Works by AI Systems: Art Works in the Data Mill”. Available at SSRN: <www.ssrn.com/abstract=3264742>.

Schlackman, S. 2016, “The Next Rembrandt: Who Holds the Copyright in Computer Generated Art”, *ART L. J.* (2016). Available: <alj.artrepreneur.com/the-next-rembrandt-who-holds-the-copyright-in-computer-generated-art//>.

Sobel, B.L.W. 2017, “Artificial Intelligence’s Fair Use Crisis”, *The Columbia Journal of Law & the Arts*, vol. 41, no. 1.

Thoma, M. 2016, “Creativity in Machine Learning”, *arXiv*. Available: <www.arxiv.org/abs/1601.03642>.

Yanisky-Ravid, S. and Moorhead, S. 2017, Apr 24, “Generating Rembrandt: Artificial Intelligence, Accountability and Copyright - The Human-Like Workers Are Already Here - A New Model”. *Michigan State Law Review*, Award Winning: The 2017 Visionary Article in Intellectual Property Law, Forthcoming. Available at SSRN: <www.ssrn.com/abstract=2957722> or <www.dx.doi.org/10.2139/ssrn.2957722>.

Literature

Drassinower, A. 2015, What’s Wrong with Copying? Cambridge, Massachusetts: Harvard University Press.

Farrand, B. 2014, Networks of Power in Digital Copyright Law and Policy, Routledge Ltd, London.

Geiger, C (ed.) 2015, Research Handbook on Human Rights and Intellectual Property, Edward Elgar Publishing, Incorporated, Cheltenham. Available from: ProQuest Ebook Central.

Howkins, J. 2013, The creative economy, Penguin.

Karapapa, S. & Borghi, M. 2013, Copyright and Mass Digitization, Oxford University Press, Oxford.

Litman, J. 2006, Digital copyright, 2nd ed. edn, Prometheus Books, Amherst, N.Y.

Lopez-Tarruella, A. (ed.) 2012, Google and the law, T. M.C. Asser Press, The Hague.

Mitchell, T. 1997, Machine Learning, McGraw Hill.

Murphy, K.P. 2012, Machine learning, MIT Press, Cambridge, Massachusetts.

Senftleben, M. 2004, Copyright, Limitations and the Three-Step Test: An Analysis of the Three-Step Test in International and EC Copyright Law, Kluwer Law International.

International Treaties

Council of Europe

European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14, 4 November 1950.

World Trade Organization

Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 1197 (1994).

Legislation and Regulations

EU

Charter of Fundamental Rights of the European Union, 26 October 2012, 2012/C 326/02.

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.

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

Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ("Directive on electronic commerce").

US

United States Constitution.

Official publications

Boulanger, J. et al. 2014, Assessing the economic impacts of adapting certain limitations and exceptions to copyright and related rights in the EU: analysis of specific policy options, European Commission. Available: <www.dx.doi.org/10.2780/20222>.

European Commission 2016, Commission Staff Working Document, Impact Assessment on the modernisation of EU copyright rules, 14 September 2016, SWD(2016) 301 final, Part 1/3.

Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, 14 September 2016, COM(2016) 593 final, 2016/0280 (Text with EEA relevance).

Triaille, J.et al. 2014, Study on the legal framework of text and data mining, European Commission. Available: <www.dx.doi.org/10.2780/1475>.

WTO panel decision DS160, US - s 110(5) Copyright Act.

Case law

EU case law

CJEU

C-201/13 *Johan Deckmyn v. Helena Vandersteen and others* [2014] EU:C:2014:2132.

C-203/02, *The British Horseracing Board Ltd and Others v. William Hill Organization Ltd* [2004] EU:C:2004:695.

C-360/13, *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others* [2014] EU:C:2014:1195.

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

Spain

Pedragosa v. Google Spain, S.L., Provincial Audience of Barcelona (Sec. 15), 17 Sept. 2008, WESTLAW AC 2008/1773.

Germany

Bundesgerichtshof (BGH) (German Federal Supreme Court) 29 April 2010, I ZR 69/08.

France

Editions du Seuil et autres v Google Inc et France, Paris District Court, 3rd Chamber, 2nd Section, 79 PTCJ 226, 18 December 2009 (France).

US case law

Authors Guild Inc v HathiTrust, No 11 Civ 6351 (HB), 2012 US Dist.

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

Authors Guild Inc. v. Google Inc., 954 F.Supp.2d (S.D.N.Y. 2013).

Authors Guild, et al. v. Google, Inc., 15-849.

Authors Guild, Inc. v. HathiTrust, 755 F. 3d 87 - Court of Appeals, 2nd Circuit 2014.

AV et al. v iParadigms, LLC, 544 F. Supp. 2d 473 (2008).

AV et al. v iParadigms, LLC, 562 Federal Reporter, 3d Series [2009], 630–647 (USA).

Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579, 114 S.Ct. 1164, 127 L.Ed.2d 500 (1994).

Field v Google, 412 F Supp 2d 1106 (2006).

Harper & Row, Publishers, Inc. v. Nation Enters., 471 U.S. 539, 566, 105 S.Ct. 2218, 85 L.Ed.2d 588 (1985).

Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003).

Kelly v. Arriba Soft Corp., 77 F.Supp.2d 1116 (C.D. Cal. 1999).

Leibovitz v. Paramount Pictures Corp., 137 F.3d 109, 114-15 (2d Cir.1998).

Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007).

Ringgold v. Black Entertainment Television Inc., 126 F.3d 70 (1997).

Sony Corp. of America v. Universal City Studios, Inc., 464 U.S. 417 (1984).

Sony Corp. of America v. Universal City Studios, Inc., 464 U.S. 417 (1984).

TufAmerica, Inc., v. WB Music Corp., et. al., 67 F.Supp.3d 590 (2014).

VMG Salsoul, LLC, v. Madonna Louise Ciccone, et. al., 824 F.3d 871 (2016).

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

Online sources

Amper Music website. Available: <www.ampermusic.com/>.

Angwin, J. 2016, May 23, - last update, *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks*. Available: <www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Brooks, R. 2017, Oct 24, - last update, *AI vs Machine Learning (part 3 of series)*. Available: <www.guavus.com/clarifying-ai-vs-machine-learning-part-3-series/>.

BVerfG press Release 2016, May 31, No. 29/2016. *The use of samples for artistic purposes may justify an interference with copyrights and related rights*. Available: <www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2016/bvg16-029.html>.

Chavez, S. , *10 Google tools investigative reporters can use to find information*. Available: <www.ijnet.org/en/story/10-google-tools-investigative-reporters-can-use-find-information>.

Cheng, S. 2016, Dec 7, - last update, *An algorithm rejected an Asian man's passport photo for having "closed eyes"*. Available: <www.qz.com/857122/an-algorithm-rejected-an-asian-mans-passport-photo-for-having-closed-eyes/>.

CREATE. *Statement by EPIP Academics to Members of the European Parliament in advance of the Plenary Vote on the Copyright Directive on 12 September 2018*.

Available: <www.create.ac.uk/wp-content/uploads/2018/09/Statement-by-EPIP-Academics.pdf>.

Elsevier. *These Elsevier collaborations use machine learning to turn data into knowledge*. Available: <www.elsevier.com/connect/these-elsevier-collaborations-use-machine-learning-to-turn-data-into-knowledge>.

European Parliament, *Amendments adopted by the European Parliament on 12 September 2018 on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market* (COM(2016)0593 – C8-0383/2016 – 2016/0280(COD)) [Homepage of European Parliament]. Available: <www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2018-0337+0+DOC+XML+V0//EN#def_1_1>.

Facebook. *Data Policy. How is this information shared?* Available: <www.facebook.com/about/privacy>.

Facebook. *Terms of services*: <www.facebook.com/terms.php>.

Flickr. *Flickr's terms of use*. Available: <www.flickr.com/help/terms>.

FutureTDM Project. Available: <www.futuretdm.eu/news/about-futuretdm/>

FutureTDM Project. *TDM Methods.* Available: <www.futuretdm.eu/method-list/>.

Google. *Ngram Viewer*, available: <books.google.com/ngrams>.

Google. *View web pages cached in Google Search Results*. Available: <www.support.google.com/websearch/answer/1687222?hl=en>.

Google. *What does the Ngram Viewer do?* available: <books.google.com/ngrams/info>.

Google. *What you'll see when you search on Google Books*. Available: <books.google.com/googlebooks/library/screenshots.html>.

Google's official blog 2006, June 28, *Germany and the Google Books Library Project.* Available: <www.googleblog.blogspot.fi/2006/06/germany-and-google-books-library.html>.

HathiTrust digital library website. Available: <www.hathitrust.org/>.

Jukedeck. *Fuelling creativity using musical AI.* Available: <www.jukedeck.com/>.

Julia Reda 2018, Oct 25, *Second round of trilogue negotiations. The latest compromise proposal.* Available: <www.juliareda.eu/wp-content/uploads/2018/10/Copyright-Directive_4-column-document_ARTICLES-v2-23102018.pdf>.

Kantrowitz, A. 2016, May 5, *Google Is Feeding Romance Novels To Its Artificial Intelligence Engine To Make Its Products More Conversational*, BUZZFEED, available: <www.buzzfeed.com/alexkantrowitz/googles-artificial-intelligence-engine-reads-romance-novels>.

Lea, R. 2016, *Google Swallows 11,000 Novels to Improve AI's Conversation*, The Guardian, available: <www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation>.

LIBER. *An open letter sent to the Licences for Europe organisers, signed by Nobel prize winners, technology SMEs, research councils, university associations, learned academies, publishers, libraries and law academics*. Available: <www.web.archive.org/web/20130903133142/www.libereurope.eu/sites/default/files/Extract%20from%20email%20sent%20to%20L4E%20TDM%20chairs%20140313_0.pdf>.

LIBER. *Europe Needs A Broad & Mandatory TDM Exception.* Available: <www.libereurope.eu/blog/2018/11/13/europe-needs-a-broad-mandatory-tdm-exception/>.

Margoni, T. 2018, April 25, - last update, *The Text and Data Mining exception in the Proposal for a Directive on Copyright in the Digital Single Market: Why it is not what EU copyright law needs.* Available: <www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/>.

Metz, C. 2017, May 15, - last update, *Google's AI invents sounds humans have never heard before*. Available: <www.wired.com/2017/05/google-uses-ai-create-1000s-new-musical-instruments/>.

Moody, G. 2017, *The right to read is the right to mine*. Available: <www.copybuzz.com/editorial/right-read-right-mine/>.

Moses, L. 2017, Sep 14, - last update, *The Washington Post's robot reporter has published 850 articles in the past year*. Available: <www.digiday.com/media/washington-posts-robot-reporter-published-500-articles-last-year/>.

Murray-Rust, P. 2012, *The Right to Read Is the Right to Mine.* Available: <blog.okfn.org/2012/06/01/the-right-to-read-is-the-right-to-mine/>.

Newitz, A. 2016, Sep 6, - last update, *Movie written by algorithm turns out to be hilarious and intense*. Available: <www.arstechnica.com/gaming/2016/06/an-ai-wrote-this-movie-and-its-strangely-moving/>.

Picker, R. C. 2015, "Internet Giants: The Law and Economics of Media Platforms", The University of Chicago, Coursera online subject, lecture notes, viewed Jan 2018. Available: <www.coursera.org/learn/internetgiants>.

Poetry Foundation website: <www.poetryfoundation.org/>.

PoetrySoup website: <www.poetrysoup.com/>.

Prisma Photo Editor. Available: <www.prisma-ai.com/>.

Quoteinvestigator 2011, Nov 5 - last update. *Computers Are Useless. They Can Only Give You Answers*. Available: <www.quoteinvestigator.com/2011/11/05/computers-useless/>.

Senftleben, M., *EU Copyright Reform and Startups – Shedding Light on Potential Threats in the Political Black Box.* Available: <www.innovatorsact.eu/wp-content/uploads/2017/03/Issues-Paper-Copyright-Directive-2.pdf>.

The Next Rembrandt. Available: <www.nextrembrandt.com/>.

Turnitin. *Legal FAQ* page: <www.turnitin.com/en_us/about-us/privacy#terms>.

Turnitin. *The Leader in Preventing Plagiarism.* Available: <www.turnitin.com/en_us/what-we-offer/feedback-studio>.

Turnitin. *User Agreement.* Available at the registration page: <www.turnitin.com/newuser_join.asp?svr=316&session-id=e6fe8036dd13e223b12c5eecb120b9d0&lang=en_us&r=14.8462595622565>.

Vorlage des Bundesgerichtshofs an den Europäischen Gerichtshofs zur Zulässigkeit des Tonträger-Samplings. Available: <juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=pm&Datum=2017&Sort=3&nr=78496&pos=1&anz=87>.

What is Bayou: <info.askbayou.com/>.

Wikipedia. *Sampling (music)*. Available: <en.wikipedia.org/wiki/Sampling_(music)>.

World Economic Forum 2018, *Creative disruption: the impact of emerging technologies on the creative economy.* Cologny/Geneva. Available: <www3.weforum.org/docs/39655_CREATIVE-DISRUPTION.pdf >.

List of Abbreviations

AI	Artificial Intelligence
CJEU	Court of Justice of the European Union
ECHR	European Convention on Human Rights
ECtHR	European Court of Human Rights
EP	European Parliament
EU	European Union
GAN	Generative Adversarial Networks
GBS	Google Books Settlement
GDPR	General Data Protection Regulation
InfoSoc Directive	Information Society Directive
ML	Machine Learning
OCR	Optical Character Recognition
PPP	Public-Private Partnership
SME	Small and Medium-sized Enterprises
STM	Scientific, Technical and Medical
TDM	Text and Data Mining
TRIPS	Trade-Related Aspects of Intellectual Property Rights
UK	United Kingdom
US	United States
WTO	World Trade Organization

1. INTRODUCTION

The modern era of technological development is often referred to as the Fourth Industrial Revolution and Artificial Intelligence (hereinafter AI) plays a crucial role in shaping the future of our society. However, the current legal system might not be ready to accommodate new technologies. In particular, this thesis is focusing on the application of AI and Machine Learning (hereinafter ML) in the field of creative economy, where all players have to follow often very strict rules called Copyright.

In fact, the present narrative is nothing new to a copyright-educated reader. The history knows many examples of new emerging technologies not fitting into traditional uses of protected works. For example, player pianos and Sony's Betamax VCR to name but two[1]. The reaction of copyright law and rightholders was typically predictable: "*they usually tried to force them back into the shadows"*[2]. And the story of AI vs Copyright is not promising a different scenario - rightholders would try their best to extend their rights also to new uses of their intellectual property.

The peculiar thing about ML uses of copyrighted works is that it raises a question of whether they are uses of works at all within its traditional understanding. Works are being copied, but without intention to be read. At least not by humans: computer readership of human authorship[3]. Works are used, but not as works. They are used rather as something else - carriers of data[4].

Under Copyright, a work is usually perceived as a communicative act and not as a thing. This theory especially makes sense in a digital age, when making one extra copy can be done at zero costs with only one click. To copy a work under the Copyright is thus to

[1] *See* US cases *White-Smith Music Publishing Co. v. Apollo Co*., 209 US 1 (1908) and *Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417 (1984).
[2] Mezei, P. 2013, "The Role of Technology and Consumers' Needs in the Evolution of Copyright Law – From Gutenberg to the Filesharers". Éva Jakab (Ed.): Geistiges Eigentum und Urheberrecht aus der historischen Perspektive, Lectiones Juridicae 10, Pólay Elemér Alapítvány, Szeged, 2014, p. 71-79. Available at SSRN: <www.ssrn.com/abstract=2199352> [2018, 24.11] or <www.dx.doi.org/10.2139/ssrn.2199352> [2018, 24.11]. P. 71.
[3] Grimmelmann, J. 2016, "Copyright for literate robots", *Iowa Law Review,* vol. 101, no. 2, pp. 657.
[4] Karapapa, S. & Borghi, M. 2013, Copyright and Mass Digitization, Oxford University Press, Oxford. P. 15.

recommunicate it. Therefore "*uses of the work as a mere pattern of ink in the absence of recommunication*" should not be regarded as "*uses of the work as a work*"[5].

Things are getting even more complicated when such ML use of copyrighted materials results into the creation of a new expressive work. Should authors of original works used as training data also have rights in those new creations? Is there any chance of AI-plagiarism? Can the subsequent work amount to a derivative of originals on which AI model was trained? These and other related questions are addressed in this research.

This study is mostly based on the legal dogmatic method of research. It focuses on a current state of the US Copyright law with its fair use doctrine developed through a range of court decisions. Similarly, it examines the EU Copyright law with its copyright exceptions prescribed by the InfoSoc Directive[6].

In addition, not only current law itself but also lawmaking process and related circumstances are addressed here. In particular, the European Commission (hereinafter Commission) official studies and impact assessment documentation are taken as an important source of information about the background of the ongoing copyright reform. It helps to view the TDM copyright exception through the prism of the main objectives of the proposal.

The thesis employs the law and economics approach to argue that ML uses do not harm economic interests of rightholders and thus must be allowed. Similarly, from the law and sociology standpoint, this work stresses on the importance of AI technologies for the future of humankind. Therefore, the EU copyright must adapt to facilitate its development and safeguard the EU competitiveness in the AI research on the international arena.

The current state of law does not explicitly regulate ML uses of works that are analysed here. Therefore answering the main research question is more about *lex ferenda* - what the law should be or how existing law should apply to future cases. That still requires evaluating what the law is now - *lex lata*. With that in mind, this study discusses a present legal system with an aim to foresee future developments in copyright law.

[5] Drassinower, A. 2015, What's Wrong with Copying? Cambridge, Massachusetts: Harvard University Press. P. 87.
[6] 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.

The fact that this study is targeting US and EU Copyright regimes does not imply a direct objective to compare them and identify which is more favourable for developing AI. It is a well-known fact that the US fair use doctrine is more flexible towards new technological uses. However, this discussion is not aiming to advocate an adoption of an open copyright clause in the EU or to discuss its benefits. It is rather attempting to assess copyright realities on two continents that are viewed as leaders in global economy.

In regard to limitations, this thesis encompasses specifically ML uses of copyrighted works that would lead to a subsequent generation of new expressive materials. Naturally, there are no major issues with authorized uses and the study is aiming to examine copying without a prior permission. In terms of the creative outcome, i.e. AI-generated work, it is only a relation to original works and copyright holders' rights that are addressed here. The study avoids detailed discussion of the copyrightability of such creative output despite its evident importance.

