The Fourth Industrial Revolution

The Functions and Impacts of Copyright and Related Rights on Artificial Intelligence and Machine Learning (ML)

BENJAMIN RICHARD JAMES WHITE

Doctor of Philosophy

Centre for Intellectual Property Policy & Management
Faculty of Business and Law
Bournemouth University

December 2025

Copyright Statement

Abstract

This thesis addresses the question of the extent to which the contours of copyright law, and in Europe *sui generis* database rights, affect the adoption of machine learning in four different jurisdictions – namely the EU, Japan, the United Kingdom and the United States of America. Furthermore, it explores how the answer to this question has influenced recent e-commerce developments in international trade law.

It approaches this analysis through the traditional framework of copyright law and related rights by evaluating the application of both permitted and restricted acts. First, the thesis examines how these four jurisdictions frame limitations and exceptions to exclusive rights that facilitate data analysis undertaken on copyrighted works. After this, existing legal doctrines that apply to exclusive rights in traditional computer programs are outlined before evaluating their applicability to AI model software.

The research employs doctrinal and comparative research methodologies grounded in an economic and innovation policy perspective on law. Such an approach serves to highlight not only the strengths and weaknesses of data analysis exceptions across the four jurisdictions examined, but how copyright law contributes to both economic imbalances in AI markets and recent geo-political developments in Preferential Trade Agreements and at the WTO.

This research makes an original contribution to knowledge through the doctrinal and comparative legal analysis adopted, including *inter alia* an in-depth analysis of Japanese copyright law's regulation of AI, and a doctrinal exploration of the extent to which copyright and *sui generis* database rights inhere in a range of different trained AI models, including but not limited to neural networks, SVM models, decision trees and random forests. The conclusion presented forms a concrete legal justification to explain recent developments in international trade law concerning software and algorithm disciplines, highlighting the potential role of copyright law as a mechanism to facilitate greater competition within artificial intelligence markets.

Table of Contents

Figures and Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 5

Acknowledgements

To my supervisors, Professor Maurizio Borghi, and Professor Dinusha Mendis, I could not have done this without you. Your guidance and encouragement have been invaluable. You have introduced me to new concepts and ways of thinking which have very much inspired me.

For helping me with the technical aspects of my PhD my heartfelt thanks to Dr Jock McNaught, Assistant Professor Hassan Alhuzali, Dr Martin Molan and Professor Mirco Musolesi.

This was all possible because of a PhD Studentship I was lucky enough to receive from CIPPM, Bournemouth University. Thank you!

本論文において日本関連の著作権法について指導してくださった上野達弘教授に心より感謝申し上げます。博士課程の研究のために日本に滞在した 5 か月間、上野教授ならび国際交流基金のご支援により早稲田大学にて非常に充実した日々を過ごすことができました。また、著作権情報センター及び早稲田大学法学部の図書館の方々にもとても親切にして頂き、大変お世話になりました。有難うございます。

この 5 年間、ユーモアを交えつつ気長に応援してくれた家族、ガンガン呑んで鎌倉状態になったり一緒にくさやを食べたり何度も美味しいマレーシア料理をご馳走してくれた友人エリボーにも感謝の言葉を送ります。

大英図書館を辞職し博士課程を始める決断を大いに支持してくれた今は亡き母と父、力を与えてくれてありがとう。 今もあなたたちのことを想わぬ日はありません。

I dedicate this to you.

Author’s Declaration

Some of the content of this thesis has been previously published or derived from the following publications:

Dinusha Mendis, Dukki Hong, and Benjamin White, *Copyright and Open Norms in Seven Jurisdictions: Benefits, Challenges & Policy Recommendations* (CIPPM 2024)

Maurizio Borghi and Benjamin White, ‘Data Extractivism and Public Access to Algorithms’ in Borghi M and Brownsword R (eds) *Law, Regulation and Governance in the Information Society: Informational Rights and Informational Wrongs* (Routledge 2022)

Introduction

1. Is Artificial Intelligence Intelligent?

Firmly embedded in everyday discourse, the term “artificial intelligence” brings with it misconceptions of both autonomy and sentient consciousness. If intelligent machines truly can dispense with humans and write and even sing AI-generated songs, where does this leave copyright – a centuries old law designed to protect human authorial expression? Before considering this question, it is informative to briefly differentiate between the often over-lapping taxonomies associated with this newest of general purpose technologies. The thesis will then embark on an examination of what role copyright law plays and could play in broader AI innovation policy.

The term artificial intelligence lacks a singular accepted definition despite its widespread use. Evoking concepts of so-called “strong AI”, descriptions frequently anthropomorphise AI by likening a machine’s behaviour to those of a human being.[1] Analytical approaches bring further dimensions to the question of intelligence into play. For example, Russell and Norvig propose four criteria for intelligence: thinking humanly, acting humanly,[2] thinking rationally, and acting rationally.[3] Nilsson, more functionally oriented, emphasises the causational aspects of “*making* machines intelligent”. Accordingly, he defines intelligence as the “quality that enables an entity to function appropriately and with foresight in its environment.” [4]

From the standpoint of philosophical anthropology, however, current AI models fall short of constituting “intelligence” in any meaningful sense. They cannot autonomously apply knowledge, demonstrate initiative, or “think” in the abstract. As Braga and Logan observe, AI ultimately is little more than computer engineers “providing the right training data to the right learning algorithms.”[5]

Turning to the black letter of the law, in the EU 2024 Artificial Intelligence Act,[6] a comparatively detailed definition was set forth:

---

[1] For example, see Matthew Scherer, ‘Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies’ (2015) 29 Harv JL & Tech 354, 362 (stating that AI “refers to machines that are capable of performing tasks that, if performed by human, would be said to require intelligence.”) Annemarie Bridy, ‘Coding Creativity: Copyright and the Artificially Intelligent Author’ (2012) 5 Stan Tech L Rev 1, 3 n.6 (describing a generative art system as “self-contained enough to operate autonomously.”). Daniel Gervais, ‘The Machine as Author’ (2020) 105 Iowa Law Review 2053, 2058 (“With deep learning, one could say that the computer ‘has its own brain.’ ”)

[2] The well-known Turing Test, or more aptly as it is also known the “imitation game”, according to Russel and Norvig is an “acting humanly” definition of AI.

[3] Stuart Russell and Peter Norvig, *Artificial Intelligence – A Modern Approach* (Pearson Education 2010).

[4] Nils Nilsson, *The Quest for Artificial Intelligence* (CUP 2009).

[5] Adriana Braga and Robert Logan, ‘AI and the Singularity: A Fallacy or a Great Opportunity?’ (2019) Information 10 2 73 1, 2 (quoting data scientist Daniel Tukelang.)

[6] Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 OJ L 2024/1689 12.7.2024 (hereinafter AI Act or AIA).

Art. 3 (1) " … a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments."

Focusing on the design of the model and its capabilities and outputs, this functional definition foregoes entirely all assumptions of intelligence or human-like traits being displayed by the software.

Reflective of the many different contexts that it is used in, ultimately a clear-cut definition of AI remains elusive. AI should however rightly be regarded not as a form of synthetic intellect, but as a broad umbrella term for encompassing a diverse family of software programs. Through the processing of data and receipt of human instructions, these programs are then able to generate various typologies of output, including but not limited to predictions, classifications, and associations.

2. Machine Learning – a Focus on the Science

Overlapping terminologies with AI include:

i) deep learning – generally referring to neural network models;
ii) text and data mining –used to describe machine learning techniques and processes for analysing data; and
iii) machine learning – a broad term which tends to be favoured by computer scientists.

Whilst often conflated with AI, the term *machine learning* (ML) tends to emphasise the science and process through which these software applications are developed. Arthur Samuel, a former Stanford University professor, is widely credited as defining machine learning as "the field of study that gives computers the ability to learn without being programmed."[7]

Definitions of machine learning commonly emphasise the role of computers as well as the crucial influence of data in shaping the AI code. Géron, for example, defines machine learning as the "science (and art) of programming computers so they can learn from data."[8] As a consequence, strict definitions of machine learning may not cover all AI models. This is because "lazy learning" algorithms such as K-Nearest Neighbours, logistic regression, and Naive Bayes rules, differ from

[7] The source of the quote is frequently cited as A L Samuel, 'Some studies in machine learning using the game of checkers.' (1959) IBM Journal of Research and Development 3 3. This reference however is incorrect, and the precise origins of this often-cited definition remain obscure.
[8] Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn & TensorFlow* (O'Reilly 2017) 4.

eager learning models such as SVM models, neural networks, and the like, as they do not undertake a training phase where elements of the model's code are "written" by the data.

Witten even questions whether the term "learning" is appropriate at all, as it evokes concepts of thinking and autonomy, suggesting instead that the word "training", which can be mindless and devoid of autonomy on the part of the person being trained, is more appropriate.[9]

It can be concluded that an accurate definition of both machine learning and AI should encompass a sense of improvement in the process, alongside the centrality of data, the software, and the role of the trainer-cum-author of the model. It is posited, therefore, that an appropriate definition for this newest form of software engineering is: *the field of computer science dependent on the processing of data and the selection and optimisation of the software by computer engineers*. This results in a model capable of producing unique classifications, predictions, and other forms of output, without being explicitly and directly instructed to do so by human programmers.

3. Copyright Context

Copyright looms large over this newest of general purpose technologies. Operative over much of the ML production lifecycle, copyright serves as a broad but often overlooked regulator of machine learning markets. It would all the same be naïve to attribute the high levels of market and geographic concentration evident in AI markets primarily to variations in national copyright regimes. It is, nevertheless, posited that in "copy-reliant" information technology markets the flexibility of the copyright framework does play an important economic function, alongside an entrepreneurial culture, accessibility to public and private investment, and the nexus that exists between research institutions and a highly skilled workforce.

Copyright is important for a number of reasons. Firstly, where software code is original in its expression, exclusive rights will protect the work from competitors' free riding. Secondly, copy-reliant markets are functionally contingent on the ability of technologies to lawfully reproduce a work. In AI markets, this may be either for training an eager learning model like a neural network, decision tree, or a support vector machine (SVM), or for deployment solely in the analysis phase (i.e. lazy learners).

Thus, the ease with which an organisation can access the copyrighted fuel that feeds its data-driven machines will prove economically decisive. As a result, it can be observed that countries like the US

[9] Id. 8.

and Japan, who rely on open norms to regulate the use of protected subject matter, have cemented an arguably unassailable lead in their respective AI markets. This contrasts with Europe, where the unauthorised reproduction of a copyright work is *de facto* rendered unlawful absent a contractual agreement or intervention by the courts or the legislature.

In sum, this thesis argues copyright is a key regulatory tool that governments can better utilise to control AI technologies and as a consequence incentivise the markets in which they are deployed. In the interests of public welfare, from a macroeconomic perspective, this in turn can positively affect the extent to which AI-related technologies and businesses can thrive. Chapter 1 of this thesis explores copyright law and the global market imbalances in greater detail.

4. Research Question and Aims of the Research

The research question of the thesis can be summarised as follows:

"To what extent does copyright law (and related rights) in the UK, EU, Japan and the US support and influence AI model development and competition, and as a result, how has it influenced software and algorithmic developments in international trade law?"

This main research question can be broken down into the following objectives:

i) To examine and compare the differing approaches taken by four jurisdictions towards limitations and exceptions that support machine learning. This is done with a particular focus on how they support innovation while simultaneously exacerbating pre-existing issues around excessive market and geographic concentration that are a prevalent feature of information technology markets.

ii) To analyse the extent to which exclusive rights in copyright software law protect machine learning models.

iii) As a result of these findings, analyse the impact that weak copyright protection for machine learning software and algorithms has had on software disciplines in trade law.

iv) To propose a novel AI reverse-engineering framework to support competition and development in instances where digital actors are defined by the authorities to hold excessive market power.

5. Scope of the Research

The thesis focuses primarily on four different bodies of law: the UK, the European Union, Japan, and the United States. The United States and Japan were chosen for their economic prominence in AI markets – jointly with Internet of Things, e-commerce, and robotics respectively – as well as, from the standpoint of jurisprudence, for their flexible approach to copyright exceptions which support AI. The former being known for fair use and its extensive case law, and Japan for its introduction into statute in 2018 of a novel open-ended derogation designed to support its technology and research-intensive industries.

The UK and the EU serve in some senses as a counterpoint, as both are struggling to compete in AI markets. Moreover, from a jurisprudential standpoint, they have chosen to introduce enumerated exceptions for machine learning which in the case of the UK disallows all forms of commercial data analysis, and in the EU, is dependent on a novel rights reservation.

These different regions were selected due to their leading roles in AI development, the influence of their distinctive legal frameworks on regional and global intellectual property standards, and the ready availability of statutory and case law commentary. Both Europe and America provide a rich vein for research on the intersection of AI and copyright – Europe has codified an array of differing provisions in its law, while the US is distinguished by an impressive and continually expanding body of case law.