Essentially, this thesis is an attempt to analyse a current state of copyright law in respect of ML uses of protected works and to conceptualise what it might be in a foreseeable future. In order to achieve this objective, the study goes through three main stages. Following this introductory part the second chapter introduces a reader to fundamentals of Machine Learning and Artificial Intelligence. It then discusses application of AI and ML in the field of creative economy: its benefits and challenges. An examination of copyright-related acts involved in ML uses of works closes the chapter.

The third chapter reviews available court decisions in cases of technological uses of copyrighted works. It is built in such a way that each kind of use e.g. web caching, image thumbnailing and library digitization is discussed in parallel within US and EU jurisdictions. The chapter starts with a short introduction to the US doctrine of fair use and ends with summary comments on the main cases. The case law discussed here is not new and had been sufficiently scrutinised by scholars. Therefore a reader, familiar with adaptation of the fair use doctrine to new technological uses in the US and EU courts struggling in this matters, may proceed directly to the next chapter.

Fourth and fifth chapters may be viewed as a core part of this thesis. They primarily discuss applicability of current legal system to the uses in question. Each chapter includes two main divisions assessing uses on the input and then on the output stages. The fourth chapter is dealing with the US copyright realities and its fair use doctrine.

The fifth chapter discusses the EU copyright law accordingly. It is mainly focusing on the recently proposed copyright exception for Text and Data Mining (hereinafter TDM) uses. It was necessary firstly to establish connection between ML and AI on the one hand and TDM on another. The chapter then follows with a broad discussion of the scopes of the proposal including some criticism and suggestions. The whole thesis traditionally ends with discussions and conclusions.

2. AI MEETS COPYRIGHT

"Computers are useless. They can only give you answers." (by Pablo Picasso)[7]*. Is true no more.*

2.1. Basics of Artificial Intelligence and Machine Learning

AI itself is a broad term and encompasses various applications most of which are irrelevant to this study. Nonetheless, it is necessary to understand the basics of an essential part to AI - the process it is based on - Machine Learning. Analysing activities involved in ML will help to better comprehend the challenges these new technologies pose for lawmakers today.

The most popular understanding of ML takes its origin from 1959 and was informally defined by Arthur Samuel as a technique that gives computer systems ability to "learn" and progressively improve from experience without being explicitly programmed[8]. He came up with this definition in a course of his work on a checkers program, which would learn to play checkers game by playing against itself. In other words, ML enables computers to "learn" from input data practically autonomously, without any intervention from the human side.

More modern and formal definition of ML was developed by Tom M. Mitchell and according to him "*a computer program is set to learn from an experience E with respect to some task T and some performance measure P if its performance on T as measured by P improves with experience E*"[9]. It does not require special computer science background to recognize the enormous potential of ML technologies capable to improve its performance on some particular task operating autonomously. It is especially true in regard to programs and applications that cannot be programmed by hand in a traditional way and it is especially true in the era of Big Data.

[7] Quoteinvestigator 2011, Nov 5, - last update. *Computers Are Useless. They Can Only Give You Answers.* Available: <www.quoteinvestigator.com/2011/11/05/computers-useless/> [2018, 24.11].
[8] Samuel, A. L. 1959, "Some Studies in Machine Learning Using the Game of Checkers", *IBM Journal of Research and Development,* vol. 3, no. 3, pp. 210-229.
[9] Mitchell, T. 1997, Machine Learning*,* McGraw Hill. P. 2.

There are three main types of ML that determine how the process of "learning" is actually happening: supervised, unsupervised and reinforcement learning[10].

Supervised learning is also sometimes called predictive. It requires that training dataset includes labelled input-output pairs. For example, an image depicting a human face would be labelled as the one that includes human face and images without a human face would be labelled accordingly. Then training AI model will need to learn a mapping from inputs to outputs to be able to determine whether there is a human face on a random unlabelled picture. This approach is for instance used in face detection (recognition), handwriting recognition technologies etc.

Unsupervised learning is sometimes called descriptive. In this case, an AI model is only provided with inputs and then is asked to find "*interesting patterns (structure)*" in the data. Therefore it is sometimes referred to as *knowledge discovery.* For example, pictures would be uploaded without telling any identifying information and then the program will by itself manipulate the data and produce uncovered structures. It can be grouping by human face presence/absence or any other factor that never could even be envisioned by the programmer. Unsupervised learning is particularly useful in discovering new knowledge that could not be foreseen in advance and with a huge load of data that would be difficult or even unrealistic to label. Moreover, unlabelled data is cheaper to acquire and usually contains more information[11].

Reinforcement learning is a less used type of ML and was inspired by behaviourist psychology. It is also associated with the trial and mistake method. This is particularly useful for robotic AI to teach it how to act or behave when given occasional reward or punishment signals[12]. The practical example of reinforcement learning application from 2017 is Google's AI DeepMind "*training several simulated bodies on a diverse set of challenging terrains and obstacles, using a simple reward function based on forward progress*"[13]. It proved to be an effective way to teach an AI to move through a number of different and challenging obstacles.

The common principle in ML stipulates that the more data used to train the AI model the better quality the result will have. The same is true about the quality of the training

[10] Murphy, K.P. 2012, Machine learning*,* MIT Press, Cambridge, Massachusetts. P. 2.
[11] *Ibid,* p. 10.
[12] *Ibid,* p. 2.
[13] Heess, N. et al. 2017, "Emergence of Locomotion Behaviours in Rich Environments", *arXiv*. Available: <www.arxiv.org/abs/1707.02286> [2018, 24.11].

data used. The role of developers is to set the right algorithms; however, it is the quality and quantity of training data that determines the final outcome.

Terms "ML" and "AI" are sometimes used interchangeably although it would be technically incorrect to equate them. The latter one is much broader and includes the former one. Without going into details, it should be enough to say that "*AI is a branch of analytics that goes beyond machine learning, providing the system with the ability to reason*"[14]. In other words, while ML capabilities are limited by what was used in training dataset, AI can go further from this point by making hypotheses and trying to understand new information. This thesis will refer to these two terms rather equally since they carry a similar meaning in the copyright context.

Overall, ML processes are designed to be conducted without human interventions and only with limited control. That is its greatest strength and also a big challenge since an AI that lacks self-awareness is unable to explain its output and it makes it difficult for developers or other interested parties to analyse rationale behind the result. Instead, it depends largely on the input data and applicable algorithms[15].

This shall not, however, be understood that developers do not have any idea about what their ML programs are doing. A developer using a pen, lots of paper, calculator and the same input data could arrive at the same results as machines do. However, comparing with a machine, that would take much longer time for the human. In other words, it is indeed possible for developers to analyse the rationale behind an AI-generated result, but that would be a very time-intensive and difficult task[16].

2.2. AI and creative economy

Approaching the main questions of this paper is well to start mentioning that the ML has been being widely applied in the field of the creative economy recently. For the purpose of this analysis creative economy is a part of the economy "*where value is based on novel imaginative qualities rather than the traditional resources of land,*

[14] Brooks, R. 2017, Oct 24, - last update*, AI vs Machine Learning (part 3 of series)*. Available: <www.guavus.com/clarifying-ai-vs-machine-learning-part-3-series/> [2018, 24.11].
[15] World Economic Forum 2018. *Creative disruption: the impact of emerging technologies on the creative economy*, Cologny/Geneva, p. 8. Available: <www3.weforum.org/docs/39655_CREATIVE-DISRUPTION.pdf> [2018, 24.11].
[16] Thoma, M. 2016*, "*Creativity in Machine Learning", *arXiv*. Available: <www.arxiv.org/abs/1601.03642> [2018, 24.11]. P. 1.

labour and capital"[17]. This covers creative activities involving music, film, literature, fashion, paintings, architecture, design etc. Copyright plays a central role in governing these endeavours and has to change constantly in response to new technological challenges.

It is no more surprising to read news about ML creating new original content throughout multiple industries. For instance, *Washington Post* developed an AI called *Heliograf* to write articles to cover sports and political news[18]; at Google, AI invented sounds that humans have never heard before[19]; a movie was made out of script written by an algorithm[20]. Completely new business models were established that operate solely on AI platforms, such as *Amper*[21] and *Jukedeck*[22] services offering low-cost high-quality AI-created music that is in demand among bloggers, game developers etc.

Furthermore, according to some researchers, AI will be able to write high-school essays by 2026, compose top 40 pop-songs by 2027 and write *New York Times* bestselling books by 2049[23]. This is clearly an evidence of how ML is disrupting creative industries nowadays on a large scale.

One of the benefits of the new technological solutions in respect of creative economy appears to be in transforming the value chain of the creative content production. There is no more a need to hire a professional translator with the availability of high-quality and, what is more important, free translation service provided by Google. Similarly, in some cases, it is more efficient to use *Jukedeck* services instead of paying to a professional musician. Removing some intermediary stakeholders from the value chain makes content creation more affordable and opens new possibilities for small and medium creators. It is a rational measure applied within the incentive of cost minimization and benefit maximization that is so essential for any business activity.

---

[17] Howkins, J. 2013, The creative economy, Penguin. Chap 1.
[18] Moses, L. 2017, Sep 14, - last update, *The Washington Post's robot reporter has published 850 articles in the past year*. Available: <www.digiday.com/media/washington-posts-robot-reporter-published-500-articles-last-year/> [2018, 24.11].
[19] Metz, C. 2017, May 15, - last update, *Google's AI invents sounds humans have never heard before*. Available: <www.wired.com/2017/05/google-uses-ai-create-1000s-new-musical-instruments/> [2018, 24.11].
[20] Newitz, A. 2016, Sep 6, - last update, *Movie written by algorithm turns out to be hilarious and intense*. Available: <www.arstechnica.com/gaming/2016/06/an-ai-wrote-this-movie-and-its-strangely-moving/> [2018, 24.11].
[21] Amper Music website. Available: <www.ampermusic.com/> [2018, 24.11].
[22] Jukedeck. *Fuelling creativity using musical AI*. Available: <www.jukedeck.com/> [2018, 24.11].
[23] Grace et al. 2017, May 30, "When Will AI Exceed Human Performance? Evidence from AI Experts", *arXiv*. Available: <www.arxiv.org/abs/1705.08807> [2018, 24.11].

Another and probably more important value of ML is in potential discoveries that could be achieved via technological analysis of data. For example, there is a big number of publications containing valuable information from various fields of knowledge: biology, medicine, chemistry, physics etc. ML technology allows processing and analysing far more information than humans could ever do by simple reading.

Moreover, it allows creating an AI model specialising in some particular field, e.g. medicine or physics, to help human make better practical decisions in their everyday professional activities, e.g. treating patients, discovering new drugs, creating new materials with improved characteristics etc. However, most of the relevant information that could be used as training datasets is copyright protected and authorization is normally required.

As it follows, there are currently two main constraints regarding the application of AI in the creative industry. First is a lack of access to relevant high-quality data to conduct research and to make important discoveries. In order to avoid any potential copyright liability, researchers have to play safely in regard to the source of training data. It forces them to use materials from the public domain, Creative Commons-licensed works or other legally low-risk content. However, the quality of such materials is not always appropriate to what ML is aiming to achieve.

It must be remembered, that the quality and quantity of training data will determine the final AI output. One of the shortcomings associated with using "cheap" training data is a danger to create an AI system that will be biased in its outcomes. There are many cases of claiming some technological applications being racist or sexist. For example, an Asian student could not renew his passport online because the algorithm identified his eyes as being closed on the picture[24]. Some tested facial recognition technologies, trying to predict the risk of criminality, will tend to point at black people in disproportionally more cases[25].

[24] Cheng, S. 2016, Dec 7, - last update*, An algorithm rejected an Asian man's passport photo for having "closed eyes"*. Available: <www.qz.com/857122/an-algorithm-rejected-an-asian-mans-passport-photo-for-having-closed-eyes/> [2018, 24.11].
[25] Angwin, J. 2016, May 23, - last update*, Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks*. Available: <www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> [2018, 24.11].

Amanda Levendowski refers to the biased AI problem as "garbage in, garbage out"[26] emphasising on the importance of qualitative data for training socially just AI. She argues that the Copyright system locks the most valuable content for the purpose of ML and that copyright exceptions or fair use must be put in place to solve the bias problem and create fairer AI.

Benjamin Sobel, in contrast, assigns only limited role for copyright in making low-bias training data hard to access[27]. He claims first, that even copyright protected materials can bear biases and second, that genuinely high-quality datasets are "*inaccessible not because of copyright law, but because of secrecy*"[28]. Therefore, unveiling access to copyrighted data in the former case will not solve the problem of the biased AI and it is not copyright, but some other legal tools that lock valuable training data in the latter case.

Certainly, it seems reasonable to claim that copyright protected works are of a larger value for ML purposes and their use as training data will to some extent improve preconditions for creating fairer AI. It can be explained by their greater relevance which stems from being comparatively new and based on the latest more credible information. Therefore, it is important to assess whether contemporary copyright law is ready to meet the needs of these new uses of protected works or whether some legislative interventions would be required.

2.3. AI and copyrighted works

Regardless of copyright liability risks associated with processing creative materials in ML operations, in practice, very few developers would choose a pace of clearing rights for the data they need. There may be several reasons for such behaviour: insufficient level of legal awareness among computer engineers, especially in respect of intellectual property rights; lack of knowledge about licensing possibilities or simply an absence of such licensing tools in regard to some specific kind of content etc. Unsurprisingly, the most common justification for using unauthorised materials in training datasets seems to be an unwillingness to pay for something that could be easily acquired by means of

[26] Levendowski, A. 2018, "How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem". 93 *Wash. L. Rev*. 579 (2018), p. 6. Available at SSRN: <www.ssrn.com/abstract=3024938> [2018, 24.11].
[27] Sobel, B.L.W. 2017, "Artificial Intelligence's Fair Use Crisis", *The Columbia Journal of Law & the Arts,* vol. 41, no. 1, p. 47.
[28] *Ibid.*, p. 48.

internet download and genuine belief, that the fair use will justify such acts (at least under the jurisdiction of the US copyright system).

The problem with clearing rights and getting authorization from copyright holders is naturally in impracticability of doing this considering the number of individually taken pieces of original creations. It would be not only tremendously difficult to identify and contact every rightholder of, let's say fifty thousand, pictures used in training an AI model, but also unreasonably expensive to provide some monetary reward for everyone involved.

Transaction costs of such an undertaking might be too high compared to the objectives of the project, especially in the situation when the value of each separately taken work for the final outcome is comparatively low. The value of the training data is in its volume rather than in each separate work. Therefore, it will only make sense to acquire desirable content *en masse* rather than per unit.

It is important to realise, that nowadays it is possible to find an institution ready to license out copyrighted materials for the purpose of computational analysis. Most likely it is only possible with respect to texts and images and under some strictly limited conditions. However, such offerings are so limited in number that it would be too soon to call it a formed market of copyrighted materials for ML purposes. Yet, the existence of this kind of market has important meaning in the debate whether copyright should tolerate the kind of unauthorised use discussed here.

The reproduction right is often viewed as a core of copyright law. All other exclusive rights, shall it be distribution or making available to the public, involve this initial step - making a copy. In ML processes, reproduction can happen at many different stages and for the purpose of this research they will be classified into three main groups: a) copying of the materials to be used as input in the datasets; b) multiple copyings that happen in the "black box" during the actual ML processes; and c) potential reproduction when AI produces outcome similar to the work used in the training dataset.

At the first stage, copying is an important step to prepare materials and include them in the training dataset. It may require digitization of paper literary works by means of scanning and optical character recognition (OCR) to make them readable by a computer. In fact, lots of materials nowadays are created and readily accessible in

digital format, e.g. electronic books, and yet reproduction is required to insert them in the dataset.

At the same time, it is important to realise, that initial copying is not always essential for ML. An alternative way to learn from content is to use web "crawlers" that would analyse web content without actually copying them. In that case, there is no reproduction as such and thus no copyright infringement. Yet even this method of "learning" may still result in reproduction in the outcome.

The second group of reproduction acts is less problematic due to its intermediate nature and, according to Sobel, *"the spirit of copyright ... seems to exempt this type of copying*". In fact, intermediary copying in normal situations would be allowed, provided that the initial use is lawful. On the other side, if the initial copying of works is not fair use or not exempted under the EU copyright law, these intermediate copying "*... would be unlikely to further prejudice the defendant's case*"[29].

Finally, it is important to discuss the possibility of an AI model producing an infringing outcome. Firstly, the very purpose of developing an AI model may be a deliberate goal to create an application able to mimic some famous author of the original style. For example, well known *Prisma* photo editor mobile application "*transforms your photos and videos into works of art using the styles of famous artists: Van Gogh, Picasso, Levitan, as well as world famous ornaments and patterns*"[30].

While the works of aforementioned artists are likely to rest in the public domain, there are dozens of other styles imitating famous creators, for example, the style *Mononoke* which is based on Hayao Miyazaki's *Princess Mononoke* - Japanese animated fantasy film. It is unknown whether Prisma Labs, Inc. obtained a license for using the original expressions in question; however, it can be argued that every picture edited with this style is likely to amount to some kind of adaptation of original expressions.

Another remarkable example of ML use creating in style is a project conducted by Microsoft in collaboration with the bank ING in 2016 under the name *The Next Rembrandt.* In that project, an entire collection of Rembrandt's works was examined

[29] *Ibid.*, p. 17.
[30] *Prisma Photo Editor*. Available: <www.prisma-ai.com/> [2018, 24.11].

employing deep learning algorithms and face recognition technologies to "*distill the artistic DNA*" and use it to create a new work[31].

The new painting was brought to life with help of latest 3D printers and was a different work, however mimicking original artistic style. Many commentators question who owns the copyright in the new production[32], while it is no less important whether this new portrait shall be regarded as a derivative work and would require authorisation from Rembrandt if he was still alive. The answer may be not that obvious considering the fact that the new work is different from all pre-existing ones. It is not a copy of original expression, but rather a copy of artistic style, which is not protected by copyright or any other law. At least it is not protected yet.

Secondly, even if an AI model was not intentionally designed to imitate original works, it may happen to do so to the level amounting to copyright infringement[33]. Normally, it would be an undesirable outcome signifying problems with learning process or just improbable incident. For example, an AI model was trained on thousands of books to learn how to write new original novels, but at some point, it reconstructs whole sentences from training data. Since it is very difficult to predict the outcome, the risk of getting AI-plagiarist still exists.

Furthermore, even more complex infringement may happen when AI produces an outcome similar to work that was not used in training data. It is hypothetically possible when the input itself was based on other works. The probability of such indirect copying is higher than it seems to be due to the cumulativeness of creativity[34]. The principle "on the shoulders of giants" is equally relevant in this situation for humans and creative AI. However, arguably humans can be more careful in choosing expression in order to avoid copying.

Unlike with intermediate copying, the legal implications of reproduction in the outcome do not entirely depend on the lawfulness of initial use of copyrighted works in a training

---

[31] *The Next Rembrandt*. Available: <www.nextrembrandt.com/> [2018, 24.11].

[32] *See* e.g. Yanisky-Ravid, S. and Moorhead, S. 2017, Apr 24, "Generating Rembrandt: Artificial Intelligence, Accountability and Copyright - The Human-Like Workers Are Already Here - A New Model". *Michigan State Law Review*, Award Winning: The 2017 Visionary Article in Intellectual Property Law, Forthcoming. Available at SSRN: <www.ssrn.com/abstract=2957722> [2018, 24.11] or <www.dx.doi.org/10.2139/ssrn.2957722> [2018, 24.11]; and Schlackman, S. 2016, "The Next Rembrandt: Who Holds the Copyright in Computer Generated Art", *ART L. J.* Available: <alj.artlawjournal.com/the-next-rembrandt-who-holds-the-copyright-in-computer-generated-art> [2018, 24.11].

[33] Sobel, 2017, p. 18.

[34] *Ibid*., p. 20.

dataset. In particular, owners or users of a creative AI risk facing liability even in the case when an AI was trained on authorised materials or when such original use was lawful for other reasons. This is an unfortunate consequence of unintentional copying that might take place in the case with AI designed to create original content.

3. TECHNOLOGICAL USAGE OF COPYRIGHTED WORKS

3.1. General characteristics

With the emergence of computer technologies more and more aspects of people's lives have been moving from physical to digital world. This trend periodically gets accelerated with disruptive innovations like it was the case with the emergence of the internet, Wi-Fi, 2-3-4-5G etc. There hardly can be named a field of human activities that avoided the touch of a "digital hand", and intellectual property in general and copyright, in particular, are far from those. In fact, a majority of intellectual creations nowadays take place in the digital world, and works from the pre-digital era are being successfully transferred from analogue to computer-readable format, i.e. digitised.

The peculiar thing about digital versions of works is that they have to be reproduced every time they are processed by computer even when making a new copy is not a primary goal of such processing. There are nowadays many examples of technological tools that are based on the processing of copyrighted works without asking permission from appropriate rightholders. Technically, those uses constitute unauthorised reproduction of copyrighted works but, depending on the character, purpose or other criteria, some of them either exempted from copyright liability in the EU[35] or deemed to be a fair use under the US copyright law.

The ML use of copyrighted works shares many characteristics with other types of technological uses. Different scholars call these uses differently. For example, professor Matthew Sag refers to them as "*copy-reliant technologies*" meaning "*technologies that copy expressive works for non-expressive ends*"[36]. This study will now discuss specific attributes of those uses to outline their scopes and move to the analysis of their judicial assessment in the US and Europe likewise.

The first important feature of technological uses of copyrighted works is a high volume of transactions or, as some refer to it, a *bulk use*. Works have to be used on a large scale

[35] *See* for example temporary reproduction exception under the Art. 5(1) of the InfoSoc Directive.
[36] Sag, M. 2009, "Copyright and Copy-Reliant Technology" (April 9, 2009). *Northwestern University Law Review*, Vol. 103, 2009; The DePaul University College of Law, Technology, Law & Culture Research Series Paper No. 09-001. Available at SSRN: <www.ssrn.com/abstract=1257086> [2018, 25.11], p. 3.

to ensure the proper functioning of the process. For example, Internet search engines rely a lot on Web Cache technologies that make web page copies in a view to provide access to them when original websites do not function properly or make it possible to track changes on the web pages. At the same time, websites often contain original content protected by copyright. Hence, web caching *ipso facto* constitutes a massive copying of copyright-protected content without prior authorisation from respecting website owners.

Furthermore, works are being reproduced not only on a large scale but also entirely and verbatim. For example, plagiarism detection systems like *Turnitin* have to copy works completely word by word otherwise it would not be possible to effectively identify similarities between works that are being compared. This is very important from the copyright perspective since one of the fair use factors checks the amount of used work to determine whether a use can be considered fair.

The example with web caching technologies is also great for another reason: there is a technical mechanism available for website owners to avoid being cached. In other words, copyright holders are provided with a possibility to *opt out.* However, it is contrary to an *opt-in* principle that copyright reproduction is based on.

The author of this thesis is of opinion, that it is reasonable in a world of mass digitization to introduce an opt-out system for dealing with copyright protected works. Firstly, when certain bulk use qualifies with fair use criteria like it is the case with web caching in the US, it will be also fair towards rightholders to equip them with a right to withdraw from such engagement with their works. Secondly, in the EU copyright realities, provided that web caching is covered with a statutory exception, the opt-out mechanism would pay respect to authors' moral rights enabling them to avoid any technical use that is in conflict with their feelings.

Overall, rightholders should be better off with a right to avoid copying, given that such copying is lawful. However, in many instances, the lawfulness of a use may not be that clear. In addition, fair use is a defence based doctrine, meaning that fairness of the use can only be established in the court in each separate case. Furthermore, "*copyright is a system of ex ante permissions*" and, as some commentators put it, the introduction of the opt-out regime is "*turning copyright on its head*"[37].

[37] Karapapa & Borghi, 2013, p. 2.

Nevertheless, the availability of the opt-out mechanism should have no effect on the lawfulness of use. To put it differently, leaving rightholders with a right to opt out shall neither make the use lawful nor diminish any possible liability. In order to avoid a copyright infringement, a user must either obtain prior permission or solely rely on copyright exceptions or fair use. The possibility to opt out is rather a credit of respect to author's right of integrity and shall have no independent copyright meaning in the legality assessment.

Probably the most important feature of technological uses is that they do not aim to demonstrate any expressive elements of the works to the public. This is a strong argument weighing in favour of exempting these uses from copyright infringement since they do not encroach on what is considered to be an exclusive monopoly of authors - a privilege to communicate their expressiveness to the public.

It can be argued, however, that Internet search engines by providing access to cached web pages make them available to everyone upon request. They communicate information that otherwise could not be accessible, for instance, if the original source was deleted. That tends to look like a clear copyright infringement case, yet it is not prohibited due to the very different purpose such cached pages serve. More discussion will follow later in a corresponding section.

3.2. Technological usage case study

Before starting a case-by-case analysis of technological uses of copyrighted works it would be helpful to provide some introduction to the US doctrine of the fair use to make a further discussion more comprehensive. The doctrine is codified in the §107 of the US Copyright Act and is used as a defence tool in cases with unauthorised copying. Courts have to consider four main factors to determine whether the copying is non-infringing, namely:

1) *the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes;*

2) *the nature of the copyrighted work;*

3) *the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and*

4) *the effect of the use upon the potential market for or value of the copyrighted work.*[38]

It must be remembered, those four factors are not cumulative and non-exclusive, thus they provide courts with a broad flexibility in determining the fairness of a use in each case. More specifically, those factors do not weight equally and, traditionally, courts have given the most weight to the first and fourth factors[39]. In fact, it is up to a court to decide which factors have a determinative value. Courts would also refer to some general principles, other factors like a market failure or public benefit of a secondary use.

The doctrine was extensively analysed and interpreted by a sitting judge Pierre N Leval in his influential article "Toward a Fair Use Standard"[40]. To him, the first factor, "the purpose and character of the use", is "*the heart and soul of a fair use case*"[41]. The secondary use of work has to be transformative, meaning that:

> *"The use must be productive and must employ the [original work] in a different manner or for a different purpose than the original . . . If . . . the secondary use adds value to the original — if the quoted matter is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings — this is the very type of activity that the fair use doctrine intends to protect for the enrichment of society."*[42]

This interpretation was mainly aiming to stimulate creative transformations of original works adding to an ultimate goal "*to promote the progress of science and useful arts*"[43]. Later the doctrine was further expanded to address purely technological uses in a way that more attention was paid to the "*different manner or a different purpose*".

To demonstrate the difference between two rationales, the old approach would focus on how the original work was changed (transformed) into the secondary creation. The

[38] United States Constitution, § 107.
[39] Compare *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 579, 114 S.Ct. 1164, 127 L.Ed.2d 500 (1994) (focusing primarily on first factor and whether use is transformative) and *Leibovitz v. Paramount Pictures Corp*., 137 F.3d 109, 114-15 (2d Cir.1998) (affirming summary judgment of fair use for parody based primarily on the first fair use factor) with *Harper & Row, Publishers, Inc. v. Nation Enters*., 471 U.S. 539, 566, 105 S.Ct. 2218, 85 L.Ed.2d 588 (1985) ("[The fourth] factor is undoubtedly the single most important element of fair use.").
[40] Leval, P.N. 1990, "Toward a Fair Use Standard", *Harvard Law Review,* vol. 103, no. 5, pp. 1105-1136.
[41] Karapapa & Borghi, 2013, p. 22.
[42] Leval, 1990, p. 1111.
[43] Article 1, sec 8, clause 8 of the U.S.C., often referred to as the "Copyright Clause".

second one addresses the question "*in which way?*" the work is used. The former requires some creative input and adding new meaning[44], while the latter one is pleased with a new application, new "*function*" of the use even when little if any, changes are made.

This later application of the doctrine to technological uses is often associated with a so-called "*functionality test*". The test essentially stipulates that the fair use defence may be invoked when a secondary work performs a different function than that of the original, regardless of any similarities between the two of them.[45]

3.2.1. Web page caching

Web caching is a particularly good example of widespread technological uses of copyrighted works without prior authorisation of rightholders. For instance, the Google search engine allows users to see cached versions of web pages included in search results. To that end, Google takes a snapshot of every web page as a backup in case the original page is not available. Those copies then become part of Google's cache and are aimed to be used in cases when original web pages are not responding[46].

It is important to make a distinction between cached copies made and stored by Google as an internet service provider on its servers and temporary cached copies made and stored on a user's computer. While the latter one was recently ruled to comply with a copyright exception under the Art. 5(1) of the InfoSoc Directive[47], the former activity is very controversial in this respect and seemingly cannot benefit from the same European exception.

It may be a coincidence that Google was first time challenged in courts with respect to its cached web page copies both in Europe and the US in the same year of 2006. The factual backgrounds in two cases are also similar: a web page owner claimed a copyright infringement in making cached copies of their websites available through Google Search.

---

[44] *See Campbell* case in terms of parody.
[45] Karapapa & Borghi, 2013, p. 24.
[46] Google. *View web pages cached in Google Search Results*. Available: <www.support.google.com/websearch/answer/1687222?hl=en> [2018, 25.11].
[47] *See* C-360/13, *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others* [2014] EU:C:2014:1195. Available: <www.curia.europa.eu/juris/document/document.jsf?docid=153302&doclang=EN> [2018, 25.11].

The US case involved a lawyer, Blake A. Field. He composed around fifty brief stories within few days, registered copyright in each of them, uploaded them to his website making them freely available online and then sued Google over cached copies of his works requesting $50,000 in statutory damages for each work[48]. Although having a bad faith character, it seemed to be a promising business model. However, the court dismissed Field's claims on four separate grounds. The most relevant two are implied licence and the fair use.

The implicit consent was inferred by the Court from the fact that Field knew how Internet search engines operate and could use the "no archive" Meta tag, or a *robots.txt* file, to ensure that Google does not display cached links to his pages[49]. This is an opt-out mechanism that was discussed above, and, because the plaintiff did not take that measure, it was interpreted as the grant of a license to Google to make cached copies[50].

On the fair use account, the Court exempted Google from copyright liability with an argumentation that follows. The first fair use factor weighted well in Google's favour because its presentation of cached links to the copyrighted works at issue served different functions than the original work, namely: a) it enabled users to access content when the original page was inaccessible; b) it allowed users to detect changes that had been made to a particular web page over time; c) it allowed users to understand why a page was responsive to their original query.

The second fair use factor was found to weigh only slightly against the fair use. It is because even assuming that the works were creative, Field deliberately made them completely available for free to everyone.

The third fair use factor was neutral, as it was necessary for Google to make entire copies of web pages to serve its transformative purposes. The Court remarked that even copying of entire works should not undermine a fair use finding where the function of the new use differs from the original.

Finally, the fourth fair use factor strongly supported the finding of the fair use as there was no evidence that Google's cached links had any impact on the potential market for Field's copyrighted works. Field argued that the market for his works was harmed as he was deprived of potential revenue he could have obtained by licensing Google the right

[48] *Field v Google*, 412 F Supp 2d 1106 (2006).
[49] *Ibid*. p. 188.
[50] *See Field* at 1116.

to make cached copies of his web pages. The Court, however, did not see any evidence of a market in which someone could license search engines the right to present cached links to web pages containing protected works, or evidence that such a market could develop in future.

The EU case took place in Spain, where an owner of the website "www.megakini.com" claimed that Google violated her copyright by means of its cached links[51]. In particular, the plaintiff asserted two copyright infringement acts undertaken by Google: the short "snippet" below the link to her website in the search results page and reproducing and making available of a cached copy of the website via cached links.

The court of first instance rejected that claims on the grounds of temporary reproduction exception (InfoSoc Directive) and safe harbour clause (e-Commerce Directive) transposed into the local Spanish law. The Court of Appeals upheld that ruling although on different grounds. In particular, short "snippets", in the court's view, were *de minimis* and, hence, non-infringing. However, the "cached" pages could find protection neither under the safe harbour nor under temporary reproduction exception.[52]

Instead, the Court did something peculiar to find Google not liable in that case. The Court first asserted the need to use the three-step test[53] in a fashion similar to the fair use doctrine: not only as a basis to define scopes of copyright exception but also as a guideline to determine limits of the concerned exclusive rights. With this in mind, the Court applied a Roman law principle *ius usus innocui* - the "*right of using someone else's property in a way that does not harm its owner, whose rationale is to prevent an overreaching protection of the owner's right*"[54].

In 2012 the Spanish Supreme Court confirmed the ruling of the Appellate Court and also emphasised the need to revert to general principles of the law when considering cases which are not specifically regulated by statute[55]. Such principles may include, for example, good faith, prohibition of an abuse of rights and, lastly, *ius usus innocui.*

[51] Lopez-Tarruella, 2012, p. 191 referring to the case *Pedragosa v. Google Spain, S.L.*, Provincial Audience of Barcelona (Sec. 15), 17 Sept. 2008, WESTLAW AC 2008/1773.
[52] Karapapa & Borghi, 2013, p. 41.
[53] InfoSoc Directive, art. 5(5).
[54] Lopez-Tarruella, 2012, p. 193.
[55] Karapapa & Borghi, 2013, p. 42.

3.2.2. Thumbnails

The second important kind of technological uses of copyrighted works that were challenged in courts involved reproduction of images - copied, reduced in size, stored and reproduced by search engines - mostly known under the name of "*thumbnails*". The thumbnailing process essentially includes several steps: internet "crawlers" would make copies of original images, then computer software would use them to create small-sized copies (thumbnails) and, finally, copies of full-sized originals would be deleted. Thumbnails would be then stored on servers and presented in return to a search query.

There were a number of cases assessing the legality of thumbnails and again it appeared to be a more difficult task for EU courts than for their US counterparts. Two most cited US rulings in the context of thumbnails came from *Kelly v Arriba Soft*[56] and more recent *Perfect 10 v Amazon*[57] case. Arriba Soft operated image search engine website and eventually was sued for a copyright infringement by a professional photographer, Leslie Kelly.

The district court[58] and later Ninth Circuit concluded that the contested use was a fair reproduction and thus non-infringing. Specifically, the Ninth Circuit held that the function of Arriba's images was different from the Kelly's use because thumbnails were used to improve access to information on the internet. In addition, the Court emphasized that it is doubtful that someone would use low-quality thumbnailed images for illustrative or aesthetic purposes. Thus the Court concluded that because of serving a different purpose Arriba's use was transformative[59].

The second case was similar to *Kelly* in respect of the use of thumbnails, however, the circumstances differed and led to a different ruling. Perfect 10 was in a business of serving adult magazines with pictures of nude models. The key difference from *Kelly* was that Perfect 10 licenced its pictures in reduced-size format to some third parties to be used on mobile phones. In other words, there was an existing market for small-sized pictures and the plaintiff claimed that Google's thumbnails could be used as a substitution. This argument was taken by a district court as the main reason to deny fair use defence for Google.

---

[56] *Kelly v. Arriba Soft Corp*., 336 F.3d 811 (9th Cir. 2003).
[57] *Perfect 10, Inc. v. Amazon.com, Inc*., 508 F.3d 1146 (9th Cir. 2007).
[58] *Kelly v. Arriba Soft Corp*., 77 F.Supp.2d 1116 (C.D. Cal. 1999).
[59] *Kelly*, 336 F.3d, at 811, 819 (9th Cir. 2003).

On appeal, the Ninth Circuit reversed that ruling. It found that the use of thumbnails was actually a fair use[60]. As regards the market effect factor the Court stated that there was not enough evidence that any downloads of thumbnail images for mobile phone use had taken place via Google and, as a result, the potential market harm was just hypothetical[61].

In Europe, the use of thumbnails in search results was treated differently in the different Member States. Even within one jurisdiction courts lacked consistency in their approaches to how to decide on cases with similar factual circumstances. For a purpose of demonstration, this research refers to French and German rulings.

In France, in a 2008 case - *SAIF v. Google France and Google Inc* - a French collective rights management society sued Google France and Google Inc. claiming illegal reproduction and display of thumbnail images in Google's search results. A Paris court of first instance held that the applicable law, in this case, was the US law and thus applied the fair use doctrine to hold Google not liable for a copyright infringement.[62]

The ruling, however, was appealed and consequently decided in a slightly different manner. The Court of Appeal found that actually a French law must be applied in this case; however, Google was still not liable on the grounds of safe harbour principle of the e-Commerce Directive transposed into a national law.[63]

Interestingly, the Paris court of first instance took a different approach already in 2009 in the case "*H&K v. Google"*[64]. In this case, a photographer Mr. André R. and his producer H&K brought a lawsuit against Google for unauthorised reproduction of their copyrighted images in thumbnails. This time, the court rejected Google's claim about the US law applicability and found the reproduction in question infringing under the French copyright law. Moreover, the court also held that because Google's thumbnail images lacked any reference to the original author it violated the author's moral rights. Furthermore, the reduction in size *per se*, which is essential for making thumbnails, was found to infringe the right to the integrity of the work.