Looking beyond the western world, while China's position in technology markets and its case law is referenced, in East Asia, Japan forms the focus of this study. This decision is supportable for different reasons. First, unlike Japan, China attracts much academic attention which has already created a substantial body of literature in English.[10] Secondly, Japan's codified derogation for machine learning and its strong civil law tradition, renders it a clearer point of comparison with Europe and the US compared to China – characterised as it is by a hybrid legal system incorporating influences from civil, Confucian and socialist law systems.[11] T Third, Japan's status as an export-oriented technological power, the authorities' regulatory focus on ML exceptions, and the apparent spillover effect of the weak copyright protection afforded AI models on Japan's role in the development of the software and algorithms discipline within free trade agreements and the WTO Joint Statement Initiative, makes it a uniquely compelling subject for research.

[10] For example, Tianxiang He, 'Copyright Exceptions Reform and AI Data Analysis in China: A Modest Proposal' in Jyh-An Lee, Reto Hilty, and Kung-Chung Liu (eds), *Artificial Intelligence and Intellectual Property* (OUP 2021). A Zhuk, 'Navigating the legal landscape of AI copyright: a comparative analysis of EU, US, and Chinese approaches.' AI Ethics 4, 1299. Lee Ju Yoen infra n 432. Yong Wan and Hongxuyang Lu, 'Copyright protection for AI-generated outputs: The experience from China.' (2021) 42 Computer Law & Security Review, 105581. Quinru Li, 'Copyright protection paradigm in the generative AI era: originality, attribution, and infringement' (2025) 16 Advances in Social Behavior Research 8, 46. Tianxiang, He 'The sentimental fools and the fictitious authors: rethinking the copyright issues of AI-generated contents in China.' (2019) Asia Pacific Law Review 27 2, 218.
[11] John Quigley, 'Socialist Law and the Civil Law Tradition' (1989) 37 *The American Journal of Comparative Law* 4 781, 781.

Notwithstanding these justifications, this author recognises that while focussing on a limited number of jurisdictions analytically deepens the research and makes it logistically manageable, ineluctably it entails its own limitations. Most notably, the exclusive focus on jurisdictions within the Global North raises the question of the feasibility of legal transplants into less developed legal systems, as discussed in Chapter 6, constrained as they often are by limited institutional capacity. Moreover, such an approach excludes consideration of other significant and rapidly expanding markets, including China, India and South Korea, therefore potentially overlooking legal innovations in these countries.

In sum, the research focuses on both exclusive authorial entitlements that may arise in the programming and training of a machine learning model, as well as limitations and exceptions that enable the use of data for training and support the analysis of copyright works, the latter being a process inherent to all machine learning models. Examining "both sides" of the copyright bargain is crucial in establishing the extent to which copyright law supports AI-driven innovation. As discussed in final chapter, particularly the analysis of exclusive rights, also shines a light on the geopolitical AI-related developments in free trade agreements and at the World Trade Organisation.

6. Outline of the Thesis

This thesis addresses the question of the extent to which the contours of copyright law and *sui generis* database rights affect the adoption of machine learning, and as a consequence, how it contributes to the relative market standing of the four different jurisdictions listed above. Furthermore, it explores how the answer to this question has influenced digital commerce related developments in international trade law.

The value of intangible assets such as copyright that underpin the extreme high market capitalisation realised by some of the world's dominant AI-reliant technology companies, coupled with the centrality of data to machine learning (also often regulated by copyright) underscore its role as a regulatory tool in the digital economy. In examining the function that copyright plays across the AI lifecycle – and as a result, how it interweaves with other competition-inhibiting features of digital markets – brings to the fore pathways that should be explored by policy makers to rectify the severe market and geographic concentration that characterises many key AI-dependent markets.
This thesis proposes that as an internationally regulated body of law central to the machine learning process, copyright law should play a crucial role in national and international AI strategies, serving as a structural remedy to lowering barriers to market entry and as a result fostering competition. *A priori* utilising copyright and para-copyright regulatory approaches has the potential to alleviate the severe economic imbalance that has characterised technology markets over the past thirty years – a phenomenon which Korinek and Stiglitz assert AI is only set to worsen. Without addressing the issue

of excessive market concentration and the geographic imbalances that have arisen, the economic and social goals of developing and second comer developed nations are likely to be severely stymied, as they are largely locked out from the economic rewards derived from these lucrative new marketplaces.

The thesis is structured as follows:

In the first part of Chapter 1, the fundamental characteristics of both AI technologies and AI-reliant markets are examined, with a particular emphasis on the role that copyright plays in the economic and geographic concentration of AI markets. The changing landscape of extra-copyright algorithmic regulation in the European Union, which from a regulatory standpoint is the most active jurisdiction examined in this thesis, is also briefly examined.

From the perspective of incentivising the widest possible adoption of AI, Chapter 2 analyses the AI-specific copyright interventions that have taken place in the UK and the European Union. Namely, a provision in UK law concerning computer-generated works, as well as the introduction of text and data mining exceptions in both jurisdictions.

Chapter 3 undertakes a parallel doctrinal and comparative analysis of the US and Japanese exceptions that facilitate machine learning, before concluding with an evaluation of the efficacy of the various different legislative responses.

Turning from exceptions to exclusive rights, Chapter 4 outlines the main copyright software doctrines which have emerged through legislation and case law to protect computer software. This forms the basis of Chapter 5, which applies existing doctrines to the unique workflow of machine learning software in order to examine the extent to which protection is afforded to a trained model that relies in large part on programs sourced from an openly licensed software library / AI toolkit.

Having examined the extent to which copyright law and related rights in the UK, EU, Japan, and the US support and influence AI model development and competition, the final chapter turns to how copyright law has affected trade law. Seemingly linked to the weak protection afforded much AI software, Chapter 6 examines new para-copyright AI developments in trade law. The chapter ends with the proposal and evaluation of a novel reverse engineering provision undertaken in the light of the Berne Three Step Test. Such an article should ideally be introduced at the international level aimed to heighten competition in technology markets and facilitate knowledge spillovers.

A glossary of machine learning terminology is provided on page 238. Compiled iteratively in the course of writing this thesis, the glossary draws on academic literature, including Alpaydin,[12] Goodfellow et al,[13] as well as Google's Machine Learning Glossary.[14] Terms were selected on the basis that they were likely to be unfamiliar to a non-technical reader, or, where some familiarity can be assumed, where definitional precision was nonetheless considered beneficial.

7. Methodology

This research primarily adopts a mixed doctrinal and comparative law methodology in order to draw out points of divergence and convergence between the four jurisdictions examined in this thesis in relation to their respective copyright approaches to machine learning. As not engaging in empirical research methods, the hallmark of original American legal realists,[15] or observing judicial outcomes for reasoning derived from beyond the black letter of the law,[16] the research cannot be said to follow a formal legal realist approach. Nevertheless, as in that tradition, with its emphasis on the social consequences of the law, the research turns its eye outward from the "law in books" to the society it governs.[17] In this sense, centred around the economic impacts of machine learning jurisprudence the thesis can perhaps be characterised as falling within what has been described by Calboli and Montagnani as "the offspring of American legal realism."[18]

Doctrinal Methodology

A doctrinal research methodology is the foundation and starting point for much legal research.[19] Often referred to as study of the "black letter of law", broadly speaking it involves a three-part process. Firstly, as was the case in this research, authoritative legal sources including statute, relevant case law, official documents, parliamentary records etc are located and assembled, prior to performing a critical interpretation and analysis of the relevant texts. Subsequent to this, based on the research, reasoned conclusions are presented. [20] Twining highlights that one potential drawback of doctrinal research is that it is autotelic in nature – purely theoretical, and conducted in isolation of the problems the law seeks to address.[21] It is submitted however that the strong economic policy focus of the research, contextualising the analysis in the effect of copyright law on the uptake of AI technologies, serves to mitigate such criticisms.

---

[12] E Alpaydin, *Machine Learning* (MIT Press 2021).
[13] I Goodfellow, Y Bengio and A Courville., *Deep Learning* (MIT Press 2016).
[14] https://developers.google.com/machine-learning/glossary
[15] Mark Suchman and Elizabeth Mertz, 'Toward a new legal empiricism: empirical legal studies and new legal realism' (2010) 6.1 Annual Review of Law and Social Science 555, 565.
[16] Dan Priel, 'Legal Realism, Legal Positivism, and Determinacy' (2026) 71 The American Journal of Jurisprudence 1 17.
[17] Roscoe Pound, 'Law in Books and Law in Action' (1910) 44 American Law Review 12.
[18] Irene Calboli and Maria Lillà Montagnani (eds), *Handbook of Intellectual Property Research* (OUP 2021) 5.
[19] Terry Hutchinson, 'Doctrinal Research' in Dawn Watkins and Mandy Burton (eds), *Research Methods in Law* (Routledge 2026) 37.
[20] Id. 18.
[21] W Twining, *The Taylor Lectures 1975* printed as *Academic Law and Legal Development* (University of Lagos Faculty of Law 1976) 20.

Comparative Law Methodology

As a study of statute and case law in four different jurisdictions, by its very nature the study is not only doctrinal but also comparative. Calboli defines comparative law as "the body of law focussing on the comparison of different legal systems … to detect their similarities and differences".[22] According to Eberle, finding its methodological roots in the quest for an understanding of foreign legal systems in order to identify universal legal principles – historically stemming from private law – this thesis steps outside the traditional mission of a comparative law study.[23] It aims at broader pursuits by shedding light on how economic and political policy developments intersect with the development of law. In doing so, through the lens of copyright, related rights and trade law, it seeks to contribute normatively to solving an important public policy issue that transcends national borders – namely the excessive concentration of technology markets.

It does this by evaluating the relative strengths and weaknesses of the limitations and exceptions that facilitate machine learning across the selected jurisdictions, assessed from the perspective of maximising the uptake of this newest of general purpose technologies. Furthermore, the thesis sets out the contours of a novel reverse engineering provision that global policy makers can adopt to encourage new market entrants and greater competition in technology markets. A further objective of this study is to contribute to the field of comparative law itself – which has historically prioritised European and North American legal systems[24] – through the inclusion of East Asian jurisdictions, with a particular emphasis on Japan. Moreover, with its focus on competition in markets reliant on machine learning it seeks to make a theoretical contribution to the fields of both comparative law and development,[25] as well as comparative law and innovation economics.[26]

As with doctrinal research, comparative law methodologies also are not without their inherent limitations.[27] The author has read the English and Japanese literature in the original, thereby as highlighted by Glanert, minimising the potential impact of linguistic barriers to comparison and interpretation.[28] Furthermore, while comparison of very different territorial laws and legal traditions affected by the same technologies brings with it its own interpretational challenges, because copyright is a globally comparatively harmonised body of jurisprudence, it is arguably less prone to the conceptual misunderstandings that may arise from textual extraction and translation when compared

22 Irene Calboli, 'Comparative Legal Analysis and Intellectual Property Law: A Guide for Research' in Calboli and Montagnani supra 18, 47 (quotation marks omitted).
23 Edward Eberle, 'The Methodology of Comparative Law' (2011) 16 Roger Williams University Law Review 1 51, 53.
24 Mathias Siems and Po Jen Yap (eds), *The Cambridge Handbook of Comparative Law* (CUP 2024) 2.
25 Andrew Harding and Yong-Shik Lee, 'Law and Development', id. 510.
26 Bruno van Pottelsberghe de la Potterie, 'Innovation Economics Approach to Intellectual Property Law Scholarship' in Irene Calboli and Maria Lillà Montagnani (eds) supra 18, 423.
27 See Irene Calboli, 'The role of comparative legal analysis in intellectual property law: From good to great?' in Graeme Dinwoodie (ed) *Methods and Perspectives in Intellectual Property* (Edward Elgar 2013) 10.
28 Simone Glanert (ed), *Comparative Law - Engaging Translation* (Routledge 2014).

to other areas of law. As widely acknowledged in the literature, all the same, it remains the case that law is deeply embedded within its own cultural and social setting.[29] Accordingly, detaching law from its environment for the purposes of comparison will involve its own inherent shortcomings. It is acknowledged therefore that the author's own cultural, political, linguistic, and legal background will have had an influence on the analysis presented herein.