[60] *Perfect 10*, 508 F.3d, at 1146, 1168 (9th Cir. 2007).
[61] *Ibid*.
[62] Lopez-Tarruella, 2012, p. 183.
[63] *Ibid*., p. 184.
[64] *Ibid.*

In Germany, thumbnails were embraced with more loyalty than in France, although no favouring copyright provision could be found there either. An exemplary case in this respect is known as "*Vorschaubilder*" (thumbnails) case. The plaintiff was an artist who had photographs of her paintings on her website. She filed a complaint against Google because Google displayed some of her pictures in form of thumbnails. Here the saga of courts' interpretation began.

The court of first instance dismissed the claim on the grounds of implied licence, similar to that in the US *Field* case. The Court of Appeal also dismissed the case, however, on the grounds of abuse of law. It rejected implied licence argument stating that mere publishing and making available of works on the Internet is not enough to infer implied licence. However, it concluded that the plaintiff abused her rights when suing Google because she set her website in a way that attracted Google's crawlers (search engine optimization).[65]

The German Federal Supreme Court reached yet different decision on this case[66]. It rejected the abuse of law theory and still held Google not liable. This time it was some kind of a simple consent that was found by the Court to exempt Google from liability. The Court explained that in principle Google's thumbnailing process is a *prima facie* infringement and no copyright exception or limitation could cover it. The implied licence doctrine could not be invoked because the plaintiff asserted her copyright by a specific notice next to her pictures, meaning that she did not intend to allow any kind of unauthorised reproduction.

The simple consent, that the Court invoked in this case, was inferred from the mere plaintiff's behaviour that included posting pictures without prevention of internet crawlers and search engine optimization. This simple consent did not confer any rights on Google with respect to pictures in question but was enough to exempt it from a liability. To the Court, this consent can be revoked by the plaintiff anytime by means of website setting measures she can take to prevent crawlers from copying her pictures.

3.2.3. Plagiarism detection

Yet another example of technological uses of copyrighted works, that received some judicial review, is a well-known plagiarism detection service *Turnitin.* The service helps

[65] *Ibid.*, p. 185.
[66] Bundesgerichtshof (BGH) (German Federal Supreme Court) 29 April 2010, I ZR 69/08

to check the originality of a newly written work by comparing its text to other works previously uploaded to the *Turnitin's* database. According to the publicly available information, this database comprises today of 62 billion of indexed web pages, 734 million student papers and 160 million of scholarly journals and articles[67].

It is not hard to determine how *Turnitin* is actually building its database. Information from web pages must be copied by crawlers and stored on servers similarly to how Google search engine is caching web pages. At least in the US after *Field*, this activity does not really raise further copyright questions and is a clear fair use. Scholarly journals and articles are most probably licenced from respecting publishers and proprietary databases.

As for student papers, the user agreement grants *Turnitin* a licence to use any uploaded student paper for a purpose of providing their services and improving quality of the services generally. The papers in question would be archived by default and stored in the database for an undefined period of time[68].

In fact, institutions may choose to not have their papers archived[69] and students may request through their respecting educational institutions to have their works removed from the service's archive[70]. Hereby some opt-out mechanism is available within *Turnitin*. However, it is hard to believe that any educational institution would choose to opt out from archiving since they are also interested in preventing copying from previous student's works. Therefore, Institutions need all student works to be stored as a benchmark of originality.

In 2006 few students in the US filed a lawsuit against the company that controlled the *Turnitin* website - *iParadigms LLC.* They claimed unlawful copying of their works to the *Turnitin's* archive and demanded compensation. One of the students' papers had a clear notice asking not to be used for archiving.

The District Court rejected students' claims on two grounds. Firstly, there was a contractual agreement between plaintiffs and defendant which entered into force when

---

[67] Turnitin. *The Leader in Preventing Plagiarism*. Available: <www.turnitin.com/en_us/what-we-offer/feedback-studio> [2018, 30.04].
[68] Turnitin. *User Agreement*. Available at the registration page: <www.turnitin.com/newuser_join.asp?svr=316&session-id=e6fe8036dd13e223b12c5eecb120b9d0&lang=en_us&r=14.8462595622565> [2018, 25.11].
[69] Turnitin. *Legal FAQ* page: <www.turnitin.com/en_us/about-us/privacy#terms> [2018, 30.04].
[70] User Agreement. *Supra* note 68.

students clicked "I agree" under the user agreement before submitting their papers. In addition, the written notice against archiving was not considered to change the provisions of that contract.

The second and most important in this analysis argument in favour of iParadigms was a fair use. The court concluded that the purpose of archiving works was to prevent plagiarism by comparative use and therefore it was transformative. In addition, the use in question could not in principle damage the market value of high school term papers and other such student works[71].

Plaintiffs appealed, claiming inter alia that the use could not be regarded transformative because it does not add anything new to their works. The Appeal Court, however, citing *Perfect 10* case, asserted that "*the use of a copyrighted work need not alter or augment the work to be transformative in nature. Rather, it can be transformative in function or purpose without altering or actually adding to the original work*"[72].

3.2.4. Books digitization

Starting from 2004 Google entered into an agreement with over 40 libraries around the world to obtain digital copies of their books and to make them available for an online search. According to agreements, Google would perform all the scanning-related work and libraries would retain an electronic copy of each digitized book as a reward for the access they provide. It looked like a win-win situation: libraries would get their collections digitized - otherwise costly undertaking that not every library can afford; and Google would obtain digital copies of books for their project with an investment they can actually afford to make.

The ultimate goal of the project was to make books accessible through Google internet search. The result page would render some general information about the book, similar to that of library cards, and in addition:

a) full-text view of out-of-copyright books and those, where permission was granted by respecting authors or publishers;

b) limited preview with authors'/publishers' permission;

[71] *AV et al. v iParadigms*, LLC, 544 F. Supp. 2d 473 (2008)
[72] *AV et al. v iParadigms*, LLC, 562 Federal Reporter, 3d Series [2009], 630–647 (USA).

c) snippet view of few sentences to display user's search term in context.[73]

In 2005 the Authors Guild Association of America sued Google over so-claimed massive copyright infringement. In 2008 parties to the dispute reached a Settlement Agreement. The Agreement was not approved and in 2013 the District Court ruled in Google's favour based on fair use[74].

The Authors Guild appealed in 2014 and in 2015 the Second Circuit Court supported the ruling in Google's favour[75]. The Authors Guild requested a writ of certiorari from the Supreme Court asking to revise the appellate decision. The Supreme Court left the ruling unchanged in 2016 closing the dispute[76].

With respect to the fair use analysis, it is interesting to mention that the same judge Leval, who originally played a big role in developing this doctrine, provided his opinion in this case on the appellate stage. The transformativeness of Google Books, as part of the first factor, was found to be based on a different purpose that this service provides in comparison to that of original works. The Judge held that it is particularly transformative to make copies when the purpose is to enable search and identification of books that contain a term of searcher's interests. Moreover, he stated that because of snippet view only allows seeing the frequency of the sought term appearing in the book, it makes the highly transformative search function even more valuable.

As regards the second requirement, the nature of works did not preclude the finding of a fair use because Google Search provides information about the book, rather than "*replicating protected expression in a manner that provides a meaningful substitute for the original*"[77]. Similarly, judge Leval justified making entire copies of books: "*it does not reveal that digital copy to the public*"[78]. Entire copies are required to enable the search functions, and the snippet view is not significant enough to offer a competing substitute. Therefore Google also satisfies the third factor test.

As for the fourth factor, the Judge recognised that the snippet function may cause some loss in sales of the original books. However, that is mainly attributable to interests not

[73] Google. *What you'll see when you search on Google Books*. Available: <books.google.com/googlebooks/library/screenshots.html> [2018, 25.11].
[74] *Authors Guild Inc. v. Google Inc*., 954 F.Supp.2d (S.D.N.Y. 2013).
[75] *Authors Guild v. Google Inc*., 804 F.3d (2d Cir. 2015).
[76] *Authors Guild, et al. v. Google, Inc*., 15-849.
[77] *Authors Guild v. Google Inc*., 804 F.3d (2d Cir. 2015).
[78] *Ibid*.

protected by copyright. A searcher may satisfy its demand in some factual information like a historical fact from a snippet view, but that information is not protected by copyright and the activity is thus noninfringing.

In the course of discussing books digitization projects, it is also important to mention the case *Authors Guild v HathiTrust*[79]. The legal action, in this case, was filed and resolved in 2012 - long after the start of litigation against Google, but also before the *Google Books* was decided. Arguably, the ruling in *HathiTrust* favoured Goole's position.

HathiTrust is a partnership of main American research institutions and libraries aiming to ensure preservation and access to cultural works. Its repository comprises of public domain and copyrighted content from different sources like Google, the Internet Archive, Microsoft, and others[80].

Like Google, HathiTrust enabled full-text search of books, but, unlike Google, they also provided free access to books for people with print disabilities. Unsurprisingly, District Court and later Appellate body found the use of books for this purposes complying with fair use requirements. What is interesting in this case, the full access to digital copies of books for print-disabled people was not found to harm authors, because "*the present-day market for books accessible to the handicapped is so insignificant*"[81] that copyright holders normally do not even consider it as a source of revenue. To put it differently, the Court did not deny the existence of the market of works adapted for print-disabled people, it only held that it was not significant to harm.

In Europe, courts embraced books digitization in a slightly different manner. In 2005 Google was sued in France by a publishing group La Martiniere for an unauthorised reproduction of French books. In 2009 the Paris Court delivered its decision finding Google's practices illegal[82]. Apart from claiming applicability of the US law, in the present case, Google also asserted that its conduct was complying with French copyright law. In particular, they argued that the snippet view of French books displayed in a Google Books search was consistent with French copyright exception permitting short quotations for informative purposes.

---

[79] *Authors Guild Inc v HathiTrust*, No 11 Civ 6351 (HB), 2012 US Dist.
[80] HathiTrust digital library website. Available: <www.hathitrust.org/> [2018, 25.11].
[81] *Authors Guild, Inc. v. HathiTrust*, 755 F. 3d 87 - Court of Appeals, 2nd Circuit 2014.
[82] *Editions du Seuil et autres v Google Inc et France*, Paris District Court, 3rd Chamber, 2nd Section, 79 PTCJ 226, 18 December 2009 (France).

However, the law requires some creative input on the part of a user and an aim to illustrate the subject matter of the original work. Hence, that argument was rejected by the court. In addition, the Court also held that Google Books violated authors' right of integrity. Google representatives firstly planned to appeal that decision; however, they settled the dispute in 2011. It can be explained by very poor prospects to win that case in French copyright realities.

Interestingly, a German publisher WBG together with the German Publishers Association sued Google over its Library Project nearly at the same time as the French one. However, they withdrew their claims shortly after they were told by the Copyright Chamber of the Regional Court of Hamburg that their legal action was unlikely to succeed[83]. It was said, that the Court decided to draw a comparison between the book's short excerpts and snippets used in Google web search. However, it is not totally clear, what legal basis the German Court would apply to find such use not infringing copyright law.

3.3. Case law summary comments

An analysis of cases from *Kelly* to *Google Books* allows outlining some general judicial approach to technological uses of copyrighted works. That may help to envision possible future judicial treatment of uses that involve ML technologies in the US and Europe.

To begin with, the first factor in the fair use analysis includes two seemingly unrelated components: first is a purpose and character of the use and, second is whether the use is commercial or not for profit. These two elements gained different decisive weight in the cases discussed here.

The purpose and character of a secondary use have been always assessed in the light of transformativeness within a meaning of the famous formula from the judge Leval. It is needless to cite every court decision to demonstrate that every judge was stressing on a different purpose or function in assessing the first factor and finding the use non-infringing. Hence, the functionality test found its direct application in all of the courts' rulings. The different function of the use was enough to declare it transformative even

[83] Google's official blog 2006, June 28. *Germany and the Google Books Library Project*. Available: <www.googleblog.blogspot.fi/2006/06/germany-and-google-books-library.html> [2018, 25.11].

in absence of any changes to original works. Consequently, it became one of the decisive factors in the whole analysis.

It is important to discuss the role of commerciality as a part of the first factor. In fact, it is a lack of any significant role of this test in a fair use analysis that grabs attention. In particular, courts would conclude that commercial motivation cannot undermine the claim of fair use when the use itself is highly transformative[84]. This rule stems from the principle set down in the *Campbell* case, literally holding that "*the more transformative the new work, the less will be the significance of other factors, like commercialism, that may weigh against a finding of fair use*"[85].

In the Second Circuit ruling on *Google Books* judge Leval held that "*many of the most universally accepted forms of fair use, such as news reporting and commentary, quotation in historical or analytic books, reviews of books, and performances, as well as parody, are all normally done commercially for profit*"[86]. It is in comfort with an interpretation that the very inclusion of a commerciality test in the statutory fair use provisions does not infer that commercial uses are presumptively unfair. However, this test may gain its relevance in cases where the secondary use lacks transformativeness in purpose.

The second factor, the nature of the copyrighted work, received limited attention in judicial assessment and never determinative. Thus for example, in *Field*, the nature of works weighted slightly against the finding of fair use. Similarly, it was only slightly in favour of the plaintiff in the *Kelly* case.

Further, the bulk, entire and verbatim copying was normally assessed by courts within the third fair use factor, namely, the amount and substantiality of the portion used in relation to the copyrighted work. It is the scale of copying that amazed Authors Guild when they claimed a massive copyright infringement by Google. However, in none of the cases analysed here, this factor played any determinative role. Technological uses, based on web caching, make billions of web page copies. Google scanned millions of books in their entirety including cover pages. Yet these factors alone were neglected in courts.

---

[84] *Authors Guild, et al. v. Google, Inc*., 15-849.
[85] *See Campbell*, 510 U.S. at 579.
[86] *Authors Guild v. Google Inc*., 804 F.3d (2d Cir. 2015).

Consequently, neither the nature of a work nor amount of the use would prevent courts from finding a fair use in cases of technological uses. This outcome suggests that the mere fact that ML requires a large number of highly creative works would not matter at all in finding this activity non-infringing.

There has been some controversy around the fourth fair use factor, namely the effect of the use upon the potential market for or value of the copyrighted work. It becomes understandable when thinking about any secondary use of a copyrighted work as a derivative use. From this perspective, plaintiffs tend to claim that almost any new use of their work is part of an unexplored derivative market[87]. Indeed, following a broad concept of a derivative work, it is possible to assume that everything is a potential market effect.

In order to preserve the functionality of the fair use doctrine and bring some clarity into the potential market concept, courts adopted some guiding rules. First, "*the market for potential derivative uses includes only those that creators of original works would in general develop or license others to develop*"[88]. Thus in *Field*, the Court could not identify any existing licensing market for making cached links to web pages with protected works, neither it found evidence that it would develop in future.

Second, the important limitation laid down in *Campbell* case and referred to in *Google Books* and *HathiTrust* suggests that "*the market harm analysis is concerned with only one type of economic injury to a copyright holder - the harm that results from the secondary use serving as a substitute for the original work*"[89]. In other words, any market harm resulted from secondary transformative uses that do not serve as substitutes for the original work normally would not count in courts.

Third, the market harm claimed by plaintiffs must be cognizable under copyright. To be more specific, if there is any negative effect on the market or on the potential market of a copyrighted work occurs, it must be examined whether that harm results from a use that is under the exclusive right of a copyright owner. To put it differently, if the harm derives from a use which is not protected by copyright, that market effect will normally be not considered in court.

---

[87] Sag, 2009, p. 43.
[88] *Campbell* at 569.
[89] Matulionyte, R. 2016, "10 years for Google Books and Europeana: copyright law lessons that the EU could learn from the USA", *International Journal of Law and Information Technology,* vol. 24, no. 1, p 59.

For example, in *Google Books* judge Leval argued that some loss in sales may occur because users would satisfy their search for a historical fact from a snippet view. It is also possible to assume that authors would like to develop a market for information about their books and license it to search engines. However, that kind of market harm "*will generally occur in relation to interests that are not protected by the copyright*"[90].

Fourth, argumentation of Leval also suggests that the magnitude of market harm matters as well. It must be meaningful or significant to matter[91]. Similarly, in *HathiTrust* the court emphasised that, if the existing licensing market for uses in question is very insignificant (like in case with books adapted for print-disabled), it will not be considered by courts as sufficient[92].

Finally, as it was mentioned in the introduction to the fair use doctrine, apart from the known four criteria courts sometimes assess other factors like a public benefit of a secondary use. Some commentators and judges refer to it as a fifth fair use factor.

Indeed, in *Kelly*, the court emphasised that search engines "*benefit the public by enhancing information-gathering techniques on the internet*"[93]. Later, the same argument was raised in the *Field* case. In *Perfect 10*, Google was praised for a public benefit that search engines provide by using an original work into a new work as an electronic reference tool[94]. In *iParadigms,* the court stressed on "*a substantial public benefit through the network of educational institutions using Turnitin*"[95]. Lastly, in *Google Books*, the District Court judge Denny Chin dedicated a separate section in its decision called "The benefits of the Library Project and Google Books" and discussed how useful this service is for an information society.

The requirement of public benefit can be traced back to the *Campbell* case when judge Leval held that the fair use factors should be considered in light of the purpose of copyright which is namely "*the Progress of Science and Useful Arts*". The established case law thus may suggest, that the more the secondary use contributes to the progress of science and useful arts, the more it adds to the finding of fair use. It shall be wrong

[90] *Authors Guild v. Google Inc.*, 804 F.3d 202, 224 (2d Cir. 2015).
[91] *Ibid*.
[92] *HathiTrust*, 755 F3d 87, 23–24.
[93] *See Kelly*, 336 F.3d at 820.
[94] *Perfect 10*, at 1165.
[95] *AV et al. v iParadigms, LLC*, 562 Federal Reporter, 3d Series [2009], 630–647.

though to imply that a lack of public benefit may undermine the defendant's position because it is not a mandatory requirement of the doctrine as it is codified in law.

Moreover, some argue that courts are not always the right place to determine to what extent or whether at all some new emerging technology is of benefit to the public, as it is especially unclear at that early stage[96]. There is always a danger that a judge will not fully recognize the ultimate value of the system in question and hamper its further development. As it was once rightly articulated, "*the copyright law can make or break emerging technologies...*"[97]

As regards the EU copyright, there is no uniform approach on how to treat unauthorised technological uses of the kind discussed here. There is simply no suitable EU copyright exception that could cover those uses that are deemed to be fair in the US.

Some courts, in order to avoid hampering technological progress, would choose to apply the US law where it is possible. Other courts, for example in Spain, would revert to general principles such as *ius usus innocui*. However, this latter approach does not seem to be a reliable solution for technological use challenges. It is rather a last resort for courts to protect legitimate interests in the absence of specific copyright rules.

As can be seen, the copyright exception for a temporary reproduction prescribed by the InfoSoc Directive also cannot cover most of the technological uses, such as thumbnails and cached copies of web pages. For instance, the kind of caching in the Spanish *Google Cache* case is different from the one mentioned in the Recital 33 of the InfoSoc Directive[98]. For the search engine to function effectively, it is not essential to have a cached copy of a website, especially when the original source was deleted or changed over time.

In addition, courts, in general, refuse to apply an implied licence doctrine to the content uploaded on the Internet. A too broad interpretation of an implied consent would contravene with a general opt-in principle, meaning that a specific consent must be sought prior to making any copy of protected works. A German alternative innovation

[96] Reese, A. 2005, "The Problems of Judging Young Technologies: A Comment on Sony, Tort Doctrines, and the Puzzle of Peer-to-Peer", *Case Western Reserve Law Review* 55 (2005) 877, 887.
[97] Sobel, 2017, pp. 5 and 33.
[98] Recital 33 of the InfoSoc Directive: "*...this exception should include acts which enable browsing as well as acts of caching to take place, including those which enable transmission systems to function efficiently...*"

of a *simple consent* concept in the *Vorschaubilder* case[99], just like the Spanish *ius usus innocui* principle, seems to be too vague to serve a reliable copyright exemption function.

Consequently, the European case law demonstrates that the different EU Member States treat technological uses of copyrighted works differently. Spain and Germany would find some other than copyright rules to protect interests that seem to be legitimate from a general law principles perspective. France, in contrast, would stick to a narrow copyright law interpretation and forbid uses that in most countries would be regarded as lawful.

It can be explained by two main reasons. The first one is an outdated EU copyright law. The last attempt to modernize copyright rules adapting them to new technologies was undertaken in 2001 InfoSoc Directive. Already in 2006 Google was challenged in court for its cached copies of web pages and could not rely on any copyright exception introduced in the Directive. Certainly, a rapid development of new technologies requires the law to follow up accordingly.

The second reason could be a lack of harmonization of copyright rules between the EU Member States. Although the InfoSoc Directive was aimed to bring closer varied national copyright laws, most of the exceptions had an optional character meaning its voluntary implementation. Moreover, Member States were also granted some degree of national discretion in implementing the Directive's rules.

Apart from the need to update the EU copyright law with some mandatory technological copyright exceptions, there could be also an alternative solution to the described copyright issue. It is a possibility to invoke fundamental human freedoms as a limitation mechanism to far-reaching exclusive rights, what both ECtHR and CJEU stressed upon for multiple times[100].

Copyright as an element of an intellectual property must be respected and protected under the Article 17(2) of the Charter of Fundamental Rights of the European Union (hereinafter the EU Charter). However, that protection must be balanced against the

[99] Bundesgerichtshof (BGH) (German Federal Supreme Court) 29 April 2010, I ZR 69/08.
[100] For more analysis on the intersection of human rights and intellectual property *See* in general Geiger, C (ed.) 2015, Research Handbook on Human Rights and Intellectual Property, Edward Elgar Publishing, Incorporated, Cheltenham. Available from: ProQuest Ebook Central. [2018, 25.11].

protection of other fundamental rights, such as e.g. the right to freedom of expression and information guaranteed by the Article 11 of the EU Charter.

For instance, it could be argued that Google's cached copies of web pages allow users to access information in times when it is no longer available, to compare changes on the web page and understand why the page was responsive to their search. It is an undisputed fact that analysing cached copies of web pages is a particularly useful tool for investigative journalists[101]. Therefore, this Google caching function could have great prospects of finding legality under the freedom of expression and information if duly assessed by courts in that fashion. Apparently, in 2012 the Spanish Supreme Court did not consider this way of reasoning and had to rely on general principles of law instead.

[101] Chavez, S. , *10 Google tools investigative reporters can use to find information*. Available: <www.ijnet.org/en/story/10-google-tools-investigative-reporters-can-use-find-information> [2018, 05.12].

4. PROSPECTS OF MACHINE LEARNING USES IN THE US

Why robots get for free something that humans have to pay for?[102]

4.1. The input reproduction assessment

Assessing legality of unauthorised uses of copyrighted works for Machine Learning purposes, relying solely on the existing case law, does not seem to be that straightforward task. The complication inherently comes from differences in how copyrighted works are used.

All of the cases discussed in the previous chapter dealt with uses that include partial (e.g. snippet view) or complete (e.g. cached copies of web pages) demonstration of original works to the public, although serving a completely different purpose by doing so. At the same time, ML uses do not intend to make any display of the original expression. However, the possibility of such display shall not be disregarded.

This study will further analyse separately two categories of reproductions associated with ML uses. The first group of reproductions takes place when original works are initially copied and then analysed as a part of a training dataset in order to discover important patterns. It would be correct to refer to this kind of copying as a *non-display use* because they do not display any part of original works to the public. These reproductions take place only due to an automation factor involved in the process of analysis of a vast amount of information.

The second kind of reproductions is relevant for the final outcome produced by ML models. It is not, however, always a case that those reproductions will take place, but since an expressive AI model is meant to create new images or generate human-like language, some similarities with original copyrighted input might occur.

For the first time, non-display uses gained its copyright relevance from the rejected settlement agreement in the *Google Books* case. According to the Agreement, "*Non-Display Uses*" meant "*uses that do not display expression from digital copies of books or inserts to the public. For example, display of bibliographic information, full-text*

[102] The fair use discourse in the US often culminates over the unsettling fact that while humans have to pay for a copy of protected work, making same copy for computational analysis (computer readership) may be exempted from a copyright infringement.

indexing without display of expression, algorithmic listings of key terms for chapters of books, and internal research and development using digital copies are all non-display uses"[103].

While the "*Non-Display Uses*" term is rather broad, the "*Non-Consumptive Research*" is more specific and meant "*research in which computational analysis is performed on one or more books, but not research in which a researcher reads or displays substantial portions of a book to understand the intellectual content presented within the book*"[104]. Thus, the latter one seems to be covered by the former one and, according to the Agreement, it includes such uses as analysis of images and texts, extraction of information, automated translation, linguistic analysis, and others.

If approved, the Agreement would permit Google to conduct non-display uses without prior authorisation of authors. However, the only fact that the Agreement was not approved by the Court does not mean that Google would refrain itself from a temptation to do just that with the database of books it has compiled.

In this regard, Pamela Samuelson suggested, that Google would likely continue non-display uses of books even if the Agreement failed. She opined, that "*non-display uses ... would likely result in advancing knowledge and/or in the creation of new noninfringing works of authorship, such as new tools to aid in the translation of texts from one language to another*"[105]. To her, because these uses do not demonstrate any protected expression to public "*they are unlikely to bring about any harm or potential harm to the market for the underlying works*"[106].

In the way of example, in 2016 researchers at Google used 11 000 novels to train an AI-based *Smart Reply* function of their email services. The aim was to improve the program's ability to "*generate coherent novel sentences*" and thus to upgrade *Smart Reply*'s conversational skills[107]. In this case, Google seemingly did not take a risk to use books from its own database. The training data originated from a different source.

---

[103] *Amended Settlement Agreement*, Authors Guild v Google, Case No 05 CV 8136-DC, 13 November 2009, § 1.94.
[104] *Ibid*. § 1.93.
[105] Samuelson, P. 2010, "Google Book Search and the Future of Books in Cyberspace", (2010) 94(5) *Minnesota Law Review*. P. 1363 (footnote 280).
[106] *Ibid*.
[107] Bowman, S.R. et al. 2016, May 12, *Generating Sentences from a Continuous Space*, *arXiv 1*. Available: <www.arxiv.org/abs/1511.06349> [2018, 25.11]; Kantrowitz, A. 2016, May 5, *Google Is Feeding Romance Novels To Its Artificial Intelligence Engine To Make Its Products More Conversational*, BUZZFEED, available: <www.buzzfeed.com/alexkantrowitz/googles-artificial-intelligence-engine-reads-romance-novels.> [2018, 25.11].

However, it turned out that at least some of the novels were copied without the consent of respecting rightholders. In fact, novels were available for a free download, but under the licence "*for your personal enjoyment only*"[108].

A Google spokesman asserted that when the ML research is conducted on free e-books "*it doesn't harm the authors and is done for a very different purpose from the authors', so it's fair use under US law*"[109]. This argument is to some extent echoing with words of Samuelson, but with one very important difference. While both suggestions emphasise on the absence of any harm to authors of original works, Samuelson asserts, it is because of the non-display character of the secondary use, and Google spokesperson attributes it to the fee-free nature of the works in question.

The concept of harm proved to be one of the decisive factors in the fair use analysis of technological uses. Samuelson directs our attention to a more fundamental rationale that if a reproduction of a work does not communicate its protected expression to the public, authors will not suffer any harm from such use and therefore it must be deemed lawful. From this perspective, it shall not really matter whether the books were offered for download free of charge or for pay. It is only an act of displaying original expression to others that is protected by copyright.

In the context of the *Smart Reply* research example, Sobel rightly mentioned, that Google could train a brilliant, expressive AI using its entire Google Books library containing millions of electronic copies of books[110]. However, unlike providing information about the books, which is not restricted by copyright, training AI intends to harvest "*authors' varied and rich expression of ideas*" which is a very "*essence of copyrightable subject matter*"[111].

Nevertheless, if *Smart Reply* function does not replicate original expressions of the training datasets, the mere act of analysing that expression shall not lead to a copyright infringement. Moreover, as the cases dealing with *thumbnails* and *snippets* demonstrated, even use of protected expression by displaying it can be fair if it is done for a different purpose (functionality test) like it is also the case with the *Smart Reply*.

---

[108] Lea, R. 2016, *Google Swallows 11,000 Novels to Improve AI's Conversation*, The Guardian, available: <www.theguardian.com/books/2016/sep/28/google-swallows-11000-novels-to-improve-ais-conversation> [2018, 25.11].
[109] *Ibid*.
[110] Sobel, 2017, pp. 23-24.
[111] *Ibid*.

It is interesting to point out, that in the *Google Books* case most of the debate was focusing on the uses that included a partial display of those books, while one type of use was not objected by the plaintiff. Moreover, it was even praised by the judge.

In particular, judge Denny Chin, in the course of highlighting benefits of the Google Books project, mentioned that "*Google Books greatly promotes a type of research referred to as "data mining" or "text mining*"[112]. The judge here meant a *Google Books Ngram Viewer*[113] - an online service that allows users to conduct research on the massive amount of books. The tool simply renders a graph illustrating how those phrases were used in books over a selected period of time[114].

The common feature of the *Ngram Viewer* tool and the *Smart Reply* function is that neither of them is meant to reproduce the author's original expression. The first one is more obvious because it only provides statistical information about a search term. It counts words and returns a number. The latter one is more intricate in this sense because it analyses human expression to produce another human-like expression.

In the context of technological uses, some commentators focused on *expressive* and *non-expressive* use classification. The line of cases from *Kelly* to *Google Books* dealt with instances of "*copying expressive works for non-expressive ends*"[115] and the *Ngram Viewer* is exactly that type of use. However, with the progress in ML, it is time to assess the copying of expressive works for highly transformative and innovative expressive ends. Can non-display technological expressive use of copyrighted works invoke fair use defence, provided that the secondary expression is distinct from the original one?

As it was discussed in the previous section, technological uses of copyrighted works would normally be tested against transformativeness, market effect, and public benefit factors. The study will now address each of them in the context of expressive ML uses and try to create some guiding principles towards finding a fair use.

First of all, ML uses employ copyrighted works in a completely different manner and for a completely different purpose than the original use. On the account of the different manner, a computer program does not read a work, let's say a book, in a way that humans would do it. It is also more evident that computers do not perceive an image in

[112] *Authors Guild Inc. v. Google Inc.*, 954 F.Supp.2d (S.D.N.Y. 2013)
[113] Google. *Ngram Viewer*, available: <books.google.com/ngrams> [2018, 25.11].
[114] Google. *What does the Ngram Viewer do?* Available: <books.google.com/ngrams/info> [2018, 25.11].
[115] Sag, 2009, p. 3.

the same way as humans. Similarly, the purpose of human reading typically would be to enjoy an original author's expressions or retrieve some useful information. For ML, in contrast, it is a discovery of unique patterns in the process of building an AI model.

Some may argue, that at least in some cases machines do not do more with copyrighted works than humans would do. A quick example may be a text mining to extract some valuable information. Exempting machine "reading" from a copyright infringement would encourage humans to outsource their reading to computers[116].

However, the scale of machine processing outreaches human reading possibilities in countless times. Furthermore, greater possibilities are not simply in a greater number of works that computer would process. Some value can be extracted only when a large number of works is being combined together. That is what some call a "*collective intelligence of a large digital library*"[117]. That intelligence might never be accessible within reach of human reading capabilities. Hence, the purpose just like a manner of the use would differ dramatically from a traditional original use.

Applying language of the judge Leval, ML would use copyrighted works "*as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings...*"[118] Hence courts should not find it difficult to conclude on a highly transformative nature of such uses.

In the market effect analysis, it is necessary first to identify whether such a market exists or is likely to develop. This question may appear more controversial than it seems to be. From a first sight, it is hard to spot any signs of such a market mainly because writers do not write books with an anticipation to get royalties from licensing their works to AI developers. The same implication applies to other fields of artistic creation as well.

Nevertheless, without prejudice to its effectiveness, it is possible to argue, that some market for training data already exists[119]. For example, *Elsevier*[120] offers its large volumes of publications for text and data mining, encourages ML application to "*turn*

[116] Grimmelmann, 2016, p. 674.
[117] Karapapa & Borghi, 2013, p. 49.
[118] Leval, 1990, p. 1111.
[119] Sobel, 2017, p. 27.
[120] Originally a publishing company that now positions itself as an information and analytics company. Provider of scientific, technical, and medical information publications. Available: <www.elsevier.com/> [2018, 25.11].

data into knowledge"[121]. While it is a great place to conduct some scientific research, it might not be suitable for training expressive AI models, since its database is limited only to scientific, technical and medical information.

Other examples of the potential market of training data are internet platforms such as Facebook, Twitter, and Google etc. They aggregate user-generated content, subject to intellectual property protection under "*a non-exclusive, transferable, sub-licensable, royalty-free and worldwide licence to host, use, distribute, modify, run, copy, publicly perform or display, translate and create derivative works of [users'] content*"[122]. That content in pictures, videos, sounds etc. is potentially useful in ML projects.

Indeed, in their Data Policy Facebook says "*we ... provide information and content to research partners and academics to conduct research that advances scholarship and innovation that supports our business or mission ...*"[123] However, this formulation implies that the content can be used only in relation to a research that supports Facebook's business or mission, which is by itself a significant constraint for independent research entities pursuing their own goals. Therefore, it is not offered for anyone for their research purposes.

An interesting recent example of using content from social internet platforms is a research directed to develop an AI model able to generate poetry from images[124]. They used two datasets in the research. To build the first one they crawled content from groups in *Flickr* that use images illustrating poems. The second one was a large poem corpus crawled from websites dedicated to poetry, such as *Poetry Foundation*[125], *PoetrySoup*[126], "best-poem.net" and "poets.org".

As a rule, scraping content from internet platforms by means of web crawlers would be prohibited by terms of use, except as otherwise expressly permitted[127]. On the one hand, it is not known whether permissions were obtained in this particular project. On another hand, social media platforms do not explicitly offer their content for this kind of

[121] Elsevier. *These Elsevier collaborations use machine learning to turn data into knowledge.* Available: <www.elsevier.com/connect/these-elsevier-collaborations-use-machine-learning-to-turn-data-into-knowledge> [2018, 25.11].
[122] Facebook. *Terms of services*: <www.facebook.com/terms.php> [2018, 25.11].
[123] Facebook. *Data Policy. How is this information shared?* Available: <www.facebook.com/about/privacy> [2018, 25.11].
[124] Liu, B. et al. 2018, "Beyond Narrative Description: Generating Poetry from Images by Multi-Adversarial Training", *arXiv*. Available: <www.arxiv.org/abs/1804.08473> [2018, 25.11].
[125] Poetry Foundation website: <www.poetryfoundation.org/> [2018, 25.11].
[126] PoetrySoup website: <www.poetrysoup.com/> [2018, 25.11].
[127] For example Flickr's terms of use. Available: <www.flickr.com/help/terms> [2018, 25.11].

analysis. There is simply no information available about such services and it is only possible to assume that requests could be negotiated individually.

Consequently, the aggregation of content within separate holders, like publishers or internet platforms, suggests that some market for training data already exists and other markets could develop in the future. However, those opportunities available today fall short of satisfying demands of researchers in the field of creative AI. Internet platforms engage in mining content and developing AI models for their own commercial purposes, rather than offering their great databases for every interested party. That could be explained by the need to preserve a competitive advantage in the race of AI research and weights against the development of a meaningful and functional market.

Therefore, the current state of affairs with a market of training data for ML suggests that a potential harm to respecting rightholders is unlikely to be sufficient to preclude the finding of the fair use. As the court in *HathiTrust* pointed out, even if some market for the use in question exists but it is very insignificant, it shall not amount to sufficient harm to the market of original works[128].

As regards the impact of the secondary use on the traditional market of that copyrighted works, it is important to remember, that only harm caused by secondary use serving as a substitute for the original work would be considered as relevant in the fair use analysis. With a high degree of certainty, it can be argued, that AI-generated works would be different enough or used in a different way to avoid any substitution claim.

First of all, secondary works would be new original creations. Similarly to new human-created works they would be capable of competing with original works. Substitution is principally different from a competition. Creating substitution means competing with original authors in their own expression and is restricted by copyright. In contrast, competition in a new different expression is normally encouraged by copyright law. The principle "on the shoulders of giants" should apply equally to human and machine creations after all.