8. Contribution to Knowledge

The varying strands of this thesis draw on existing bodies of literature that vary in both age and level of scholarly depth. The first academic evaluations of copyright and computer generated works can be traced back to the 1960s.[30] As outlined in Chapter 2, since then numerous scholars have offered interpretations regarding who should be afforded copyright in the creative outputs of machines.[31] Analysis of the scope and legal certainty provided by machine learning limitations and exceptions is however comparatively new, prompted in Japan and Europe by the codification of limitations and exceptions into law. The works of Ueno,[32] Geiger,[33] Kretschmer, Margoni[34] are particularly relevant in this context. In the United States, it was the cases involving the mass digitisation of books by Google that first prompted much scholarly commentary in this area.[35] The literature generally concurs that forms of AI such as text and data mining that does not expose copies of the works being processed to the public, or only provide very small amounts (so-called "non-consumptive" uses), will fall as fair use.[36] However, many US scholars prudently maintain that the legal position of generative AI systems that make extensive use of the works of creative authors will turn on the particular facts and circumstances of the model and the nature of the outputs in question.[37]

---

[29] For example, Günter Frankenberg, 'Critical Comparisons: Re-thinking Comparative Law' (1985) 26 Harvard International Law Journal 4. Pierre LeGrand, 'European Legal Systems Are Not Converging' (1996) 45 International and Comparative Law Quarterly 52.
[30] Infra n. 371.
[31] For example, Daniel Gervais, 'The machine as author' (2019) 105 Iowa Law Review 2053 (pointing to a Benthamite view of the law that with rights come responsibilities and therefore an AI cannot be afforded copyright ownership.) Andrew Wu, 'From Video Games to Artificial Intelligence: Assigning Copyright Ownership to Works Generated by Increasingly Sophisticated Computer Programs' (1997) 25 AIPLA Q J 25 131, 174 (arguing that copyright for purely machine generated works should be granted to the programmer or owner of the software.) Pamela Samuelson, 'Allocating Ownership Rights in Computer-Generated Works' (1986) 47 U Pitt L Rev 1185, 1207 (making the case that incentive rationales for copyright better align with affording ownership to users of AI software.)
[32] For example, Tatsuhiro Ueno,『平成 30 年著作権法改正について』年報知的財産法２０１８－１９(On the 2018 Amendment to the Copyright Act) in Intellectual Property Law Annual Report 2018-19 (Hyōronsha 2018). 人工知能と機械学習をめぐる著作権法上の課題: 日本とヨーロッパにおける近時の動向 (特集 AI がもたらす知的財産法の変容と未来）「２０１９」91 法律時報 33 | 'Copyright Law Issues Regarding Artificial Intelligence and Machine Learning: Recent Trends in Japan and Europe' (Special Feature: The Transformation and Future of Intellectual Property Law that AI Brings) (2019) 91 Legal Times 33.
[33] For example, Christophe Geiger and others, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU' (2019) Propiedad Intelectual Y Mercado Único Digital Europeo 27.
[34] Thomas Margoni T and Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) 71 GRUR International 685.
[35] Michael Carroll, 'Copyright and the Progress of Science: Why Text and Data Mining is Lawful' (2019) 53 Uc Davis L Rev 893. Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning' (2019) 66 Journal of the Copyright Society of the USA 29. See also, Maurizio Borghi M and Stavroula Karapapa, 'Non-Display Uses of Copyright Works: Google Books and Beyond' (2011) 1 Queen Mary Journal of Intellectual Property 21.
[36] For example, Michael Carroll supra 6. Matthew Sag, 'Copyright and Copy-Reliant Technology' (2009) 103 Northwestern University L Rev 1607. James Grimmelmann, 'Copyright for Literate Robots' (2016) 101 Iowa L Rev 657. Mark Lemley and Bryan Casey, 'Fair Learning' (2021) 99 Tex L Rev 743.
[37] Benjamin Sobel, 'Artificial Intelligence's Fair Use Crisis' (2017) 41 Colum JL & Arts 45 (making the case that fair use determinations in support of generative AI may damage authors, but excluding fair use will damage societal progress.) James Grimmelmann and others, 'Talkin' Bout AI Generation: Copyright and the Generative-AI Supply Chain.' (2025) 72 J Copyright Soc'y 251 (arguing there are too many

Building on this foundation, the thesis makes an original contribution to knowledge by undertaking the first in-depth cross-jurisdictional comparison of codified limitations and exceptions in the UK, EU, Japan and the US together with an analysis of their interplay with contracts and anti-circumvention provisions. Most distinctively, it conducts an exhaustive examination of a doctrinally novel open norm that was introduced into Japanese copyright law in 2018. The analysis of original Japanese-language sources further enhances scholarly understanding of this distinctive L&E and informs a historical interpretation to the background of the provision's introduction. Furthermore, through a systematic evaluation of the comparative strengths and weaknesses of these distinctive approaches, the thesis offers fresh insights for scholars and policymakers concerned with fostering competition and technological innovation within the framework of copyright law.

The subsequent two chapters shift focus to exclusive rights, situating the analysis within the rich literature of the 1980s and 1990s that analysed software copyright law and case law – in particular the scholarship of Samuelson[38] in the US and Nakayama in Japan.[39] Against this backdrop, in applying copyright doctrines to AI model production workflows the thesis constitutes the first cross-jurisdictional and comprehensive study into the copyrightability of AI software, thereby addressing a marked gap in the existing scholarship. Furthermore, based on this unique analysis, the thesis proposes a framework designed to guide courts in identifying appropriate lines of inquiry for determining whether, and if so where, copyright subsists in trained machine learning model software.

Notably, the examination of both restricted and permitted acts diverges from the existing literature which almost exclusively focuses on a single category of AI model – namely, neural networks. Based on a wide reading of the technical literature, this study draws original conclusions based on an analysis of an array of models including support vector machines (SVMs), linear and logistic regression models, decision trees, random forests, as well as lazy learner models that require no training such as K-Nearest Neighbours (KNN).

The final chapter examines the spillover effects on trade law that have emerged, it is posited in part, due to the weak copyright protection afforded AI model software. Foundational contributions by Lee-Makiyama,[40] Burri,[41] and Irion[42] provide critical insights into the ways in which laws governing

model types to make a generic determination around the applicability of fair use to generative AI.) Matthew Sag, 'Fairness and Fair Use in Generative AI.' (2024) 92 Fordham Law Review 5 1887 (suggesting that even where outputs are not direct expressive substitutions, nevertheless courts may consider wider harms to copyright's economic incentives.)

[38] Inter alia, Pamela Samuelson and others, *Software and Internet Law 4th Edition* (Aspen Law & Business 2011).

[39] Inter alia, Nobuhiro Nakayama,『ソフトウエアの法的保護』(Legal Protection of Software) (Yūhikaku 1990).

[40] Hosuk Lee-Makiyama, *Briefing Note: AI and Trade Policy* (2018).

[41] Inter alia, Mira Burri, *Current and Emerging Trends in Disruptive Technologies: Implications for the Present and Future of EU's Trade Policy* (Directorate General for External Policies 2017).

[42] Inter alia, Kristina Irion, *AI Regulation in the European Union and Trade Law: How Can Accountability of AI and a High Level of Consumer Protection Prevail Over a Trade Discipline on Source Code?* (Verbraucherzentrale Bundesverband 2021).

access to software and algorithms as a condition of import have emerged in international trade agreements.

Informed by patent law and growing EU regulatory algorithmic intervention, a novel AI reverse engineering provision designed to support technology transfer, competition, and economic development is proposed. AI technologies are further entrenching the excessive geographical and market concentration of information technology companies. Absent significant policy interventions this trajectory is likely to deepen the large gap between dominant AI incumbents, who frequently operate within winner-takes-all market structures, and all other actors; a new paradigm which transcends the established Global North-South divide.

Taken together, these contributions advance the field by not only shedding light on different legal approaches to machine learning and the protectability of AI software, but also by moving the debate on AI and copyright beyond the narrower question of how AI affects purely commercial creators. Building on the existing literature and analysing the role copyright plays in promoting the greater adoption of machine learning technologies, the thesis promotes a holistic understanding of the macroeconomic and knowledge transfer role that limitations and exceptions and access rights could play in service of economic development across the globe.

Chapter 1: Examining the Role Copyright Plays in Artificial Intelligence Technologies and Information Technology Markets

Introduction

This thesis proposes that copyright and para-copyright law regulatory interventions can act in tandem with other policy tools to ameliorate high levels of asymmetry in AI markets. Structured in three parts, this chapter examines the points at which copyright can arise in the AI production lifecycle and, as a result, how it comprises one important facet of the digital market conditions that are the object of ongoing regulatory scrutiny across the world. First, the chapter explores the four essential pillars of machine learning (ML), highlighting where they intersect with copyright. Second, a concise overview of the key legislative developments in Europe over the last three decades is undertaken, enacted in response to the societal and economic consequences of our growing reliance as a society on algorithmic decision-making. With a particular focus on copyright, the chapter concludes with an analysis of the structural characteristics of a globally imbalanced technology market – a phenomenon exacerbated by the adoption of AI and the high market valuations attributable to intangible assets.[43]

Section One

1. The Four Elements of AI

As an author-centric body of law, any evaluation of AI involving copyright requires an understanding of the role of the data scientist viz. author and the production environment in which s/he works. Broadly speaking, the deployment of AI can be described as a complex interplay of four factors: i) human ingenuity, ii) software, iii) data, and iv) hardware. In brief, with the exception of lazy-learning models, AI is the outcome of data scientists selecting and cleansing data, prior to using it to refine the machine learning model code. Sophisticated hardware solutions are also a prerequisite, given the ever-increasing processing power required to process big data.[44]

[43] See Jonathan Haskell and Stian Westlake, *Capital Without Capital* (Princeton University Press 2018). Herman Schwartz, 'Global secular stagnation and the rise of intellectual property monopoly' (2022) Review of International Political Economy 29 5 1448 (arguing that firms with high levels of IPRs have a lower marginal propensity, which has contributed to low levels of growth globally over the last 30 years in OECD countries.) Cédric Durand and William Milberg, 'Intellectual monopoly in global value chains' (2020) Review of International Political Economy 27 2 404 (concluding that IPRs lock in market power by creating barriers to market entry and knowledge transfer.)
[44] Alberto Lerner and Gustavo Alonso, 'Data Flow Architectures for Data Processing on Modern Hardware' 2024 IEEE International Conference on Data Engineering DOI: 10.1109/ICDE60146.2024.00439 (examining the requirement for developments in data processing to accommodate the complex architectures that have evolved as a response to every growing data volumes.)

As a body of law that protects tangible original human expression, *a priori* copyright is therefore particularly pertinent to two of the four factors: namely the software and the data that is used to train and / or be analysed by the model.

A clear explanation of the four core elements of AI is intended to anchor the legal analysis and normative discussions undertaken in the subsequent chapters in the practical realities of machine learning production.

1.1 Copyright Context

As already examined, the various definitions of machine learning and artificial intelligence remain relatively silent on the precise role of its creator.[45] As copyright pivots on the contribution of a human author, the presence or absence of the data scientist in the AI lifecycle has critical juridical ramifications to which scholars and policy makers must be attuned.

As is well known, the original expression of a human author is the *sine qua non* of copyright jurisprudence. Where informational matter lacks original human expression, as a rule, it will fall beyond the scope of protection.[46] The cornerstone role of a natural person is enshrined in the Berne Convention, which limits the duration of copyright to a minimum of the "life of the author and fifty years *after his death*."[47] Consequently, in the UK, copyright expires "70 years from the end of the calendar year in which the author dies."[48] Furthermore, in the UK and the European Union, copyright is the result of an "author's own intellectual creation",[49] and as is evident from the raft of substantiating case law from the Court of Justice of the European Union (CJEU), where the human author is absent, so too is copyright protection.[50]

[45] In this thesis the terms computer programmer, software engineer, data scientist etc will be used interchangeably to signify that while a data scientist focuses on the processing of data they will also have a background in computer science and programming. Additionally, in smaller organisations such distinctions are academic as employees will be tasked with all aspects of machine learning from data collection to programming.
[46] As discussed in Chapter 2, significant exceptions to this are the UK, Ireland, New Zealand, India, South Africa and Hong Kong where quasi-copyright provisions exist for autonomously produced computer-generated works.
[47] Article 7 (1).
[48] s.12(2), Copyright, Designs and Patents Act 1988 (hereafter CDPA).
[49] See Article 1 (3). Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs [2009] OJ L111/16 (codified version of 91/250/EEC) (hereafter Software Directive or Computer Programs Directive); Article 6 and Recital 17 of the Council Directive 93/98/EEC of 29 October 1993 harmonizing the term of protection of copyright and certain related rights [1993] OJ L290/ 9 (hereafter Term Directive). Recitals 15, 16 and Article 3 of the Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L77/20 (hereafter Database Directive). See also s.3A, CDPA.
[50] For example, in non-software or database related cases, Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* ECR I-06569 (hereinafter *Infopaq 1*), Case C-145/10 *Eva-Maria Painer v Standard Verlags GmbH and Others* ECR I-12533 (hereinafter *Painer*), Case C-310/17 *Levola Hengelo BV v Smilde Foods BV* [2018] ECLI:EU:C:2018:618 (hereinafter *Hengelo*), Case C-683/17 *Cofemel – Sociedade de Vestuário SA v G-Star Raw* CV [2011] ECLI:EU:C: 2011:648 (hereinafter *Cofemel*) etc.