Some scholars argue, that creative AI "*not only jeopardizes the market for the works on which it is trained, it also threatens to marginalize authors completely*"[129]. The threat comes from a possible, and at least partial, automation of the creative industry.

[128] *HathiTrust*, 755 F3d 87, 23–24.
[129] Sobel, 2017, p. 30.

Unrestricted creative AI could produce works at little or no cost oversupplying market with cheaper content. Arguably, it may discourage human authors to create in such new circumstances. As the result, a doctrine that promotes computer authorship and deprives humans of incentives to create looks inconsistent with what copyright law was originally meant to achieve.

Automation or "AI-ization" of some aspects of human activities is inevitable and has already started, including the field of creative industry. Naturally, it may lead to some human professional groups losing their traditional markets and sources of income. One recent example is an AI model called *Bayou* that can write actual computer code[130]. It is claimed to be a significant breakthrough in the chain of attempts to develop a computer program able to code another computer programs. If further developed, it can disrupt the industry and oust human programming or at least alter it to a great degree.

Another danger of creative AI is that it could be used to appropriate author's artistic personality. As it was pointed out, authors risk to "*lose control over their own expressive personality, as embedded in their works*"[131]. The possibility of such artistic identity theft is more realistic than imaginative. The example of *The Next Rembrandt* demonstrates the technological potential to analyse an entire collection of one author and then create a new work mimicking an original artistic style.

With AI capabilities it would be possible to create a digital avatar that would generate new works following particular artistic patterns contained in previous works of an artist. In the positive scenario, authors could utilise this technology for their own benefit to generate subsequent creations. It would be enough to create a personal distinguishable style and then outsource it to AI to do the rest.

In the case of artistic identity theft, a traditional analysis of a copyright infringement may not always apply because it is a style and not original expression that is being copied. Therefore it might be impossible to establish any reproduction in part or in whole. At the same time, interests of the author who fall victim to AI avatar would be affected. It may cause significant market harm and also reputational losses.

However, in the same manner, as with negative effects of the AI-zation, a potential harm to rightholders' markets, in this case, would derive from an activity not restricted

[130] *What is Bayou*: <info.askbayou.com/> [2018, 25.11].
[131] Sartor, G. et al. 2018, "The Use of Copyrighted Works by AI Systems: Art Works in the Data Mill". Available at SSRN: <www.ssrn.com/abstract=3264742> [2018, 25.11]. P. 12.

by copyright law. So it should have no meaning for the fair use discussion. Moreover, as the *Sony Betamax* case also illustrated[132], when there are legal uses of a technology in place, the mere possibility that someone will misuse it for a wrongdoing should not hamper the development of that technology.

After all, current copyright law does not yet grant protection to AI-generated works. Hence, while fair use may give AI more possibilities to create, humans are in a better position with respect to protecting their creations. Although developing two different legal systems for machines and humans[133] entails a number of associated problems, e.g. ensuring that a work was written by human and not by AI, the very idea to deprive AI-generated works of copyright protection seems to be an adequate price for a facilitated access to copyrighted works. It must be remembered though, that the future of copyrightability of AI-generated works is far from being certain.

The public benefit of creative AI can be approached from two perspectives: first, it is a public utility of a specific AI model trained on copyrighted content and; second, a progress of science in general, which stems from new discoveries in the field of computer science.

The level of public utility will vary from model to model depending on the specific practical application. For instance, a sophisticated AI-based online tool generating translations indistinguishable from human-made, or some reliable AI model able to generate new computer code would be of great public benefit. At the same time, the public utility of some other applications, e.g. the Prisma App kind, could still be questioned.

On the account of the progress of science and useful arts that copyright law is aimed to promulgate, it is clear enough that greater access to training data would foster advancements in the field of AI. Therefore it would be a strong argument in a potential assessment of the so-called fifth fair use factor.

---

[132] In *Sony Betamax* the use of timeshifting video cassette recorders was found legal regardless the fact that they could be used to create infringing copies of TV broadcasted videos. *See Sony Corp. of America v. Universal City Studios, Inc.*, 464 U.S. 417 (1984).

[133] Grimmelmann (2016) in this respect notes: "*We have created a two-tracked copyright law: one for human readers and one for robots. Uses involving human readers receive close and exacting scrutiny to make sure that no market belonging to the copyright owner is being preempted. Uses involving robotic readers are fast-tracked for fair use*", p. 667.

4.2. Reproduction in the output

Fortunately, the US copyright has developed a solid case law dealing with copying of creative elements of one work into another. It would go beyond of the objectives of this paper to discuss into details elements of the classical copyright infringement case analysis. However, it is worth pointing out specific features of applying existing doctrines to ML cases.

First of all, it is always necessary to prove that a claimed copying actually took place. That can be done, *inter alia*, by demonstrating that a defendant had access to the original work of a plaintiff. While it may be a tricky task in a case with human creators, it normally would require a mere inspection of a training dataset in the case of ML use.

Further, the well discussed fair use doctrine can also apply to some cases of reproduction in the outcome. As it was emphasised before, even literal copying of original works in part or in whole and its subsequent demonstration to the public may find its protection under the fair use doctrine, provided that the use is highly transformative. The functionality test may find its best application in cases involving AI-generated works.

Just like the use of copyrighted images in thumbnails was recognised a transformative in purpose, an AI-generated work might also serve a different purpose from the original one. For instance, the use of an original expression in translations is quite different from the use in a literary work. Another, more coherent example may be the aforementioned *Prisma App*. The application uses some original expression from different works like paintings or videos in a totally different way - as a style of a particular photo editing filter.

ML uses may offer various innovative ways to reuse pre-existing copyrighted works that would be found highly transformative. The transformativeness of a secondary use would normally add to another important fair use factor - a market harm element. The more transformative use is - the less effect it would have on the market of original works.

Even in situations where a secondary use can hardly be recognised transformative and therefore not compatible with fair use requirements, a defendant still may avoid

copyright infringement claims when the amount of borrowed expression is trivial. That is very likely in cases of incidental reproductions.

In particular, according to the *de minimis* doctrine in the US, in order to establish infringement, the secondary use must exceed a trivial level. The US copyright law "*does not concern itself with trifles*"[134]. The doctrine has long applied to various copyrighted subject matters and only recently its reach was extended also to sound recordings[135]. Therefore, AI-generated music, by analogy with samplings, in some cases may excuse the act of copying, provided that it is not significant in quantity and quality.

[134] Mezei, P. 2017, "De Minimis and Artistic Freedom: Sampling on the Right Track?", *Jagiellonian University Intellectual Property Law Review*, vol. 139, no. 1/2018, footnote 4, p. 2., citing *Ringgold v. Black Entertainment Television Inc.*, 126 F.3d 70 (1997), p. 74.
[135] *See TufAmerica, Inc., v. WB Music Corp., et. al.,* 67 *F.Supp.3d* 590 (2014) and *VMG Salsoul, LLC, v. Madonna Louise Ciccone, et. al*., 824 *F.3d* 871 (2016).

5. PROSPECTS OF MACHINE LEARNING USES IN THE EU

Why robots have to pay for something that humans can do for free?[136]

5.1. TDM copyright exception for the input copying

5.1.1. TDM term and its legal framework

As it became clear above, European copyright law is a system of exceptions and limitations. In the absence of a clear copyright rule that would allow a use of copyrighted materials for a specific purpose, the legality of such uses would be always associated with uncertainty. That is exactly the case with uses of copyrighted works for machine learning purposes in Europe.

The situation, however, is about to change with an ongoing copyright reform. In particular, the reform intends to introduce a copyright exception for text and data mining (TDM) purposes. Since all the copyright discussion within the EU operates the term TDM, it is important to establish a connection between TDM and ML to understand whether and how copyright exception for the first one will affect the second one. Therefore, the research will next discuss the meaning of the TDM term and how it corresponds to ML and then analyse the current legal status of such activities in the EU.

According to the draft Copyright Directive proposed by the European Commission (hereinafter Proposal), TDM is defined as "*any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations*"[137]. However, this definition shall not be misunderstood as limited only to text and data. As the recital 8 of the Proposal explains, TDM generally shall be viewed as "*the automated computational analysis of information in digital form, such as text, sounds, images or data*"[138]. In other words, the subject matter of the TDM can be any copyrighted material.

---

[136] The TDM discourse in Europe often narrows down to the point, that once a lawful access to copyrighted materials is provided, humans can freely analyse that content - the same must apply also when analysis is done by computer.

[137] *Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market*, 14 September 2016, COM(2016) 593 final, 2016/0280 (Text with EEA relevance). Art. 2(2).

[138] *Ibid*., Recital 8.

Furthermore, words construction "*any automated analytical technique*" suggests that there may be many ways to conduct this computational analysis of information. Indeed, the *FutureTDM* project, an organization aiming to foster the use of TDM in the EU[139], provides information about different methods used in TDM analysis[140]. The list by now includes 29 methods, among which are Artificial Intelligence, Machine Learning, Text Mining, Data Mining, Deep Learning, and others.

So as it appears, the Commission defines TDM in a rather broad fashion. It is limited neither to text and data as subject matters, nor it is limited to text mining and data mining as methods of computational analysis. ML shall be regarded as one of the methods of TDM. Therefore, the new exception will be applicable also to acts of analysing materials for the ML purposes.

Although the last statement holds true, the relationship between two terms shall not be viewed as a part and a whole. The TDM term does not cover all activities associated with ML. Neither ML enfolds the TDM term completely. The acts of analysing a vast amount of data lie in the intersection of those two terms and are particularly relevant for this study. ML may go further after the computational analysis and pattern recognition stage to e.g. apply acquired insights into a new context, but that would go beyond the scopes of TDM.

Under such circumstances, it is necessary to analyse the newly proposed TDM exception as a potential legal justification for using copyrighted materials on the input stage in ML. The thesis will next mostly refer to the TDM term accordingly.

It is worth mentioning that the Study on the legal framework of text and data mining[141], conducted on the request of the Commission, suggested referring to "data analysis" rather than to "text and data mining" in a future legislative intervention. It was argued, that the definition of text and data mining must be "*technology-neutral, evolutive and made for changing technologies*"[142].

---

[139] *About the FutureTDM Project*. Available: <www.futuretdm.eu/news/about-futuretdm/> [2018, 25.11].
[140] *TDM Methods*. Available: <www.futuretdm.eu/method-list/> [2018, 25.11].
[141] Triaille, J.et al. 2014, *Study on the legal framework of text and data mining*, European Commission. Available: <www.dx.doi.org/10.2780/1475> [2018, 25.11].
[142] *Ibid*., p. 9.

It was also emphasised that "to mine" means "*to extract data from texts qua informational resources*"[143], while "data analysis" covers a larger number of processes than the mere extraction of data. Furthermore, to mine content means to go deep into the mined subject matter, while some data analysis techniques stay on the surface of that content[144].

It is particularly true in regard to the web crawling technologies that often simply copy the content for caching purposes. As it was illustrated in third chapter, European courts have difficulties with cases involving copying for web caching needs. Therefore, use of the "data analysis" term could go beyond of discovering knowledge and also regulate other legitimate uses of copyrighted content bringing some clarity and consistency into the legal doctrine and European courts' rulings.

It is argued, that the process of analysing information carried in various content is not copyright infringing activity by itself. It is because ideas are not protected by copyright as such[145]. Copyright protects only the creative expression - the form, but not the information contained in that protected expression - the content[146]. With this said, no copyright holder has a right to oppose text and data mining done manually with a pen. Neither should this activity be restricted simply because it is performed by computers and on a much greater scale.

As it was rightly articulated, "*It is a universal truth that once lawful access is granted to a reader of an analogue book or journal they are free to extract information, imagine and innovate. The same must be true for computers in the modern information society*"[147]. This argument is generally referred to as "*the right to read is the right to mine"*[148].

---

[143] Karapapa & Borghi, 2013, p. 47.
[144] Triaille et al., 2014, p. 10.
[145] A general rule prescribed both in TRIPS Agreement (Art. 9(2)) and WIPO Copyright Treaty (Art. 2).
[146] Geiger, C. et al. 2018, "The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects" (March 2, 2018). Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2018-02. Available at SSRN: <www.ssrn.com/abstract=3160586> [2018, 25.11] or <www.dx.doi.org/10.2139/ssrn.3160586> [2018, 25.11], p. 2.
[147] LIBER. *An open letter sent to the Licences for Europe organisers, signed by Nobel prize winners, technology SMEs, research councils, university associations, learned academies, publishers, libraries and law academics*. Available: <www.web.archive.org/web/20130903133142/www.libereurope.eu/sites/default/files/Extract%20from%20email%20sent%20to%20L4E%20TDM%20chairs%20140313_0.pdf> [2018, 25.11].
[148] Moody, G. 2017, *The right to read is the right to mine*. Available: <www.copybuzz.com/editorial/right-read-right-mine/> [2018, 25.11] and Murray-Rust, P. 2012,

In this regard, Professor Martin Senftleben suggests that TDM process must be understood as a mere consulting of a work and, hence, has no copyright relevance[149]. He mentions that even the Commission itself referred to TDM as a process "*through which vast amounts of digital content are read and analysed by machines*"[150]. Therefore, TDM activities should be exempted from rightholders control altogether.

However, in most of the cases, when TDM research is performed by a computer, the information cannot be extracted without a reproduction of protected works and, hence, it may lead to a copyright infringement. Some call it a paradox for a copyright law:

"*On one side, automated processing presupposes the repeated copying of whole works; in this respect, it is an exemplary prima facie case for infringement. On the other side, however, the purpose of this reproduction is to extract information from texts and about texts, an activity that does not normally amount to an infringement in copyright law*".[151]

Furthermore, Copyright is not the only legal regime involved in the process of TDM research. Surely, the reproduction right, prescribed in the Art. 2 of the InfoSoc Directive[152], is the most relevant in the context of copying materials for automated analysis. Besides that, because most of the data used in the TDM would be taken from existing collections rather than copied and OCR-ed one by one from different individual analogue sources, the rules governing legal status of databases come into play as well.

In this respect, the Database Directive[153] provides for two distinct means of legal protection of databases:

- copyright protection of expressiveness of a database as such in case if selection or arrangement of contents constitute the author's own intellectual creation;[154]

The Right to Read Is the Right to Mine. Available: <blog.okfn.org/2012/06/01/the-right-to-read-is-the-right-to-mine/> [2018, 25.11].

[149] Senftleben, M., *EU Copyright Reform and Startups – Shedding Light on Potential Threats in the Political Black Box*. Available: <www.innovatorsact.eu/wp-content/uploads/2017/03/Issues-Paper-Copyright-Directive-2.pdf> [2018, 25.11], p. 9.

[150] *Ibid*., p. 9 citing European Commission, 9 December 2015, Doc. COM(2015) 626 final, p. 7.

[151] Karapapa & Borghi, 2013, op. cit., note 6, p. 51.

[152] The art. 2 of the InfoSoc Directive defines the reproduction right in a very broad fashion: *"the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part."*

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

[154] *Ibid*., Art. 3.

- *sui generis* database right in case of qualitatively and/or quantitatively substantial investment in either the obtaining, verification or presentation of the contents.[155]

The TDM-related activities restricted by above-mentioned rules are a reproduction of database expressiveness (copyright) and an extraction of substantial parts of contents from a database (*sui generis* right). While the reproduction of database expression incorporated in its selections and arrangements would rather be a side effect of copying the contents, an act of extraction is normally necessary to copy a large number of materials. The CJEU noted that an act of extraction takes place when a data is transferred from one medium to another and then integrated into the new medium[156]. In both cases, an authorisation of respecting database owners would be required.

In order to facilitate TDM projects in Europe the proposed Directive offers a mandatory exception to the exclusive rights discussed above, namely:

- Reproduction of copyrighted works in general (InfoSoc Directive art. 2);

- Reproduction of copyrighted expression of databases (Database Directive art. 5 (a));

- Extraction of contents from databases (Database Directive art. 7 (1));

Additionally, the exception would cover also the newly proposed copyright for publishers of press publications concerning their digital uses[157]. However, it would be too soon to include this aspect into the discussion since the proposed right is highly controversial and was met with strong opposition from civil society stakeholders.

It would be a significant step forward if the principle "the right to read is the right to mine" was implemented into the EU copyright law by virtue of the said exception. For that purpose, it is necessary to discuss such possibility by evaluating the compatibility of a broad TDM exception with a three-step test.

The three-step test was implemented into the EU copyright law from the WTO Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS). It

[155] *Ibid*., Art. 7.
[156] *See* e.g. C-203/02, *The British Horseracing Board Ltd and Others v. William Hill Organization Ltd* [2004] EU:C:2004:695.
[157] Art. 11(1) of the Proposal.

requires copyright exceptions to be "*applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightholder*"[158].

In order to understand the meaning of each step, it is necessary to refer to the WTO interpretation of the requirement it had provided in its dispute resolution report in 2000. According to the Panel, a first step "*certain special cases*" requires that a copyright exception must be clearly and narrowly defined as regards its scope and reach. A "clearly defined" should not, however, require identifying each and possible situation explicitly, but at least it shall guarantee a sufficient level of legal certainty. On the other side, limited scope and reach shall be understood as a need of being narrow in a quantitative and qualitative sense[159].

A "*normal exploitation of the work"* within the second step was deemed to include forms of exploitation that currently generate significant or tangible revenue for the rightholder as well as those, which are likely to acquire considerable economic or practical importance in the future. The Panel also clarified that this requirement will be generally met when exempted uses are outlined in a way that precludes them entering into economic competition with non-exempted uses[160].

"*Legitimate interests*" in its turn were defined as an economic value of the exclusive rights conferred by law on their holders[161]. In terms of InfoSoc Directive, it is the right of reproduction (article 2), the right of communication to the public of works and right of making available to the public other subject-matters (article 3) and the distribution right (article 4). While a prohibition of an "*unreasonable prejudice*" shall also imply that a certain amount of prejudice can be justified as reasonable, the Panel did not though elaborate on what degree or level of harm to rightholders' interests would reach a threshold of "reasonableness".

Hence, the given interpretation set a rather narrow space for introducing any copyright exception to exclusive rights. As Rita Matulionyte notes, "*under the three-step test, a use may only be allowed in the absence of any actual, foreseeable or potential harm to the right holder's sources of revenue*"[162]. However, it still remains to answer whether

[158] Art. 5(5) of the InfoSoc Directive.
[159] WTO panel decision DS160, US - s 110(5) Copyright Act, para 6.112.
[160] WTO panel, para 6.180-6.181.
[161] WTO panel, para 6.227.
[162] Matulionyte, R., 2016, p. 61.

licencing out copyrighted materials for the purpose of TDM analysis is a normal exploitation of those works. Then it will become clear also whether rightholders' economic interests in this form of exploitation are legitimate.