Similar to Europe, in the United States, the anthropocentric nature of copyright is evident, as in the view of the courts it resides in the "original intellectual conceptions of the author".[51] Irrespective of the operativity of international treaties, any remaining doubts about the anthropocentric nature of copyright law have been definitively resolved by the well-known *Naruto* case. Referencing the hominal requirement of copyright law, the Ninth Circuit Court of Appeals opined that "this monkey – and all animals, since they are not human – lacks statutory standing under the Copyright Act."[52] Thus, reinforcing the principle that copyright law is rooted in the presence of a human author. As a foundational norm of the international framework, to qualify for protection, any work – including those generated by AI – must therefore embody tangible original expression created through human intellectual effort and labour.

2. The Role of the Author-Programmer in a New Programming Paradigm

2.1 Software 1.0

To evaluate how copyright supports and incentivises AI model creation, it is essential first of all to understand the modalities of traditional program development, around which the norms of the present-day software copyright framework have been built. By understanding the precise role of the author-programmer in a rules-based software environment, it provides a foundation for the evaluation undertaken in Chapter 5, as to the extent to which the contemporary legal norms for software align with the realities of AI model design.

Taking first a typical example of traditional rules-based programming, if a computer engineer were to develop a program for analysing who to choose for pre-diabetes screening, having selected certain common functions from a software library, s/he would then proceed to write code for each of the factors relevant to a diabetes diagnosis. This would be based on the programmer's research and interaction with specialists in the field, and as a minimum would include age, gender, race, co-morbidities, BMI, diet etc. The software engineer would write explicit instructions specifying that if certain combinations of factors, or a total number of factors are present, the individual should be called forward for testing. Depending on the number of factors selected as relevant, the code could be lengthy and contain complex instructions which may be difficult to maintain and amend if changes relating to the diagnosis of diabetes were to emerge after the program were completed.

[51] *Feist Publications Inc v Rural Tel Serv Co*, 499 US 340 (1991) 347 quoting from the *Trade Mark Cases*, 100 US 82 (1879).
[52] *Naruto v Slater,* 888 F 3d 418 (9th Cir 2018) 420. (A case concerning whether copyright subsisted in a selfie taken by a macaque.)

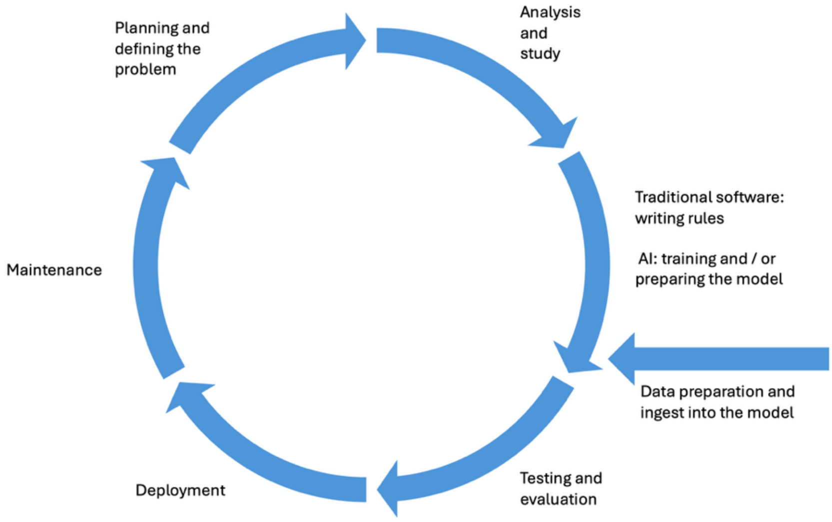

Fig. 1-1. AI Development Life Cycle

Having written part or all of the software, in what can be a potentially long and convoluted process, the code is then stacked together to form the overall program.

2.2 Software 2.0

Although as shown in the diagram above many commonalities with traditional software workflows remain, in contrast to rules-based programming, a machine learning, or as Karpathy has called it a "Software 2.0"[53] approach, would first see the data scientist select one, or a number of pre-written models, from an open source software library, commonly referred to as an "AI toolkit".[54]

Derived largely from a software library, and therefore requiring comparatively minimal programming (as explained in Chapter 5), the models are then trained on a range of patient data representing the factors that are known to influence diabetes.[55] This would include labelled data for patients who did and did not develop diabetes, enabling the model to learn which data combinations are effective predictors for diabetes. Thus, rather than a human programmer laboriously writing if-then statements

[53] Andrej Karpathy, 'Software 2.0' (*Medium*, 11/11/2019) https://karpathy.medium.com/software-2-0-a64152b37c35.
[54] While software libraries date back to the 1940s, their popularity has increased significantly over the last 20 years owing to the evolution of object-oriented programming in the 1980s, coupled with the rise in the 1990s of the internet and the Free and Open Source Software (FOSS) movement.
[55] Taking the diabetes example given above, the data scientist would almost certainly have relied upon one of the well-known open source frameworks such as PyTorch, PaddlePaddle, TensorFlow (supported by Meta, Baidu and Google respectively) or, Scikit-Learn.

for predicting diabetes, the rules are fixed in the code during the training phase through the interaction of the largely pre-provided software and the labelled data.

2.3 The Central Role of the Data Scientist

Whilst the role of the computer engineer as author of the code is more or less evident in a traditional programming paradigm, a question deserving more attention than it has hitherto received in the literature, is whether the traditional role of the author-programmer in copyright law still holds true in the same way for AI software. Karpathy argues that Software 2.0 is radically "altering the programming paradigm", as data scientists are spending most of their time curating, cleaning, and labelling data, and far less time on programming.[56] Thus, while data scientists are still of central importance, the jurisprudential question becomes the extent to which their actions in creating an AI model trigger authorial entitlements, and as a consequence, whether sufficient incentives are afforded investors in this newest of frontier technologies.

From the standpoint of copyright, the work of a data scientist *qua* model author can be divided into two broad categories. The first relates to the selection, cleansing, and pre-processing of data, either intended to train the model, or for the model to analyse and generate outputs from. Whilst beyond the scope of this thesis, it is of note that where original in arrangement, compilations of data may attract copyright protection. In Europe, *sui generis* rights may also be triggered subject to "substantial investment in either the obtaining, verification or presentation of the contents" of the data in a database.[57]

The second, more relevant, category for this discussion are acts associated with algorithm selection, AI software optimisation, and programming of code. Though discussed extensively in Chapters 4 and 5, at this juncture it is nevertheless worth clarifying that according to jurisprudential theory copyright will be afforded to human authors where the code they write, or the choices they make in the structure and organisation of the program, are original. Thus, whilst copyright will frequently be afforded the many routines written using traditional programming methods that are found in AI software libraries, this cannot necessarily be assumed to be the case for the final ML model, as by and large, it is the result of refining pre-written openly licensed software in the light of the results returned by the data and the application of commonly known heuristics.[58]

[56] Karpathy, supra 53.
[57] Article 7 (1), Database Directive.
[58] This sentence is not intended to imply that copyright will not inhere in the pre-written sub-programs available in the toolkit.

The conspicuous divergence from the programming paradigm upon which TRIPS,[59] and contemporary computer software copyright law is based, may be evinced from sectoral reports. For example, according to the 2016 CrowdFlower Data Science Report, while 82% of a computer engineer's time is spent collecting and organising data, by contrast only 4% of their time is occupied with refining algorithms, and only 9% mining data for patterns. While the report fails to provide a definition of these two terms, they will unquestionably encompass copyright-relevant activities such as writing or amending code, inputting instructions relating to the hyperparameters of the model via a graphical user interface (GUI), as well as undertaking model optimisation. In contrast to data collection and organisation, the relatively small amount of time dedicated to these copyright-relevant activities indicates the marked difference between AI and the traditional programming norms around which current copyright software protection has been based.

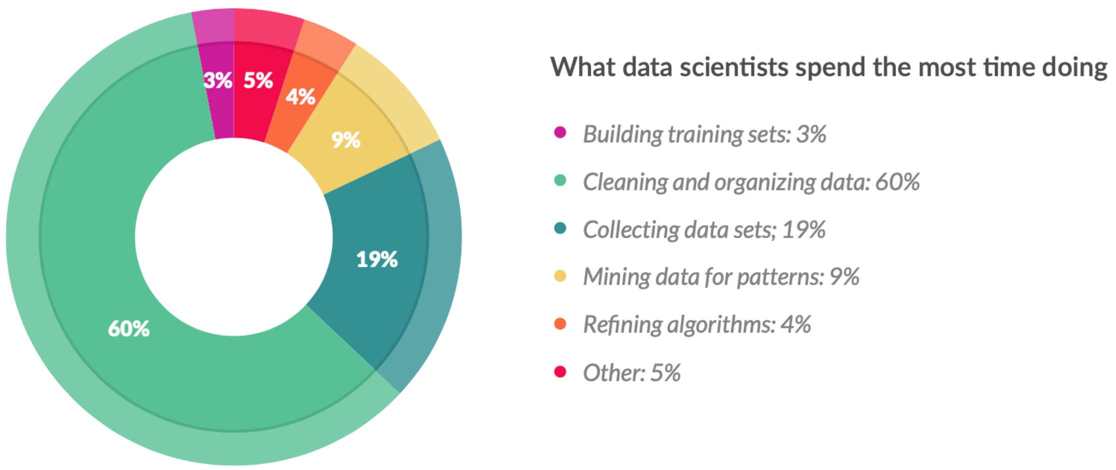

Fig. 1-2. 2016 Data Science Report (CrowdFlower)[60]

None of this is to say that the appropriate kind of inputs that induce copyright protection are necessarily absent. Nevertheless, as explored in Chapter 5, it does highlight the need to re-evaluate the extent to which copyright protection for computer programs applies to the creation of AI models. From this brief evaluation alone, it may already be discerned that the role of data as a co-author in code generation, coupled with the widespread use of openly licensed AI toolkits, has engendered less uniform protection for this new type of computer program, compared to the rules-based programming paradigm around which software copyright law was originally fashioned.

[59] Agreement on Trade-Related Aspects of Intellectual Property Rights, (April 15, 1994) (Annex 1C) of the Marrakesh Agreement Establishing the World Trade Organization (1994) (hereinafter TRIPS).
[60] CrowdFlower, '2016 Data Science Report' (CrowdFlower 2016) 7.

3. Big and Small Data – the Fuel for the Machine

Alongside the computer engineer and software, the third essential element of AI is data. Data is required for two distinct purposes: first, in the case of eager learning models, for training. Secondly, for the analysis and production of a model output.

Before data is of any value to a data scientist, it must be collected and refined. Pre-processing involves many different activities, starting first with creating a working copy of the data obtained, followed by the storing of the original data for future use or validation purposes.[61] The data is inevitably going to be noisy and require cleansing, so that the model can work effectively. Inconsistencies and discrepancies in the data can be the result of numerous issues including faulty instrumentation, inputting errors, or naming inconsistencies. Consequently, filling in missing data, converting nominal data into numeric representations and removing outlying data points may all be required.

After cleaning, the data undergoes transformation, where its format or structure is altered for the software to process. This can involve reducing or increasing the number of data points, creating new features (feature construction) through aggregation,[62] or applying feature scaling (normalisation) to ensure all features contribute equally to the model.[63] Data may also be batched to improve processing efficiency.[64, 65]

For eager learning models, the processed data is split into training, validation, and test sets. Models learn from the training set, are further honed on the validation set, and ultimately evaluated on the test set to verify their ability to generalise and avoid "overfitting".[66]

According to CrowdFlower's 2016 Data Science Report (Figure 1-2) the collection and processing of data stands on average at 82% of a data scientist's time. This represents a significant economic commitment in data collection and curation, raising the question of how intellectual property rights can protect this investment. Importantly, while copyright or *sui generis* database protection may be triggered, it is not necessarily afforded in all instances. For example, original artistic, literary, and audiovisual works authored by a human being, alongside original compilations of information and data, will attract copyright protection. At the other end of the scale, in their raw form, automatically

[61] For an in-depth overview of data pre-processing see Jiawei Han et al, *Data Mining Concepts and Techniques* (Morgan Kaufmann 2011) Chapter 3 and Géron, op. cit. 8, 33-78.
[62] Géron, op. cit. 8, 25 ff.
[63] Id. 66.
[64] Joseph Nelson, 'You Might Be Resizing Your Images Incorrectly' (*Roboflow,* 31/1/2021) https://blog.roboflow.com/you-might-be-resizing-your-images-incorrectly/
[65] Paul Sayak, 'Learning to Resize in Computer Vision' (*Keras*, 30/4/2021) https://keras.io/examples/vision/learnable_resizer/
[66] Géron, op. cit. 8, 29 ff.

produced machine data,[67] event data, or spatiotemporal data, chemical formulas and the like, may not be assigned protection by any form of intellectual property right.

4. Hardware

The final pillar of the AI ecosystem is the hardware. Semi-conductors are essential for many electronic devices, not least machine learning technologies. For this reason, at the time of writing, two semiconductor companies (Taiwan Semiconductor Manufacturing Company (TSMC) and NVIDIA) are in the top ten list of the world's most valuable companies by market capitalisation.[68]

As AI models have become increasingly sophisticated, the demand for greater processing power to handle ever larger datasets has grown exponentially. High-powered processing chips like graphic processing units (GPU) and quantum chips, drive the processing of ever more complex models and large sets of data. The strategic importance of semiconductors is such that many developed countries are seeking to rapidly expand their domestic production of chips, rather than having to rely on the vagaries of supply from a handful of suppliers – one, in the case of TSMC, susceptible to increasingly tense relations between Taiwan and China.[69]

To conclude, while much of the non-technical literature on AI emphasises the role of data in machine learning, the reality is that it is only one of four equally important factors. Like any unprocessed matter, data alone has no worth. Only once it has been coupled with software, running on an appropriately robust technical infrastructure, can its value be realised. The overemphasis on data in the legal literature has arguably overshadowed the crucial role of the author-programmer within the AI ecosystem. It is only thanks to the human author that copyright can be afforded to ML models and their outputs. In turn, as a body of law that accords monopoly rights to all concretised expressions of human originality, the copyright framework influences the contours of the wider digital economy.

As background to the proceeding chapter, which examines copyright-related AI interventions by UK and European regulators, the following section turns to the European Union's extra-copyright response to the broader societal and economic challenges presented by algorithmic technologies. This analysis underlines the wide-ranging and novel impacts of AI-reliant "superstar" firms on contemporary society. Building on this, with a particular focus on the role of copyright, the foundations for the very high levels of market concentration associated with AI markets are explored in the concluding sub-section.

---

[67] Article 43. Regulation 2023/2854 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) [2023] OJ L2023/2854 (hereafter Data Act).
[68] Q4 2025. New York Stock Exchange.
[69] Chad Bown, 'How the United States marched the semiconductor industry into its trade war with China' (2020) 24 East Asian Economic Review 4 349.

Section Two - Europe

1. Regulation over Competition?

The acute economic dislocation that has emerged around AI speaks to the necessity of a less fragmented approach to regulation than is typically seen when governments are faced with challenges posed by new technologies. The "mechanistic trap", a term coined by Capra and Mattei, refers to a situation where little if any coherence exists between different laws and regulations.[70] If not addressed, the lack of a comprehensive legal and political framework surrounding AI will further entrench geographical asymmetries and the market dominance of a few superstar companies that define today's ML-dependent technology markets. With this in mind, this section now turns to analyse the development of algorithmic regulation in Europe over the last 30 years.

As highlighted in the Draghi Report, Europe lags economically ever further behind its technology competitors.[71] In the first half of 2024, the EU only attracted 6% of global investment into AI.[72] Despite these deep systemic issues, an overarching and coordinated economic strategy nevertheless appears elusive. Measures to date by the authorities can be divided broadly into three overlapping areas: financial markets, product safety, and fundamental rights. With the notable exception of the 2019 Copyright in the Digital Single Market Directive, it can be observed that the majority of the EU's algorithmic policy initiatives aim to mitigate the negative social welfare impacts resulting from the monopolistic practices of established digital market players, rather than focusing on broader issues of market competition.

As a prelude to the next chapter which focuses exclusively on copyright law, this section will outline the legislative developments in the European Union that relate to digital platforms, health and safety, market trading, and fundamental rights. It aims to serve as a brief historical record of the evolution of algorithmic regulation in Europe, and in doing so underline that by bringing in *ex-post* mechanisms to deal with public welfare harms, the focus of regulation has by and large been reactive. The centre of attention has largely been one of amelioration and, by implication, acceptance of extreme market disequilibrium, at the expense of a positive innovation agenda that seeks to lower barriers to market entry and engender algorithm-driven competitiveness and innovation.[73]

[70] Ioannis Lianos and Alexey Ivanov*, Digital Era Competition: A BRICS View* (BRICS Competition Law and Policy Centre 2019) (Based on the work of Fritjof Capra and Ugo Mattei, the use of the term "mechanistic trap" signifies systems thinking – that reality, and therefore law, is not an aggregate of separate parts, but a network of interrelationships, and that by working within predefined institutional restraints and silos the law is unlikely to be able to deal with major societal challenge.) 54.
[71] *The Draghi Report: In-depth Analysis and Recommendations* (Publications Office of the European Union 2025) Part B. 3.2.
[72] Bertin Martens, *Catch-up with the US or Prosper Below the Tech Frontier? An EU Artificial Intelligence Strategy* (Bruegel Policy Brief 25/2024).
[73] See Haskell and Westlake, infra 158 (arguing the need to concentrate on lowering barriers to market access to increase productivity rather than focussing on *ex post* regulation.)

Two phases of algorithmic regulation in Union law can be identified. The first, spanning over two decades, is characterised by private actors self-governing their own software. The second, more interventionist phase, originates from 2020 when the Commission issued the draft proposal of the Digital Services Act 2022. Significant in terms of new policy thinking, this second wave creates an environment where corporations with an online presence are required to provide the authorities with access to their source code and data under defined circumstances. While, as will be seen, these developments are restricted in their application, nevertheless, they may act as a harbinger for more interventionist and competition-oriented initiatives in the years to come.

1.1 Access to Code – From Self-Governing to More Interventionist Algorithmic Regulation

Prior to 2020, the European Union adopted a largely self-governing approach to algorithmic regulation. Like the 1995 Data Protection Directive (95/46/EC) before it, the 2016 General Data Protection Regulation (2016/679) acknowledges that automated processing of personal data can adversely impact individuals, and requires that data subjects are informed about the existence of automated decision-making and are given meaningful information about the logic, "significance and the envisaged consequences of such processing". However, neither law mandates that data subjects be given access to the code or parameters processing their data.[74]

Drawing from German rulings, Wachter argues that not only do decisions relating to a particular individual not have to be explained, but that the Regulation only requires providing generic information about the logic of the software.[75] Thus, in the case of AI, details including the code, feature weightings, or parameters may remain undisclosed.

A similar self-governing approach to algorithmic regulation typifies EU financial regulations. The 2014 Directive on Markets in Financial Instruments (2014/65/EU) (MiFID II)[76] addresses high frequency algorithmic trading and obligates firms to report such trading to the competent authorities and conduct annual self-assessed internal compliance tests. Furthermore, pursuant to the Commission Delegated Regulation C 2017/589, investment firms must be able to apply kill-switches to their trading software in order that they can "cancel immediately, as an emergency measure … executed orders submitted to any or all trading venues".[77] Despite the calamitous negative welfare effects instability within financial markets can have on both the economy and society, nevertheless, as with

[74] Article 13.2 (f).
[75] Sandra Wachter and others, 'Why a right to explanation of automated decision-making does not exist in the general data protection regulation' (2017) International Data Privacy Law 7 2 76, 87, n.54.
[76] Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU [2014] OJ L173/349 (hereafter MiFID II).
[77] Article 12.1. Commission Delegated Regulation (EU) 2017/589 of 19 July 2016 supplementing Directive 2014/65/EU of the European Parliament and of the Council with regard to regulatory technical standards specifying the organisational requirements of investment firms engaged in algorithmic trading [2017] OJ L87/417.

EU data protection law, MiFID II fails to provide national authorities (or any other entity) access to the algorithms themselves.[78]

Moving to the area of consumer protection law, numerous cases of algorithmic manipulation – such as search rankings or price-fixing – has prompted a legislative response from the EU.[79] The 2019 Omnibus Directive (2019/2161), part of the EU's "New Deal for Consumers" initiative, *inter alia* requires platforms to disclose on their website the logic of their ranking parameters. However, in common with other pre 2022 legislation, access to the relevant algorithms themselves is still denied.[80] Similarly, the Regulation on Promoting Fairness and Transparency for Business Users of Online Intermediary Services (2019/1150) mandates online intermediaries and platform providers to reveal ranking criteria. However, as with the Omnibus Directive, the Regulation creates no clear pathway for those affected by the ranking of goods and services to access the underlying algorithms.[81]

Motivated in part by broader existential concerns about the impact of online disinformation campaigns on the European body politic,[82] since 2020, the Commission has however started to recalibrate its self-governing approach to algorithmic regulation. As part of the programme, "A Europe Fit for the Digital Age", a new wave of EU legislation has sought to deal with digital consumer protection and market competition issues in a more interventionist manner. In the words of the DSA, the seemingly unassailable market position of the very large online platforms means they are in a unique position to "strongly influence … safety online, the shaping of public opinion and discourse, as well as … online trade."[83]

Three pieces of legislation heralded this new wave of bolder algorithmic governance. The first is the 2022 Digital Services Act (DSA).[84] Amending the Directive on Electronic Commerce (2000/31/EC), it focuses on the activities of "very large platforms"[85] in four broad areas: online dissemination of illegal content, freedom of speech, competition harms, and the manipulation of online services by bad

[78] Article 17.5 subparagraph four of MiFID II states "The competent authority of the home Member State of the investment firm may require the investment firm to provide, on a regular or ad-hoc basis, a description of the systems and controls referred to in first subparagraph and evidence that those have been applied."
[79] For an overview of algorithmic collusion see *Algorithms and Collusion* (OECD DAF/COMP/WD (2017) 4 2017).
[80] Recital 23 "Traders should not be required to disclose the detailed functioning of their ranking mechanisms, including algorithms."
[81] For example, Article 5.6 states "Providers of online intermediation services and providers of online search engines shall, when complying with the requirements of this Article, not be required to disclose algorithms or any information that, with reasonable certainty, would result in the enabling of deception of consumers or consumer harm through the manipulation of search results."
[82] Significant events include the European Parliament's investigation into the Snowden revelations, and the questioning of Mark Zuckerberg in 2018 about Facebook's role in the Cambridge Analytica scandal. In 2019 a Commission report into online manipulation, micro-targeting, psychographic profiling, and fake news, found that such techniques are well established in the online environment, citing numerous cases of their use to influence voters in the UK, Germany, US, Sweden, Hungary, Estonia, Latvia and Ukraine. *Polarisation and the Use of Technology in Political Campaigns and Communication* (European Parliamentary Research Service 2019).
[83] Recital 79.
[84] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC [2022] OJ L 277/1 (hereafter Digital Services Act or DSA).
[85] Article 33 (1) defines as average monthly users in the EU equal or greater than 45 million users.

actors. The second is the Digital Markets Act (DMA),[86] which builds on the Omnibus Directive by specifically targeting the largest digital platforms and technology providers known as "gatekeepers."[87] This ambitious piece of legislation "aims at preventing gatekeepers from imposing unfair conditions on businesses and end users and at ensuring the openness of important digital services."[88] Obligations include those relating to interoperability,[89] consumer freedom to use software of their own choice,[90] non-preferential treatment of their own goods and services,[91] pricing and advertising transparency,[92] as well as prohibitions on cross-fertilisation of personal data obtained from other services offered by them and third parties.[93]

Stemming from a growing understanding by the European legislature of the societal-wide effects that technologies have over all aspects of modern life, subject to hefty fines for non-compliance,[94] unlike previous legislation, significantly, both these Acts provide the authorities with access to any relevant documents, data, algorithms and information necessary to conduct investigations and monitor conformity with the Regulation.[95, 96]

1.2 Artificial Intelligence Act

The third Regulation in this wave of more interventionist algorithmic governance is the Artificial Intelligence Act (2024/1689). It is the first piece of legislation in the world to establish an approval scheme for AI algorithms as an *ex ante* market access requirement for AI-powered products and services. As such, it represents a further shift in how the oversight of frontier technologies is being tightened in response to existential fears about the impact of algorithms on citizens' lives.