According to the study conducted by the Charles River Associates[163], the current market for TDM licences in the EU can be qualified as underdeveloped[164]. Just 1%-2% of all publishers own around 70% of all scientific journals. Some of those few publishers with an important number of journals are not available for mining projects. As a result, licencing possibilities narrow down only to few powerful publishers.

Furthermore, mining requests are normally considered on a case-by-case basis. Publishers are wary of TDM that may result in derivative information products. As a result, they would require a description of each mining project. Furthermore, while mining for a non-commercial purpose by research institutions would be normally approved free of charge, commercial miners would need to pay a licencing fee[165].

Under those circumstances, it can be concluded, that although it is happening on a little scale nowadays, offering copyrighted content for the purpose of text and data mining is very likely to acquire a significant economic importance in the future. Therefore, it may amount to normal exploitation of works for publishers and other aggregators of large collections of content.

It must be kept in mind, however, that this kind of business practice is relevant only in regard to database owners and not individual authors. Only aggregators of big collections of copyrighted works are likely to engage in offering their content for mining purposes and hence would have legitimate interests in remuneration. Authors *a priori* are excluded from beneficiaries of the mining use because of the limited number of works they could offer.

Furthermore, not all kinds of copyrighted content are present on the TDM licencing market nowadays. It is only scientific, technical and medical publications that were discussed within the copyright reform. It could be claimed, that images also fall into the range of mining interests. However, it is unclear whether fictional literature, music, and videos would enter that market in the future.

---

[163] Boulanger, J. et al. 2014, *Assessing the economic impacts of adapting certain limitations and exceptions to copyright and related rights in the EU : analysis of specific policy options,* European Commission. Available: <www.dx.doi.org/10.2780/20222> [2018, 25.11].
[164] *Ibid*., p. 62.
[165] *Ibid*., p. 64.

As regards legitimate interests of rightholders, it is undoubtedly true, that database owners should have a right to extract economic value from their content by means of licencing its use for mining purposes. It is not much different, however, from granting access to their content in a traditional way.

Nevertheless, according to the concept "the right to read is the right to mine", rightholders' interests to charge a fee for mining content, to which users already have access, cannot be deemed legitimate. It is argued, that when a lawful access to content was granted, nothing shall restrict a user from mining that content. In other words, it is not legitimate to charge first for reading articles and then charge a second time for mining the same articles.

The factual conditions of mining practices are not that straightforward though. In order to conduct a TDM project, researchers often need to download a vast amount of content. That can significantly increase a load on database servers and affect the stability of the whole system. It may lead to situations when traditional users will face access problems because of few miners extracting the contents at the same time.

To tackle this issue, some publishers decided to invest into a separate network infrastructure to specifically facilitate TDM downloads. It is those investments that rightholders should have legitimate interests to recoup. Here is the main question deriving from this: how to accommodate the freedom to mine with rightholders' interests to return investments they make in connection with TDM?

The author of this thesis is of opinion, that everyone shall have a right to analyse the content, to which they already have lawful access. That must be recognized as a basic principle in copyright law. That right, however, shall not confer any obligation on rightholders to bear additional costs in connection with simplifying content downloads. In other words, some restrictions in terms of speed or amount of downloads may be justified by the need to preserve the stability of the database platform. It shall not be viewed, though, as copyright enforcement measures, but rather as legitimate restrictions of a legitimate freedom.

In case, when rightholders make mining-related investments to facilitate the process, they simply should price respecting costs into the access fee. It may lead to the creation of separate purpose-specific mining licences. Again, higher access fees would be

justified by legitimate interests to recoup investments and not by the requirement of separate authorisation for TDM.

The situation would be similar, although not identical, to airport fast lanes. Everyone with a valid ticket can proceed to the security control. However, you can pay extra to get on your plane faster. Similarly, everyone with a lawful access may mine the content, but you may wish to pay extra to expedite the process and get your results sooner.

On the negative side, the proposed solution would have its shortcoming in forcing researchers to move to facilitated TDM platforms with higher prices in a long term. It would essentially mean that users with a traditional access would not be able, in fact, to use that content for computational analysis effectively, although not explicitly prohibited.

On the positive side, it would be recognized that the right to read is also the right to mine, what would bring about more fundamental benefits. It would add significantly to the legal certainty with respect to TDM uses. In addition, in situations, when access is not restricted and rightholders are not pursuing to explore the market of TDM licensing, users would be able to fully exercise their freedom to mine, although some precaution measures against abuses would be required.

Furthermore, as it was mentioned before, some reasonable prejudice can be acceptable in respect of rightholders' legitimate interests. Therefore, adopting a broad TDM copyright exception prohibiting rightholders from a discrimination of users based on a subscription model has a room for discussion. Simply speaking, it is worth considering an option of letting everyone with a lawful access to use publishers' mining facilities without any extra fee or change in a subscription model. That would deprive rightholders a right to demand compensation for their TDM-related investments if they made any.

In this respect, some commentators refer to the third step as a "proportionality test"[166], which is often used by legislators and courts to balance conflicting rights. On the one hand, there is the right to property and its copyright component and, on the other hand, there is the freedom of the arts and science.

[166] Senftleben, M. 2004, Copyright, Limitations and the Three-Step Test: An Analysis of the Three-Step Test in International and EC Copyright Law, Kluwer Law International, p. 226.

Applying the proportionality test to conflicting rights, the CJEU would inter alia assess potential harm to rightholders as opposed to public benefits of the exception. It must be remembered, that the only losses, that should be considered here, are the ones that derive from a need to make TDM-related investments. However, there are reasonable grounds to claim that the damage would be minimal and easily outweighed by public interests.

First, the Commission admitted in its Impact Assessment, that publishers tend to regularly include TDM in the subscription licences without increasing licensing fees significantly. It suggests that TDM does not have a significant extra value in the context of current subscription licences[167].

Second, the popularity of TDM researches is increasing. That means that databases adapted to this kind of use would attract more users in the future. The increase in the number of subscribers even without a significant rise in licencing fees would eventually generate higher profits and compensate investments into a mining infrastructure.

There is no need to discuss the public utility of the freedom to mine at this stage to conclude, that there is a clear way to adopt a broad copyright exception allowing everyone with a lawful access to databases to analyse their contents. To put it differently, there are reasonable grounds supporting compliance of the broad TDM exception with the three-step test requirements.

5.1.2. Scopes of the proposed exception

In the context of TDM, the explanatory memorandum of the Proposal declares an objective to facilitate digital uses of protected content for the purpose of scientific research. This is a foundational limitation of the new exception that deserves a separate scrutiny before delving into the details.

Essentially, there may be many reasons to analyse content and scientific research is only one of them. According to the principle "the right to read is the right to mine", the mining activity *per se* shall have no copyright relevance. It means that it must be allowed to undertake a computational analysis of materials to which one has lawful access regardless of the purpose. Therefore, the Commission's approach to introducing

[167] European Commission (2016), Commission Staff Working Document, *Impact Assessment on the modernisation of EU copyright rules*, 14 September 2016, SWD(2016) 301 final, Part 1/3, p. 114.

the TDM exception only for the purpose of scientific research to a great extent undermines the said principle.

Further, the Article 3 introduces copyright exception for text and data mining only when made by research organizations[168]. Such organizations include universities, research institutes or any other organizations, the primary goal of which is to conduct scientific research (may also provide educational services). Moreover, such organizations must either operate on a not-for-profit basis or pursuant to a public interest mission[169].

Hence the Commission limited the proposed exception not only by the purpose of scientific research but also by a kind of beneficiaries: only research organizations and only non-profit. It is interesting to realise, that this approach is somewhat different from what was already adopted in some Member States. For instance, UK adopted a TDM exception for a non-commercial research but for everyone with a lawful access to the content. Some other Member States later, with minor deviations, followed the same approach[170].

Arguably, the proposed rule would amount to the same non-commercial use and even be more restrictive. It is because "*a "non-commercial" limitation would allow a business acting for non-commercial purposes (e.g. research, criticisms, news reporting, etc.) to benefit from the exception, something that is not possible under Art. 3*"[171]

On the other hand, there are many borderline cases, where uses cannot be unequivocally classified between commercial and non-commercial. Thus the Commission chose to limit beneficiaries by their status rather than by the character of the use, what is meant to make the applicability of the exception more clear.

In the explanation of this policy option, the Commission focused solely on the licencing market of TDM uses exploited by STM publishers. The major demand in this market is presented by big life-science companies and not-for-profit research institutions. It was argued, that a broader exception benefiting commercial companies would deprive

[168] Art 3(1) of the Proposal.
[169] *Ibid*., Art. 2(1).
[170] For example Estonia (2017), France (2016), Germany (2017) opted for a non-commercial character of use but did not discriminate beneficiaries in their corresponding TDM copyright exceptions. *See* Geiger et al., 2018, pp. 17-18.
[171] Margoni, T. 2018, April 25, - last update*, The Text and Data Mining exception in the Proposal for a Directive on Copyright in the Digital Single Market: Why it is not what EU copyright law needs*. Available: <www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/> [2018, 25.11].

rightholders their ability to subject TDM research to their licensing agreements and as a result lower the value of those contracts.

In fact, the Commission admitted the ability of STM publishers to compensate potential losses in revenue by increasing licencing fees by virtue of the lawful access rule. However, it simply failed to see "whether and to what extent they would manage to do so"[172].

Consequently, it concluded that a broader copyright exception would require rightholders to renegotiate their business agreements with commercial customers resulting in high compliance costs. That was seemed to the Commission as an unjustified intervention into a purely commercial market for TDM, especially in a view that those commercial players had "*generally not asked EU intervention in this area*"[173].

This approach of the Commission deserves some substantial criticism. Firstly, it is principally wrong to refuse to recognise that the right to read is also the right to mine. Many legal academics on numerous occasions strongly recommended that the TDM exception under the Article 3 should apply to all lawful usage regardless of the purpose and character of the use[174]. As minimum it seems to be irresponsible to ignore recommendations of the European scientific community in matters of its competence.

Secondly, while limiting the objectives of the copyright intervention only by the need to facilitate scientific research, it would be times more progressive to focus only on these purpose-specific uses. It can be understood, that the Commission aimed to prevent commercial exploitation of TDM uses that add little if any to the state of science. For that reason, a TDM exception for a research purpose would be sufficient limitation. In this case, a bigger number of researchers such as entrepreneurs, individual post-graduate students, journalists and others could engage in TDM.

Thirdly, excluding commercial companies from beneficiaries appears to be also wrong. This commerciality factor is similar to the one discussed as part of the fair use doctrine in the US. There, in many cases, it was argued that commercial motivation cannot

---

[172] Commission (2016), *supra* note 168, Part. 1/3, p. 118.
[173] *Ibid*., pp. 116, 117.
[174] *See* CREATE. *Statement by EPIP Academics to Members of the European Parliament in advance of the Plenary Vote on the Copyright Directive on 12 September 2018.* Available: <www.create.ac.uk/wp-content/uploads/2018/09/Statement-by-EPIP-Academics.pdf> [2018, 25.11].

undermine the claim of fair use when the use itself is highly transformative[175]. Indeed, most of the universally accepted forms of transformative uses such as news reporting, quotation, and parody are normally done commercially for profit[176]. Moreover, commercial companies, start-ups, and SMEs are more likely to invest in research and innovation accelerating the progress of science and technology.

As regards the commercial market of rightholders, it was discussed in the previous section of this chapter, that rightholders' legitimate interests only narrow down to mining-related investments. It was explained, that those losses could be compensated from either increase in licencing fees or in a number of subscribers, or both altogether. Therefore, the harm to rightholders from a broader exception would be insignificant if any.

Moreover, the need to preserve a purely commercial market for TDM was probably the main argument for the chosen policy option. The intervention into this market was deemed by the Commission as unjustified. However, how can this argument apply to instances of mining content from databases that have no licencing models? For example content crawled from the Internet or social networks. How can Commission justify allowing only non-profit research institutions to mine freely available web content or social networks data, while this activity normally would not harm corresponding rightholders if conducted by any interested party?

Fourth, even non-commercial scientific TDM research still could be interesting for a broader number of parties than only research institutions. For example journalists, independent researchers, post-graduate students, and others would have same legitimate interests to engage in TDM, especially in a view of the availability of affordable TDM technologies. Therefore, having an objective to promote non-commercial TDM uses for the purpose of scientific research, it seems slightly discriminatory to exclude mentioned individuals from the list of beneficiaries.

Next, it is interesting to refer to a real-case TDM research that took place in the US and assess whether it would be lawful if conducted in the EU. How would proposed copyright exception apply in this case? Would it solve the question of legal certainty over reproductions that took place?

[175] *Supra* note 84.
[176] *Supra* note 86.

The study was done in 2017 at Cornell University[177]. Researchers copied over 100 million pictures from Instagram and used ML algorithms in order to identify patterns in how clothing styles vary around the world. The results were published along with some samples of pictures for an illustration purpose.

Would it happen in the EU, the University would benefit from the proposed copyright exception because it is a non-profit research institution envisioned by the Directive. However, the exception does not allow reproductions of the content used in datasets in the final research results, precluding hence any communication to the public of the mined copyrighted subject matter.

In this respect, some argue, that it is irrelevant to extend the exception to the right of communication to the public, because they are unlikely to take place: "*In most cases, while being the result of the data mining process, the report will not contain or display any of the data that have been "mined*"[178]. This practical example from Cornell University proves them wrong. To be more specific, it shows that mining results would sometimes lead to reproduction of mining materials, although as a part of a scientific research paper.

Technically speaking, the Art. 5(3) (a) of the InfoSoc Directive could be applied in this situation. It allows for reproduction and communication to the public of copyrighted works "*for the sole purpose of illustration for ... scientific research*"[179]. Although it requires to indicate the source, including the author's name, this may be neglected if turns out to be impossible.

On the other hand, this copyright exception is not mandatory and its implementation may vary throughout the Member States. In addition, for the purpose of verifiability of TDM research results, researchers might need sometimes to store source materials and probably also communicate them at least inside the research community[180]. This activity would trigger the right of communication to the public, but will not benefit from the mentioned "illustration" exception.

---

[177] Matzen, K. et al. 2017, "StreetStyle: Exploring world-wide clothing styles from millions of photos", *arXiv*. Available: <*www.arxiv.org/abs/1706.01869*> [2018, 25.11].
[178] Boulanger et al., 2014, p. 28. *See* also Geiger et al., 2018, p. 7: "*the TDM output should not infringe any exclusive rights as it merely reports on the results of the TDM quantitative analysis, typically not including parts or extracts of the mined materials.*"
[179] Art. 5(3) (a) of the InfoSoc Directive.
[180] Geiger et al., 2018, p. 7.

Therefore, it is recommended, that the new copyright exception covers not only initial reproductions and extractions needed to obtain materials to be mined but also consecutive activities including redistribution and communication to the public of the corresponding research results. That would apply only to TDM activities and only in relation to the original materials necessary for the purpose of TDM research. It would be similar to parody or quotation exceptions, where "*the original work is redistributed but only as part of the parody or quotation*"[181].

Another hurdle that this kind of research could face in the EU is the new personal data protection law. Because the study was conducted on images depicting real people, the further publication of those pictures triggers on the right to privacy. In fact, even not hypothetical but this particular research may risk violating the GDPR although it was done outside the EU. Provided of course, that pictures used for illustration belong to EU citizens, which is very likely to be, considering the fact that Instagram does not know borders and researchers were targeting clothing patterns from around the world, including Europe.

Consequently, the Cornell University research could be possible in the EU under the proposed TDM exception. Some uncertainty is still present though in regard to published samples of pictures in the final research results. They could potentially benefit from the InfoSoc "illustration" copyright exception but could face legal scrutiny under the GDPR. This latter aspect would fall outside the scopes of this study though.

At the same time, the research illustrated here could be of high interest also for commercial players to conduct e.g. a market research with a purpose to identify clothing, food, cosmetics habits across different countries, different age groups of people etc. For that purpose, it would also suffice to utilize a great source of Instagram images. However, the proposed copyright exception would not cover this activity only because it is performed by a commercial entity.

In addition, pursuant to the Recital 10 of the Proposal, research organizations should also benefit from the exception when they engage in public-private partnerships (PPP). This potentially opens a way for commercial for-profit companies to benefit from the TDM copyright exception indirectly.

[181] Margoni, T., 2018.

However, it is questionable to what extent this tool would be useful for private companies since most of the research and development projects require a high degree of secrecy due to market competition reasons. The risks associated with collaboration with a third party might exceed the benefits of relying on the mentioned copyright exception. Therefore, in cases that involve some important innovation like developing creative AI, PPP may not be an appropriate solution to avoid a copyright infringement.

And last but not least, on 12 September 2018, the European Parliament (EP) adopted a position on the Copyright Directive[182]. In terms of the TDM exception, the EP suggested to include an optional copyright exception that would encourage innovation also in the private sector. It would be possible, however, only when "*the use of works and other subject matter ... has not been expressly reserved by their rightholders, including by machine readable means*"[183].

Under this EP proposal, everyone with lawful access would benefit from the TDM exception only in the absence of corresponding licencing schemes. In other words, the mining authorisation would not be required when rightholders do not expressly offer such service as part of their business. It looks like a limited version of "the right to read is the right to mine" principle.

Albeit limited by the rightholders' business model, this exception could become a viable solution for an innovation in the private sector. However, this solution would come at the expense of harmonisation of the EU copyright law, leading to legal uncertainty in cross-border uses. For example, a company in a country with broader copyright exception would have difficulties with mining content originating from a country that only adopted mandatory TDM exception in its narrow scopes. It may also lead to an unfortunate situation of having a different technological environment in the different Member States. Clearly, such optional character of the said exception would go contrary to the main objectives of creating a digital single market.

The latest suggestion from the Council as a part of trilogue negotiations would limit this optional exception even further. In particular, it offers to allow only temporary

---

[182] European Parliament*, Amendments adopted by the European Parliament on 12 September 2018 on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market (COM(2016)0593 – C8-0383/2016 – 2016/0280(COD))* [Homepage of European Parliament]. Available: <www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P8-TA-2018-0337+0+DOC+XML+V0//EN#def_1_1> [2018, 25.11].
[183] *Ibid*., Art. 3(a).

reproductions and extractions of the works used in a TDM process[184]. In contrast, non-profit research organizations as beneficiaries of the mandatory exception may retain copies for an undefined period of time in a view of the need to verify their research results. It does not appear reasonable to discriminate private actors in this respect.