As a first mover, like the GDPR before it, it stands to act as a standard-bearer globally for future AI regulation, undergirded by an extra-territorial effect. In lockstep with the GDPR, it governs situations where the personal data of EU citizens is used in third countries, either during training or in

[86] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 [2022] OJ L265/1 (hereafter Digital Markets Act or DMA).
[87] See Article 3 (2) (a). An organisation is a gatekeeper if it has a turnover of €7.5 billion within the EU over the last 3 years or a worldwide market capitalisation of over €75 billion. Gatekeepers must also have at least 45 million monthly end-users and 100,000 business users. Finally, a gatekeeper will provide a "core platform services" such as marketplaces and app stores, search engines, social networking, cloud services, video-sharing platforms, advertising services, operating systems, messaging services, virtual assistants and web browsers.
[88] 'Questions and Answers" (*European Commission* 6/9/2023) https://ec.europa.eu/commission/presscorner/detail/en/QANDA_20_2349.
[89] Article 6 (7).
[90] Article 6 (3).
[91] Article 6 (5).
[92] Article 5 (9).
[93] Article 5 (2).
[94] Non-compliance brings hefty fines. Under the DMA (Article 31 (1)) and DSA (Article 52 (4)) periodic penalties for non-compliance are up to 5% of average daily global turnover per day. Fines under the DSA (Article 52 (3)) are up to 6% of the previous year's global turnover, and under the DMA up to 10% for a first offence (Article 30 (1)), and 20% for a repeat infringement (Article 30 (2)).
[95] DMA. Articles 21 (1), 21 (3), 23 (2) (d), 23 (4).
[96] DSA. Article 69 (2) (d), 72 (1).

deployment, even in cases where the product or service is "not placed on the market, put into service, or used within the Union."[97]

Having its origins in various initiatives across the preceding half decade,[98] the AIA envisages both a European Artificial Intelligence Board and an AI Office which *inter alia* work alongside competent authorities in Member States to ensure compliance with the Regulation. It delineates between three different risk-based AI sectors: i) those defined as non-high risk that only have limited transparency obligations,[99] ii) those which are high risk and therefore require prior approval, and iii) those uses which are banned outright or require prior judicial authorisation.[100] The Act classifies high risk AI as software used as part of a safety component, a product covered by one or more of the twenty Regulations or Directives listed in Annex I, or a model deployed in one of the following eight sectors:[101]

1) Biometric identification and categorisation of natural persons
2) Management and operation of critical infrastructure such as road traffic and energy supplies
3) Education and vocational training
4) Uses related to employment
5) Public administration and private services in the area of benefits, emergency services, and credit-scoring
6) Law-enforcement
7) Migration, asylum, and border control management
8) Administration of justice and other law-related processes.

Notably absent are most public-facing social media platforms and e-commerce providers who remain regulated *inter alia* by the Omnibus Directive 2019/2161, the Regulation on Promoting Fairness and Transparency for Business Users of Online Intermediary Services (2019/1150), the Digital Markets Act, and the Digital Services Act. Similarly, high-frequency trading will continue to be controlled under MiFID II.

Although an exhaustive examination of the broader effects of the Regulation fall beyond the scope of this thesis, in brief, obligations on providers of high-risk AI software include maintenance of a risk

[97] Recital 11.
[98] Resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL) 2018 / C252 / 25. Communications on Artificial Intelligence in Europe. COM/2018/237 (25/4/2018). Coordinated Plan on Artificial Intelligence. COM/2018/795 (27/12/2018). European strategy for data. COM/2020/66. Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics. COM/2020/64.
[99] Chapter III. Article 6 (4) Non-high-risk providers of AI must register their software and create documentation outlining the reasoning for the designation. Recital 165 also encourages the creation of codes of conduct for programs that fall in this category.
[100] Article 5 "Prohibited Artificial Intelligence Practices", includes psyops activities, uses undertaken by public authorities based on social behaviour likely to result in prejudicial behaviour against individuals or groups, real-time biometric monitoring in public spaces etc.
[101] Annex III.

management system, prior conformity checks, and a requirement for data used on the models to be "relevant, sufficiently representative, and to the best extent possible free of errors and complete in view of the intended purpose."[102] Further obligations include human oversight,[103] registration of the system in a Europe-wide AI database,[104] and software explainability.[105] Supported by strong lobbying from the entertainment industries,[106] the Act also requires that "a sufficiently detailed summary about the content used for training of the general-purpose AI model" is made publicly available.[107] Moreover, providers of general-purpose models are obligated to "put in place a policy to comply with Union law on copyright and related rights, and in particular … a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790."[108]

Notably, the AIA also requires that providers create and maintain documentation that can be examined as part of a conformity assessment procedure.[109] According to Annex IV of the proposal, *inter alia* documentation must include a version record of the software / firmware, explanation of functionality, "the general logic of the AI system and of the algorithms",[110] an explanation of any third-party software, training methodologies, performance limitations and expected accuracy, lifecycle changes, and ongoing oversight activities. While conformity assessment documentation must be submitted to the appropriate authority, the conformity assessment procedure itself is a self-governing process for most systems that are designated as high-risk. This is because AI sectors that have harmonised standards (Annex I), and the many sectors listed in the Act (with the exception of biometric identification in the case of criminal investigations), only require an internal self-authorising check-off for conformity with the legislation.[111] Although the AIA represents a landmark piece of legislation, in practice it lacks any independent assessment; thus, the majority of the day-to-day oversight and procedural compliance takes place under the sole control of the organisation or company deploying the software. Once the company is satisfied that its software and procedures are compliant, it is then free to draw up a declaration of conformity and register its compliance in the EU database of high-risk AI systems.

In spite of the self-governing approach taken by the AIA, in common with the 2022 Digital Services Act and the Digital Markets Act, for compliance purposes the competent authorities do nevertheless upon a "reasoned request" where "all other reasonable means to verify conformity have been

[102] Article 10 (3).
[103] Article 14.
[104] Article 49.
[105] Article 13 (1).
[106] See amendment 399, 28 (b) 4 c of the AIA text adopted by the European Parliament on 14/6/2023. (P9_TA(2023)0236**).**
[107] Article 53.1 (d). See also Recital 104.
[108] Article 53.1 (c).
[109] Annex IV.
[110] Annex IV Article 2 (b).
[111] Article 5 (3).

exhausted”,[112] have a right of access to the source code[113] and “where relevant” any “training, verification or testing dataset”[114] used in the creation of the model. In terms of transparency, members of the public will also have access to the yet to be designed EU database of registered high-risk AI, although pursuant to Article 71(4) only limited information about the model will be made publicly available.[115]

While this more interventionist approach will no doubt result in a greater protection for EU citizens, in giving powers to a national competent authority to analyse the functioning of the AI software, several significant policy assumptions have been made. Most obvious is that the investigating bodies have the requisite experience and expertise to be able to manage the technical analysis. As highlighted by the UN’s High Level Panel on Digital Cooperation, this is something which many public organisations struggle with.[116] This includes not only the computer skills required to analyse potentially highly complex models, but also the knowledge required to understand how the software interacts with the data and architectures within which they operate. Whether this activity is undertaken in-house by the competent authorities or contracted to third parties, for these measures to be at all effective they require a significant budgetary investment by the state in trained personnel as well as in the independent technical infrastructures needed for evaluating the algorithms under investigation.

Where responsibilities are divided between different organisations, the digital cooperation architecture will inevitably be challenging. For example, certain services such as Microsoft’s LinkedIn will fall under the DSA, DMA, as well as the AIA.[117] Where issues are cross-cutting as well as legally and technically complex, coherent and effective governance can be particularly challenging.

In conclusion, the mounting evidence of online injustices crystallised in the public consciousness following the Snowden and Cambridge Analytica scandals, has brought about a step change in the European Union’s approach to algorithmic regulation. The EU has moved from a trust-based model of governance to an environment where the authorities and appointed experts now have a legal basis for accessing and evaluating AI software and any relevant data. Whereas this is a significant step forward in the regulation of the dangers inherent in digital markets, whether the statutory rights provided to competent authorities proves effective or not remains to be seen.

From the standpoint of the market, to effectively compete with powerful technology firms requires more than just transparency and procedural refinements to market trading norms. Merely accepting

[112] Annex VII Article 4.5.
[113] Articles 74 (13), 92 (3) and Annex VII Article 4.5.
[114] Annex VII Article 4.3.
[115] Article 71 (4).
[116] *The Age of Digital Interdependence. Report of the UN Secretary-General’s High-level Panel on Digital Cooperation.* (UN 2019) 23.
[117] LinkedIn provides employment related services falling under Annex III §4.

the dominant position of AI-reliant superstar companies and seeking to improve their behaviours falls short of an effective innovation strategy. A second, much-needed regulatory pillar should focus on fostering greater economic dynamism in AI-reliant markets. However, as can be seen from the suite of algorithmic regulations explored above, a coordinated legislative innovation strategy intended to nurture more effective competition in markets dominated by powerful AI incumbents appears largely absent in the EU.

Section Three – Characteristics of AI Markets

1.1 The Market Concentration of AI

Having outlined the scope of algorithmic regulation in the EU beyond that of copyright law, this and the following sub-section examine the structural characteristics of a globally imbalanced AI market. As with previous general-purpose technologies, it is evident that the benefits of machine learning are far from being evenly spread. Korinek and Stiglitz assert that "it is hard to conceive of an AI economy being competitive."[118] As will be shown, several differing factors uniquely coalesce to create both AI superstar companies and, particularly in the area of e-commerce, unassailable winner-takes-all markets.

The market concentration, and attendant global economic dislocation this causes, warrants urgent and sustained attention from policy makers. The fact that the economic dominance of a small cohort of unassailable companies, primarily headquartered in the United States and East Asia, is intertwined with IP and copyright in a number of important ways indicates that copyright law should demand broader regulatory attention than it has to date. Haskell and Westlake contest that market disequilibrium is compounded by a high percentage of these companies' value being represented by intangible assets such as know-how, R&D and, particularly relevant for this study, intellectual property.[119] This private appropriation of knowledge in the form of a concentration of intangible assets becomes, in turn, a driver of both market and capital concentration.[120] Secondly, AI companies are highly reliant on data – much of which is protected by copyright and related rights. As this thesis posits, this dynamic renders intellectual property law – and in the context of artificial intelligence in particular, copyright law – a logical focal point for policy makers.

[118] Anton Korinek and Joseph Stiglitz, *Artificial Intelligence, Globalization, and Strategies for Economic Development.* (National Bureau of Economic Research 2021) 15.
[119] Haskell, supra 43. *Intangible Asset Market Value Study* (*Ocean Tomo*, 2020) https://www.oceantomo.com/intangible-asset-market-value-study/ (The merchant bank Ocean Tomo estimating that 90% of the S&P 500 market capitalisation is represented by intangible assets such as intellectual property.)
[120] See Cecilia Rikap, *Capitalism, Power and Innovation: Intellectual Monopoly Capitalism Uncovered* (Routledge 2021) 1.

Analysis shows that digitalised markets, and particularly those reliant on data, are typified by the following phenomena which drive both market and geographic concentration:

i) natural monopolies
ii) network effects
iii) economies of scale
iv) economies of scope
v) a heavy reliance on software and data
vi) high levels of intangible assets

Before discussing the geographic concentration of AI in the following subsection, this section will examine each of these six facets of the AI economy in turn. In doing so, it will underline that lack of competition is such a notable feature of artificial intelligence markets, that without a concerted effort from global policy makers to address the economic power and high levels of profits that accrue in the hands of a tiny number of powerful AI superstates, the existing imbalances in the global economy are set to worsen. Importantly, the analysis also serves to spotlight that IP – and from the perspective of this thesis particularly copyright – percolates strongly through both factors five and six. Thus, serving to underscore why more attention to copyright law as an economic regulator of AI markets is desirable.

1.2 Natural Monopolies and Network Effects

Technology markets have a strong tendency to create natural monopolies.[121] According to Mosca, natural monopolies occur in two distinct scenarios.[122] The first is where there is a need to invest large amounts in infrastructure in order to provide products and services. High start-up costs pose a barrier for all market entrants including firstcomers, however, competition is frustrated by the presence of an incumbent whose sunk costs allow them to provide goods and services at lower marginal costs than subsequent market entrants.

For competitors, the issue of high infrastructural costs is compounded by a strategy whereby first comers may run large losses for many years, which potentially deter competitors from even entering a market. Tesla, for example, founded in 2003, and despite having spent nearly $10 billion on research

[121] Korinek and Stiglitz, supra 118, passim. Manuela Mosca, 'On the Origins of the Concept of Natural Monopoly' (2006) Università di Lecce Department of Economics Working Paper 92/45 7. Francesco Ducci, *Natural Monopolies in Digital Platform Markets* (CUP 2020) passim (observing that digital technologies have a tendency to create natural monopolies.)
[122] Ibid. Mosca.

and development,[123] only returned a full year of profit for the first time in 2020.[124] Similarly, it took Amazon six years to become profitable,[125] while Uber (founded in 2009) only returned a profit for the first time in 2023.[126] Few companies, would be able to envisage, let alone sustain, for example, the $5.61 billion loss Google made in its investments into data processing centres in 2020.[127]

Secondly, natural monopolies can be driven by the presence of network effects, something competition authorities have been particularly concerned with in relation to digital platforms.[128] With their common reliance on data, it should be noted that many of the same features that characterise and drive the market concentration of online platforms are also hallmarks of other AI markets. e.g. driverless cars, internet of things, robotics, etc.

Direct network effects, where a good or service increases in utility the more people use it, are often cited as a reason why the so-called superplatforms have been able to reach a "winner-takes-all" position in their chosen marketplace(s). However, many of these companies have also grown by virtue of indirect network effects. Indirect network effects occur where one user group (e.g. a consumer) attracts another actor (often businesses and sellers in the case of online platforms) which results in the creation of a new market.[129] For example, the more people that have an iPhone and use the App Store, the more companies design and create software for iPhones. This in turn creates more of an incentive (alongside familiarity with Apple's own software and systems) to continue to stay and upgrade to later generations of the iPhone – a phenomenon known as single-homing. Leaving one provider for another is challenging as it will mean a user has to not only acquaint themselves with new technologies, re-input a range of new information and data, but also potentially lose much of the information – and in that sense personal history – they uploaded and shared with friends, family and others. The switching costs, combined with the indirect network effects, lead to a situation where companies grow and attract more users, and customer churn is relatively low compared with non-digital businesses. Furthermore, in AI markets, as a company grows due to both direct and indirect network effects, they are afforded comparative market advantage by virtue of the volume of (often unique) data they have unique access to.[130]

---

[123] 'Research and development expenses of Tesla' (*Statista* 2022) https://www.statista.com/statistics/314863/research-and-development-expenses-of-tesla/
[124] 'Net loss of Tesla' (*Statista,* 2022) https://www.statista.com/statistics/272130/net-loss-of-tesla/
[125] 'Silicon Insider: When Did Amazon Get Profitable?' (*ABC News,* 29/1/2002) https://abcnews.go.com/Business/story?id=87393&page=1
[126] Jasper Jolly and Graeme Wearden, 'Landmark moment as Uber unveils first annual profit as limited company' (*The Guardian,* 7/2/2024).
[127] Sabine Weston, 'Google Cloud lost £4.1 billion in 2020' (*IT Pro,* 3/2/2021) https://www.itpro.co.uk/cloud/358506/google-cloud-lost-56-billion-in-2020
[128] Ducci, supra 121, passim.
[129] Bruno Jullien and Wilfred Sand-Zantman, *Network Effects* (Industrial Economics Institute 2016) 12-16 (setting out indirect network effects in analogue and digital markets.)
[130] Massimiliano Nuccio and Marco Guerzoni, 'Big data: Hell or heaven? Digital platforms and market power in the data-driven economy' (2019) 23 Competition & Change 3 312, 313.

Network effects are a major contributor to markets "tipping",[131] a situation where markets favour one dominant firm. For example, Google in search, Amazon for online ordering, and Facebook within certain segments of the social media market.[132] Once a market has tipped in favour of one player, out competing the market incumbent becomes extremely challenging, if not impossible.[133]

1.3 Economies of Scale

In addition to natural monopolies and network effects, according to Crémer (2019), a further defining feature of digital markets is extreme returns to economies of scale.[134] Economies of scale "are the cost advantages that companies obtain as the number of units produced increases."[135] Once a company has sunk its costs into establishing its infrastructure, the relatively low marginal cost of production and adding new users and customers thereafter can strongly discourage new market entrants, thereby potentially leading to the establishment of a natural monopoly. This is especially the case in markets where network effects are prevalent such as online markets, where upfront investment costs and fixed costs in the form of labour form the bulk of operational expenditure.[136] The ability to serve additional customers at extremely low to zero marginal cost, combined with network effects, has created extreme levels of market concentration whereby only a very small number of superstar platform providers can compete for markets that in some instances encompass billions of global consumers.

1.4 Economies of Scope

Fourth, digital markets are characterised by economies of scope. A term first coined by Panzar and Willig,[137] the term refers to the creation of products and services that are either made less costly or improved by combining or sharing production and resource. In terms of AI, the reutilisation of ML models, data, software and technical infrastructure across a corporation's various businesses, affiliates, and subsidiaries generate economies of scope. Amazon, for example, developed its platform service AWS out of its own internal logistics infrastructure. Similarly, Google's 2014 acquisition of the biotech firm DeepMind has helped improve its deployment of ML technologies across the Alphabet Group.[138] By cross-fertilising data across a business (e.g. Meta, Facebook, WhatsApp and

---

[131] Bruno Jullien and Wilfried Sand-Zantman, 'The economics of platforms: A theory guide for competition policy' (2021) Information Economics and Policy 54 100880.
[132] Özlem Bedre-Defolie and Rainer Nitsche, 'When Do Markets Tip? An Overview and Some Insights for Policy' (2020) 11 Journal of European Competition Law and Practice 10 610.
[133] Patrick Barwise and Leo Watkins, 'The Evolution of Digital Dominance' in Martin Moore and Damian Tambini (eds), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (OUP 2018) 22 ff. (outlining how dominant players cannot be replaced but can be eclipsed by the development of bigger adjacent markets.)
[134] Jacques Crémer and others, *Competition Policy for the Digital Era* (European Commission 2019) passim. See also Ducci, supra 121.
[135] Harald Øverby and Jan Arild Audestad, *Introduction to Digital Economics: Foundations, Business Models and Case Studies* (Springer 2021) 79.
[136] Peter Earl and Tom Mandeville, 'The Competitive Process in the Age of the Internet' (2009) 27 Prometheus 3 195, 196.
[137] John Panzar and Robert Willig, 'Economies of scope' (1981) 71 The American Economic Review 2 268.
[138] See 'Improving billions of people's lives' https://deepmind.com/impact#real_world_impact (showing the use of DeepMind's technologies by Google, YouTube, Google Health etc.) See also infra n.162-166.

Instagram sharing data,[139] Microsoft having access to LinkedIn's data[140] and so on), digital oligopolies can create significant market advantage offering new insights that otherwise would not be possible.[141]

In sum, the deployment of technologies horizontally across different businesses, creating something greater than the sum of its parts, further reinforces the unassailable economic lead held by a handful of the world's largest digital companies.

1.5 AI and Data

As already touched on, and of particular relevance to this thesis given it is often subject to copyright,[142] a fifth characteristic of the AI economy is the defining role of data.[143] Aided by network effects, incumbent digital firms reinforce their market dominance through ever greater volumes of data compared to their competitors.[144] As a consequence, this contributes to the production of more accurate algorithms resulting in increasingly more sophisticated products and services.[145] This positive feedback loop allows dominant players to further take advantage of network effects as well as economies of scope and scale, which in turn improve their algorithms and data extraction capabilities. This virtuous cycle creates what Rikap refers to as data-driven innovation rents.[146] Thus, alongside the other factors already discussed, acting to further reinforce market concentration. [147]

Economies of scale and scope, cutting edge technology, combined with comparatively greater access to data, makes competition costly. For example, vying with Google in the search market took Microsoft from 2005 to 2015 to achieve profitability with Bing, with losses estimated well in excess of $11 billion.[148] The ability of Google to return better search results and therefore attract more advertising revenue in comparison to offerings from Microsoft was in no small part thanks to Google's superior machine learning capabilities, grounded in the data advantage it had built up from its early entry into the search market.[149]

---

[139] Kate O'Flaherty, 'All the data WhatsApp and Instagram send to Facebook' (*Wired,* 24/07/2021) https://www.wired.com/story/whatsapp-instagram-facebook-data/#:~:text=Data%20WhatsApp%20can%20share%20with,%2Dcalled%20"Facebook%20companies"
[140] Case M.81-24 Microsoft / LinkedIn (6/12/2016).
[141] Inge Graef, 'When Data Evolves into Market Power – Data Concentration and Data Abuse under Competition Law' in Moore and Tambini, supra 133, 72 ff.
[142] Also in Europe, *sui generis* database rights.
[143] Crémer. supra 134. Ducci, supra 121.
[144] Lina Khan, 'Sources of Tech Platform Power' (2017) 2 Geo L Tech Rev 325, 326.
[145] Ducci, supra 121, 23 (pointing out that the scale effects of data, in some markets, will taper off after a critical mass has been established.)
[146] Rikap, supra 120.
[147] Barwise and Watkins supra 133, 29. Ducci, supra 121, 4, 20 ff, 41.
[148] Brad Reed, 'Microsoft has lost $11 billion trying to compete with Google' (*Yahoo News,* 9/7/2013). https://www.yahoo.com/news/microsoft-lost-11-billion-trying-compete-google-185008151.html
[149] Nathan Newman, 'Search, antitrust, and the economics of the control of user data' (2014) 31 Yale J on Reg 401, 421.

Although competition authorities and the European Commission have particularly focused on online markets, the unique combination of factors that induce a lack of competition are not the sole preserve of these markets alone. Automated technologies such as IoT also use AI and data to improve the nature of their product offering and to reinforce market position. Taking Tesla as just one example, the data it collects from its cars as well as its network of charging stations are utilised and fed back to its machine learning models to improve its assisted driving software.[150] On top of this, the data is used to further refine its battery capabilities,[151] and improve both traffic congestion and vehicle safety for its customers.[152] Thus, network effects have the potential to be a feature of IoT markets also, allowing early entrants to create a comparative data advantage off which they can leverage better products and services and thus further entrench market position.

1.6 Intangible Assets and IP

Finally, the oligopolies that are active in the digital economy are noted for their high levels of intangible assets.[153] The international standards body IFRS defines an intangible asset as "an identifiable, non-monetary asset, without physical substance".[154] Classic examples include intellectual property, know-how, customer lists, domain names, and R&D. According to the merchant bank Ocean Tomo, the share of the capitalisation of intangibles has grown from 17% in 1975, to now over 90% of total value of the S&P 500.[155]

Focusing on registered IP rights as one indicator of market power in AI-reliant markets, unsurprisingly large technology firms from the United States and East Asia dominate. The OECD highlights that of the top fifty companies with the highest percentage of AI-related patent and trademark registrations in their total IP portfolios, US and Chinese superplatforms (Alphabet, Amazon, Microsoft, Baidu, Tencent) as well as Japanese multinationals (Honda, Fujitsu, Toshiba, Sony, Canon, Toyota, Panasonic, Fuji etc) dominate.[156] A further indicator of the perceived importance of IP in certain AI market, is the volume of registration of AI-related patents. According to WIPO, of the top 10 applicants by number of patent families, three were from the US (IBM, Microsoft, Alphabet), one Korean (Samsung) and the remaining six Japanese (Toshiba, NEC, Fujitsu, Hitachi, Panasonic, Canon).[157]

---

[150] Rabah Amir et al, 'Oligopoly with network effects: firm-specific versus single network' (2012) 71 Economic Theory 3 1203, n.6 (observing the indirect network effects involved in the growth of TESLA.)
[151] Peter Schick, 'TESLA business model: innovation, scale and network effects' (18/1/2022) https://peterschick.eu/tesla-business-model-innovation-scale-and-network-effects/
[152] Kyuho Maeng and others, 'Network effects of connected and autonomous vehicles in South Korea: A consumer preference approach.'(2021) Research in Transportation Economics 90 100998.
[153] See Erik Brynjolfsson and others, 'Intangible assets: Computers and organizational capital' (2002) Brookings Papers on Economic Activity 137 (evaluating the link between high levels of IT related intangible assets and high market capitalisation.)
[154] See https://www.ifrs.org/issued-standards/list-of-standards/ias-38-intangible-assets/
[155] Ocean Tomo, supra 119.
[156] *Who Develops AI-Related Innovations, Goods and Services?* (OECD Science, Technology and Industry Policy Papers 121 2021).
[157] *WIPO Technology Trends 2019 – Artificial Intelligence* (WIPO 2019) 60.

Combining AI-related IP registration information with global listings of market capitalisation clearly demonstrates the nexus that exists between high market valuations, digitalisation, and the legal protection of software and data that powers AI. From a regulatory perspective, as examined later in the thesis, new developments in trade law are also serving to further reinforce market concentration in machine learning markets. As highlighted by Haskell et al and Rikap, the private appropriation of knowledge in intangible assets, often undergirded by intellectual property rents – the benefits of which frequently bypass less developed countries – has resulted in a global decline in productivity and market competition.[158]

1.7 The Geographic Concentration of AI

A further notable facet of AI markets is their geographical concentration. This is of critical relevance to this thesis, as it contends that in contrast to Europe, the more flexible approach to limitations and exceptions taken by Japan and the United States fosters technological innovation. This, in turn, reinforces the first mover advantage and economic power of these two export-oriented AI superstates. In addition to highlighting the connection that exists between copyright law and uneven economic development within information technology markets, an understanding of the geographic concentration of machine learning further calls attention to the underlying economic and political power structures that have driven the para-copyright software and algorithm disciplines in international trade law that are discussed in Chapter 6. Furthermore, from the perspective of policy development, it provides essential regulatory context for late-comer nations seeking to amend their copyright law in response to this newest of general-purpose technologies.