It can be assumed, that the Council is aiming to prevent private entities from creating a secondary market of those copies used in a course of TDM. However, that outcome is not likely due to another requirement not to use those copies for purposes other than TDM[185]. Moreover, reselling copies *per se* would amount to recommunication of original works or their adaptations and therefore would infringe on copyright.

As the LIBER organization suggested in its open letter to EU policymakers, the "temporary" requirement would go contrary with the realities of data analytics and verification of research results. It may deter researchers from making investments into TDM if they only could retain working materials on a temporary basis[186].

Finally, part of the general objective of the EU copyright reform is "*to promote digital innovation and to foster the international competitiveness of European research"*[187]. It is hard to see though, how this could be effectively achieved only with help of research organizations under such a narrow copyright exception. The Commission itself recognised in its Impact Assessment that "*TDM is still a nascent tool in the non-business sector*"[188], i.e. universities or research institutes.

On the other hand, private commercial companies normally would have resources and incentives to invest in TDM research in Europe that might eventually drive innovativeness and bring EU on the level playing field with the US, Canada, Japan, and other countries, which adopted less restrictive TDM exceptions. Disfavouring this key group of stakeholders could cause an opposite effect like a "brain drain" and outflow of investments to other jurisdictions. That seems like going contrary to what the EU is trying to achieve with its copyright reform.

---

[184] Julia Reda 2018, Oct 25, *Second round of trilogue negotiations. The latest compromise proposal*. Available at: <www.juliareda.eu/wp-content/uploads/2018/10/Copyright-Directive_4-column-document_ARTICLES-v2-23102018.pdf> [2018, 25.11].
[185] European Parliament 2018, *supra* note 184, par. 2 of the art. 3(a).
[186] LIBER. *Europe Needs A Broad & Mandatory TDM Exception*. Available: <www.libereurope.eu/blog/2018/11/13/europe-needs-a-broad-mandatory-tdm-exception/> [2018, 25.11].
[187] Commission (2016), *supra* note 168, Part. 1/3, p. 82.
[188] *Ibid*., p. 104.

5.2. Legal grounds for output reproductions

Unlike the process of analysis, reproduction in the outcome is not envisioned by the proposed TDM exception and was only addressed by some commentators as mentioned above. However, the primary interest of this section is to discuss the possible reproduction of original works in a secondary work generated by a creative AI model. This is a kind of reproduction that was deemed to be unlikely by various commentators on the TDM reform and it will be explained next why.

The TDM process itself is a simple process of analysis and extraction. ML, on the other hand, would include this stage of TDM and would go further afterward. The process of creating a new work is distinct from TDM although is based on it. Unlike the report on TDM findings similar to one described in the study of clothing patterns, a creation of a new work is the next step to it and utilises those findings in a new way.

A good example is an AI model generating poetry from images that was discussed in the fourth chapter. The process includes initial mining of the training datasets - the process envisioned by the EU TDM exception and is followed by a creation of new works, which goes beyond the EU copyright reform.

Consequently, it is important to find possible legal rules that would govern this kind of reproductions in case if they take place to the extent of copyright infringement. That is another important question: how much copying is acceptable if any? Some recent decisions of national courts, as well as CJEU, may offer us some insights in this respect.

In its well-known *Infopaq* decision, the CJEU specified that even as little as 11 words of original author's expression are protected by copyright and unauthorised reproduction would lead to an infringement. To be more specific, as soon as a portion of work constitutes author's own intellectual creation, which is for national courts in each individual case to determine, it will meet requirements of originality and, thus, copying of the said portion would amount to the "reproduction in part" within the meaning of the Art. 2 of the InfoSoc Directive[189].

Under such interpretation, first, a reproduction of some portion of copyrighted work would not infringe on copyright unless that portion meets requirements of originality. Second, normally the EU Copyright law does not offer a threshold of infringement

[189] C-5/08 *Infopaq International A/S v. Danske Dagblades Forening* [2009] EU:C:2009:465 § 51.

when original expression is reproduced in another work. As soon as it is original, an authorisation would be required.

Nevertheless, fortunately, the EU offers some space for an effective defence case. First of all, it is worth considering a concept of independent creation. It is not copyright infringing to independently create a work even identical to already existing one. "*Only the very act of borrowing infringes on the right of reproduction, not similarity per se*"[190].

It may be argued, that the process of generating a new work by an AI model is not based on copying, but rather is a new original creation. "*It is not a trivial combination of the data which was fed into the machine learning system*", AI generates results that never existed before[191]. ML would "break" original works into single units, which individually taken would not be covered by copyright. After that, it would utilise those single elements and, following certain patterns, compile a new work.

On the other side, a plaintiff may argue that similarities between AI generated output and her original work would not happen if the work was not used in the training dataset in the first place. In other words, it is possible to claim and prove a causal link between prior mining and later copying of original expressions.

It remains to wait and see what stance courts will take in such situations. It must be remembered, that AI lacks self-awareness and thus is unable to explain the rationale behind its results. Therefore, it would be difficult if not impossible to determine whether it simply copied an original expression at issue or independently created it.

If the concept of independent creation would not apply to AI reproductions, it is still possible to rely on the existing copyright exceptions as a valid defence tactic. Depending on each individual case, even the parody exception could be invoked in some situations, provided of course that parody requirements are met in the secondary work[192]. However, the most interesting with respect to creative AI and less discussed in scientific community copyright exception concerns *incidental uses* of copyrighted works.

---

[190] Cabay, J. & Lambrecht, M. 2015, "Remix prohibited: how rigid EU copyright laws inhibit creativity", *Journal of Intellectual Property Law & Practice,* vol. 10, no. 5, p. 8.
[191] Thoma, M., 2016, p. 1.
[192] For more details on the CJEU interpretation of the parody exception refer to its groundbreaking decision on *Deckmyn* case (C-201/13 *Johan Deckmyn v. Helena Vandersteen and others* [2014] EU:C:2014:2132).

The said exception allows for "*incidental inclusion of a work or other subject-matter in other material*"[193]. It used to be a part of national copyright laws long before the InfoSoc Directive. The exception was normally meant to cover situations when copyright protected works appear in photographs and films rather by chance than intentionally and are of secondary importance to the main work. The scopes of it, however, may vary significantly in the different Member States.

A rather broad formulation of the *incidental inclusion* exception in the InfoSoc Directive suggests that it could theoretically apply to new circumstances like the one with AI-generated works. One could possibly argue that an infringing outcome generated by an AI model was not intended to happen and is a pure incident. A contested original expression could be *incidentally included* in the compilation process. It can also be easy to claim that an infringing portion is of secondary importance to the whole new work.

Similarly to the concept of independent creation, such argumentation could be countered with a claim that the initial deliberate use of the original work in training datasets undermines its incidental character. Accordingly, it is not easy to analyse the rationale behind an AI outcome and courts may face serious problems in their assessments.

However, the said exception could have its stronger position in cases where a plaintiff's work was not included in the mining process. As it was discussed before, there is a possibility that an input would be based on another work because of the cumulativeness of creativity. In the case of literary works, it would normally happen through quotation, illustration etc. As the result, it would be easier to argue *incidental inclusion* of that contested original portion of expression as it was not intentional.

Further, the nature of AI-generated works could be compared with another type of transformative creations that are currently unsettled under the EU Copyright law - samplings. In music, sampling is an act of taking a portion or entire sound recording, known as a sample, and reusing it as a part of a new sound recording[194]. The need to refer to samplings is particularly relevant in cases of AI-generated music.

---

[193] Art. 5(3)(i) of the InfoSoc Directive.

[194] Wikipedia. *Sampling (music)*. Available: <en.wikipedia.org/wiki/Sampling_(music)> [2018, 25.11].

Normally, the EU copyright offers little flexibility in regard to such creative activity. However, a recent copyright reference from the German Federal Court of Justice to the CJEU in the case *Metall auf Metall III*[195] may alter this situation for better.

The *Metall auf Metall* is an old German case in which a musician Moses Pelham took a two-second sample from a song "Metall auf Metall" by a German band Kraftwerk and used it in the song "Nur mir" performed by Sabrina Setlur. The case eventually was heard by the German Constitutional Court which came to the conclusion, that "*if the artist's freedom of creative expression is measured against an interference with the right of phonogram producers that only slightly limits the possibilities of exploitation, the exploitation interests of the phonogram producer may have to cede in favour of artistic dialogue*"[196].

The key point implied from this decision is that the artistic freedom shall include a right to use parts of other authors' works to the extent that does not harm economic interests of those authors. It was also emphasised, that the mere availability of licensing possibilities cannot adequately fulfil the freedom of artistic expression because "*a right to be granted a license to use the sample does not exist*"[197]. The rightholder can unreasonably deny licensing without any explanation.

Following that decision the German Federal Court of Justice requested the CJEU, among other things, to provide a guidance on how to balance conflicting fundamental rights, namely the right of copyright protection[198] and the freedom of arts[199]. It is particularly important for the industry of transformative uses like sampling, mash-up, collage, etc. By analogy, the CJEU decision could also apply to AI generated works. However, it would be possible only if a connection between fundamental human rights and AI creativity was established.

The aforementioned task may not be that straightforward. In terms of a human creativity, the freedom of the arts can be easily invoked as a basis for interference with

[195] Vorlage des Bundesgerichtshofs an den Europäischen Gerichtshofs zur Zulässigkeit des Tonträger-Samplings. Available: <juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=pm&Datum=2017&Sort=3&nr=78496&pos=1&anz=87> [2018, 25.11].
[196] BVerfG press Release 2016, May 31, No. 29/2016. *The use of samples for artistic purposes may justify an interference with copyrights and related rights*. Available: <www.bundesverfassungsgericht.de/SharedDocs/Pressemitteilungen/EN/2016/bvg16-029.html> [2018, 25.11].
[197] *Ibid*.
[198] Art. 17(2) of the EU Charter.
[199] Art. 13 of the EU Charter.

someone's copyright. However, when we employ robots to take care of the whole creative process, there is not much left from our artistic freedom. Rephrasing words of Grimmelmann and applying them in a slightly different context, when we as people take part in non-human creativity, we suspend our human capacities[200]. Consequently, we cannot rely on fundamental human rights and there is simply no such thing as a freedom of robotic arts.

On the other side, it is not yet clear what would be a status of human involvement in AI creativity. Even if the work was generated completely by an AI model, there still has to be a human behind that process. Someone has to do some necessary arrangements whatever trivial they might be like choosing a style or colour density. Someone simply has to press a button.

What if the fundamental freedom of arts gets interpreted in a very broad way to cover also instances of creating art by means of AI with a very limited personal involvement of a human being? That is not so illusory possibility especially in a view of the doctrine "*sweat of the brow*", according to which a person may get his work protected even if it is not original or does not involve substantial creativity.

This perception of an AI as a tool in human hands looks clearer when AI is used in a different from the arts field, e.g. news reporting or online translation. It can be argued, that expressive AI is a simple tool of human engagement with information. The freedom of information does not require any level of human intellectual creativity. Therefore a user of an expressive AI model may still invoke the right to receive and impart information and ideas as it is guaranteed both by the EU Charter and the ECHR.

As a result, fundamental human rights could serve some defensive ground in case of reproductions in the outcome. However, it is uncertain at this point whether the proportionality requirements would be met in such cases.

[200] Grimmelmann, J. 2016, p. 667. The original citation is "*When we talk about nonexpressive uses, we should perhaps refer to them by another name: non-human uses. When we as people take part in these uses, we suspend our human capacities.*"

6. CONCLUSIONS

"When someone builds a new technology and people love it, the legal system will evolve to allow it." (by Randal C. Picker)[201]

The present research has delved into peculiarities of AI expansion to the creative industry. It would be fair to view this invasion as another chapter of the long-lasting story of an interaction between digital technologies and copyright laws. Accordingly, this piece of study is meant to contribute to the discussion that already started over the applicability of old laws to new uses of protected works. It also prompted a critical analysis of the ongoing EU Copyright reform.

The appearance of non-display uses on the copyright regulated stage was not sudden. More conventional technological applications such as web caching technologies, books digitization projects and others preceded its coming. Fortunately, to some extent, they paved a way for the present enquiry with a solid argumentation developed throughout judicial debates.

On the west shore of Atlantic, those technological uses successfully tested a reach of the fair use doctrine proving its right to exploit copyright protected works when it does not harm rightholders. In Europe, the same technologies were met with a traditional hostility of national copyrights. While some countries refused to tolerate unauthorised reproductions regardless of any justifications, others practiced their agility in finding some basic principles of law to protect legitimate technologies.

In particular, the US case law demonstrated that computer processing of works even on a large scale may be found lawful if it does not lead to a creation of substitutes. The requirement of transformativeness had its clear application as a functionality test: it may be lawful to make a copy when a secondary work performs a different function than that of the original, regardless of any similarities between two of them. Hence, copies made for the ML training process shall be viewed exactly in this fashion: they are not used as substitutes and perform a completely different function.

[201] Picker, R. C. 2015, "Internet Giants: The Law and Economics of Media Platforms", The University of Chicago, *Coursera online subject,* lecture notes, viewed Jan 2018. Available: <www.coursera.org/learn/internetgiants> [2018, 25.11]. In the context of *Sony Betamax* case.

Furthermore, the very nature of ML uses fit very well into the definition of transformativeness developed by judge Leval: they use copyrighted works as raw material and then transform them into new information, new expressive works. Thus it seems to be a type of use that the fair use doctrine is meant to support.

The analysis of the market of works for ML uses demonstrated a lack of significant danger for rightholders' interests arising from unauthorised uses of their works. Authors typically do not create works to be used in ML training datasets. And the market of database content for ML purposes is still underdeveloped.

The study also discussed a potential outcome that expressive AI may disrupt a creative industry and oust human authors from their traditional markets. Regardless of how significant it may happen to be, any negative impact on the human creative industry would derive from an activity not restricted by copyright. It is viewed by this paper as a logical consequence of "AI-ization" of different aspects of human life.

There appear to be no meaningful issues with possible reproduction of original human expressiveness in AI-generated outcomes. They are not likely to occur in the first place. But should they happen and exceed a trivial level they might still find protection under the fair use. In other cases, the US copyright case law has developed a number of rules on how to assess potential copyright infringement. Although AI creativity may lead to numerous copyright trials in a near future that would possibly receive media headlines like "Humans vs AI", things are more or less certain in that respect.

As can be seen, ML and expressive AI seem to receive a green light from a copyright law in the US. It becomes possible mainly owing to a flexible and technologically-friendly doctrine of fair use. It would be a consistent follow-up approach after a range of similar decisions on other technological uses.

In Europe, the InfoSoc Directive became partially outdated as soon as the new computer uses, understandably unforeseen by the lawmakers at that time, became a common practice for global Internet search engines. Quite ironically, it happened no later than the Directive was implemented by the last EU Member State - in the same year of 2006.

It is then not surprising that ML copies of works would hardly find any protection from the reach of exclusive rights of copyright holders. Copyright laws must be reconsidered

every time "*when technology renders the assumptions on which they were based outmoded*"[202].

A close study of the ongoing copyright reform in Europe demonstrated that ML uses fit well into the category of Text and Data Mining. Therefore a future of research into the field of expressive AI depends now very much on the scopes of the copyright exception that will be adopted soon. It has been discussed in this research that the mining activity *a priori* has no copyright relevance as facts and ideas are not protected. Therefore it is suggested that the freedom to read must be also the freedom to mine.

The art of making copyright laws is far from ideal though. The main stakeholders would unlikely agree to the law that would leave them worse off compared to what they currently receive. A broad copyright exception would go contrary with interests of STM publishers that get revenue from life-science companies mining their content. That was described as a main concern of the Commission when choosing a policy option.

However, this thesis demonstrated that a broad approach to exempting mining activity from a reach of copyright would likely comply with the three-step test. The more fundamental benefits of the said exception would outweigh any potential losses to legitimate interests of database owners. Therefore it appears that there are no other constraints to ensure a mining freedom except a political will of EU legislative institutions.

The latest trilogue consultations between the Council, the European Parliament, and the Commission suggest very limited possibilities for ML and expressive AI in Europe. It has been offered to adopt a mandatory copyright exception for TDM conducted only by non-profit research institutions and only for scientific research purposes. The Member States would be offered an optional right also to extend such an exception to benefit private actors but only in relation to content not exploited in this regard by its rightholders.

It remains to wait and see what would be a final result of the compromise, however at this point it appears certain, that ML and AI developments in Europe would be restricted in terms of getting training materials. As Jessica Litman rightly observed, "*the institutional and legal structure of the copyright community makes it difficult to prevent*

[202] Litman, J. 2006, Digital copyright, 2nd ed. edn, Prometheus Books, Amherst, N.Y. p. 22.

foolish approaches to new technology"[203]. Such a regrettable outcome would signify an unwillingness of the EU policymakers to accept the fact that old analogue rules may not serve the same balancing function in a new digital world.

It would not come by surprise though, as the history of EU copyright harmonisation often prioritised the interests of industry over users' rights despite alerting messages from the academic community[204]. As Rita Matulionyte observed, "*in the EU, legislators often seem to assume that creativity and innovation are encouraged by granting increasingly broad and exclusive rights to creators and industries*"[205]. They must be wondering then why most of the world-spread technologies originate from the US or other countries.

As regards possible claims of copyright infringements as pertaining to AI-generated works, national copyrights still may offer some room for effective defence strategy. Besides limited applicability of the concept of independent creation and some copyright exceptions like the one covering incidental inclusions, it is more interesting to see a potential impact of fundamental human rights in this respect. In particular, the freedom of science and arts, as well as the freedom of information may serve a balancing role against the monopoly of copyright holders. The CJEU still has to say its word in this respect.

By and large, making copies of works for the purpose of ML would normally not result in outright copyright infringement either in the US or in Europe. Emerging technologies need access to qualitative content and they are likely to receive it. Although in Europe, traditionally high level of protection of copyright would place more restrictions upon the use of protected works.

[203] *Ibid.*
[204] *See* in general Farrand, B. 2014, Networks of Power in Digital Copyright Law and Policy, Routledge Ltd, London.
[205] Matulionyte, R., 2016, p. 54.