1.7.1 United States

At the time of writing, nine of the world's ten largest companies were AI intensive companies. With an aggregate value of $15.84 trillion, all are headquartered in the United States with the exception of TSMC. Unsurprisingly, Google, Apple, Facebook, Amazon, and Microsoft (GAFAM) feature prominently.[159] Examining the strategic use of AI by these five companies reveals that they are not only integrating machine learning technologies into their core markets – online search and advertising (Google), consumer electronics (Apple), social media (Facebook), retail (Amazon), and software (Microsoft) – but are also extending them across their vast portfolio of subsidiaries. According to one study, the superplatforms have huge reach beyond their core markets with Google making acquisitions in 187 sectors, Apple in 96, Facebook in 77, and Amazon in 76.[160] Taking one example

[158] See Jonathan Haskell and Stian Westlake, *Restarting the Future* (PUP 2022) and Rikap, supra 120.
[159] On 28th November 2025 nine of the world's top ten most valued companies nine are heavily involved in AI: Apple, Microsoft, Alphabet, Amazon, Tesla, Meta, NVIDIA, Broadcom and TSMC. NYSE.
[160] See Chris Alcantra and others, 'How Big Tech got so Big: Hundreds of acquisitions' *The Washington Post* (21/4/2021).

alone, Google's holding company, Alphabet, cross fertilises its artificial intelligence technologies across numerous of its subsidiaries,[161] including DeepMind (health), Google's data centres,[162] Waymo (driverless cars),[163] Verily (health),[164] Firebase (mobile and web applications),[165] and Calico (life sciences).[166] In addition to market capitalisation, other economic indicators attest to US dominance in artificial intelligence markets. For example, it is estimated that in 2022 the US accounted for approximately 54% of all private AI investment[167] and, according to one industry report, 73 of the world's top 100 AI start-ups are based in the US.[168]

1.7.2 East Asia

Alongside the US, the AI economy is dominated by powerful countries in East Asia. The US and China comprise 90% of market capitalisation value of the world's 70 largest digital platforms.[169] China has its own AI superstar companies: Baidu, Alibaba, Tencent, and Huawei (BATH). While they operate across the same sectors as GAFAM, with the exception of Huawei, they are focused primarily on their own domestic market of 1.4 billion people. After the US, China is the second largest recipient of private AI investment globally at 15.5 %.[170] Furthermore, taking R&D as an indicator of market power, between 2016 and 2020 nearly twice the number of AI publications originated from China than from the US.[171]

Elsewhere in East Asia, Japan excels in the application of machine learning, particularly in the area of Internet of Things (IoT), vehicle automation, and driverless cars. With offerings from household names such as Toyota, Nissan, NEC, Panasonic, Fujitsu and Honda, Japan is one of the world's largest exporters of AI automated technologies. It also leads in robot technology with 7 of the world's top 10 robotics companies based there, supplying 32.3% of all exported industrial robots.[172]

1.7.3 Europe

As frequently observed, Europe is failing to compete effectively in the global AI race. Whereas between 2000 and 2019, European academics have published more journal articles than either the US

[161] https://bdtechtalks.com/2021/10/07/google-deepmind-2020-earnings/
[162] https://deepmind.google/discover/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-by-40/
[163] https://deepmind.google/discover/blog/how-evolutionary-selection-can-train-more-capable-self-driving-cars/
[164] https://support.workbench.verily.com/docs/guides/cloud_environments/alphafold/ (deploying Deepmind's Alphafold)
[165] https://blog.google/technology/ai/google-io-2024-100-announcements/ (deploying Deepmind's Gemini)
[166] https://deepmind.google/discover/blog/predicting-gene-expression-with-ai/
[167] Fig. 4.2, 10 *AI Index Report 2023* (Stanford University 2023).
[168] *AI 100: The Most Promising AI Startups of 2022* (CB Insights 2022).
[169] *Digital Economy Report 2019: Value Creation and Capture: Implications for Developing Countries* (UNCTAD 2019) 2.
[170] Stanford University, supra 167, Chapter 4.
[171] 'AI-related publications worldwide 2016-2020, by country' (*Statista*, 10/3/2023) (Between 2016-2020 China published 76,300 articles on AI compared to 44,300 in America.)
[172] Daniel Workman, 'Top Industrial Robot Exporters' (*World's Top Exports*) https://www.worldstopexports.com/top-industrial-robots-exporters/

or China, it struggles to translate this into economic impact – the so-called innovation paradox.[173] In the first half of 2024, the EU only attracted 6% of global investment in AI.[174] Not a single of the top 10 companies currently investing in quantum technologies is based in the EU.[175] Furthermore, in 2024, only 3 of the 58 most notable AI models originated from Europe.[176]

In terms of the long-term economic outlook for Europe, the vibrancy of an economy is often mirrored in its start-up, scale-up, and SME sectors. Attesting to the weak position Europe finds itself in AI technology markets, it is noteworthy that in 2020 only 5 out of the world's 100 most promising AI start-ups were based in the EU.[177]

1.7.4 The Global South and the Northward Retreat of Global Value Chains

Beyond Europe, the distribution of AI activity is even more uneven. In 2022, none of the top 100 AI startups identified by CB Insights were based in Africa, Oceania, or South America. Aside from China, only one originated in Asia.[178] Similarly, among the 10 leading recipient nations of private AI investment, only China and India are situated in the Global South.[179] Research output also reflects this imbalance, with an analysis of top machine learning conference publications showing no representation from Africa, Latin America, Oceania, or the Middle East (excluding Israel).[180]
Economic theory traditionally posits that labour arbitrage and global value chains (GVCs) facilitate technology transfer and the diffusion of benefits to the Global South. This includes both formal mechanisms (e.g. licensing, exposure to international standards) as well as informal channels (e.g. training, workforce mobility, etc). From the perspective of intellectual property law, the raison d'être of the TRIPS Agreement was the desire to protect foreign direct investment, and as a corollary, the positive externalities that arise from knowledge transfer.[181] This goal is explicitly encapsulated in Article 66.2 and the obligation on developed countries to "provide incentives to enterprises and institutions in their territories for the purpose of promoting and encouraging technology transfer to least developed country members in order to enable them to create a sound and viable technological base."
Recent evidence however challenges these traditional assumptions. UNCTAD research shows that the international footprint of e-commerce firms is comparatively asset light as they are able to serve their

---

173 *Artificial Intelligence Index Report 2021* (Stanford University 2021) 20.
174 Martens, supra 72.
175 Draghi, supra 71, 80.
176 *Artificial Intelligence Index Report 2025* (Stanford University) Fig. 7.
177 Source: *The World Factbook Field Listing: Exports – Commodities* (Central Intelligence Agency). Quoted in Daniel Workman, 'Top Industrial Robots' https://www.worldstopexports.com/top-industrial-robots-exporters/
178 *AI 100: The Most Promising AI Startups of 2022* (CB Insights 2023).
179 *Artificial Intelligence Index Report 2023* (Stanford University 2024) 190.
180 Gleb Chuvpilo, 'AI Research Rankings 2020: Can the United States Stay Ahead of China?' (20/12/2020). https://chuvpilo.medium.com/ai-research-rankings-2020-can-the-united-states-stay-ahead-of-china-61cf14b1216
181 Adronico Adede, 'Origins and History of the TRIPS Negotiations' in Christophe Bellmann and others, *Trading in Knowledge: Development Perspectives on TRIPS, Trade and Sustainability* (Routledge 2003). (See for a detailed history of the TRIPS agreement).

customers online with fewer foreign investments and fewer employees abroad.[182] As a consequence, where companies employ less people outside of their core markets, the level of both formal and informal knowledge transfer is reduced, serving to further concentrate digital markets and the intangible assets they spawn.

Compounding this further is the rise of back-shoring in high-tech manufacturing – a particular feature in robotics-intensive sectors.[183] According to the OECD, this trend is likely to be a massive "game changer, reversing the importance and length of Global Value Chains and reorienting global production and trade back towards OECD economies."[184]

In conclusion, while the widespread adoption of AI across both established and newer sectors of the economy attests to its status as a general-purpose technology, its economic benefits are far less evenly spread. Although studies estimate that by 2030 the contribution of AI to the global economy will range from $13 trillion[185] to $15.7 trillion[186] – 14% higher than would be the case without AI – the economic rewards will mostly accrue to a small number of companies predominantly located in North America and East Asia. Furthermore, the combination of back-shoring and the light international footprint of digital corporations indicate that countries are less likely to experience the same level of knowledge transfer as in the past. Ineluctably, this will exacerbate the geographic disequilibrium inherent in technology markets and frustrate the development goals of the many countries in the Global South that have traditionally benefited from the public welfare gains that arise from foreign investment.

Conclusion

The considerable wealth of the world's largest companies tied to intangible assets, combined with their strong dependence on data of all types, prompts the policy question of whether intellectual property – and as addressed in this thesis particularly copyright – should receive greater regulatory attention from governments and international bodies such as WIPO than has hitherto been the case. As we see in Europe, while heterogenous forms of competition law influenced *ex post* and *ex ante* regulation in technology markets can be envisaged, critically, anti-trust interventions fail to address the core issue of barriers to market access.[187] As a regulator of not just AI software and its outputs,

[182] *World Investment Report 2017: Investment and the Digital Economy (*UNCTAD 2017) 164 ff.
[183] Bernhard Dachs and others, 'Bringing it all back home? Backshoring of manufacturing activities and the adoption of Industry 4.0 technologies.' (2019) 54 Journal of World Business 10 10, 17.
[184] K De Backer and D Flaig, *The Future of Global Value Chains: Business as Usual or 'A New Normal'?* (OECD Science, Technology and Industry Policy Papers 41, 2017).
[185] *Notes from the AI Frontier. Modelling the Impact of AI on the World Economy* (McKinsey Global Institute 2018).
[186] *Sizing the prize. What's the real value of AI for your business and how can you capitalise?* (PWC 2017).
[187] Haskell and Westlake, supra 158, Chapter 7 (arguing that the lack of competition is the result of the importance of intangible assets and should be addressed by reducing barriers to market entry rather than antitrust law approaches such as corporate breakups.)

but a company's ability to acquire and create the training data necessary to power its algorithms, copyright law operates across critical domains that help reinforce information monopolies and the rentier behaviour of technology market incumbents.

As discussed in the proceeding chapters, by facilitating access to proprietary software and the data upon which ML depends, limitations and exceptions to exclusive rights have the potential to promote innovation and competition by lowering barriers to market entry.[188] Promoting viable competition in this way will *a priori* reduce the tendency that technology markets have to tip, and ameliorate the high levels of market and capital concentration that are a prominent feature of the digital economy.

[188] Pamela Samuelson, 'Justifications for Limitations and Exceptions' in Ruth Okediji (ed), *Copyright Law in an Age of Limitation and Exceptions* (CUP 2017) 35 ff. (outlining how a number of L&Es accomplish economic goals.)

Chapter 2: UK and the European Union – Text and Data Mining and Computer-Generated Works

Introduction

As examined in Chapter 1, the majority of European legislative interventions relating to AI to date have targeted the economic or social harms inherent in algorithmic trading and winner-takes-all technology markets. However, a more positive and innovation focused agenda can be found in copyright law. The first AI-related copyright intervention in Europe can be traced back to 1988 with the UK's introduction of a provision for computer-generated works (CGW).[189] The second, more recent development, has been the introduction of text and data mining (TDM) limitations and exceptions into statute, starting again with the UK in 2014.

The inherent nature of copy-reliant digital technologies has expanded the applicability of copyright law's reproduction right far beyond its original purpose.[190] Whereas traditionally copyright provided users with significant freedoms to utilise the works of others unencumbered, with the advent of network technologies, predicated as they are on the routine making of copies, copyright has metamorphosed into an all-encompassing regulator of modern life. As a consequence, the reproduction right has become a problematical concept as it struggles to differentiate between economically significant copies and those of a purely mechanical and procedural character.[191]

Copyright exceptions of a digital nature (such as Article 5.1 the Information Society Directive (2001/29/EC)), tacitly reinforce the hegemonic role copyright law now plays in regulating nearly all forms of modern knowledge production. For example, given the broad scope of the reproduction right in Union law (Article 2), the automatic copying undertaken by all hardware and software means that an exception to copyright's exclusive right of reproduction is required as an antecedent to accessing the underlying information – even when the information itself lies in the public domain.[192]

Despite this broad extension in the application of the reproduction right constituting a material encroachment on the public domain, unless copyright law develops to exclude entire classes of mechanical reproduction, policy makers must for the time being at least operate within the legislative

---

[189] s.9(3), CDPA.

[190] The term "copy-reliant technologies" was first popularised by Matthew Sag, 'Copyright and Copy-Reliant Technology' (2009) 103 Northwestern University Law Review 1607.

[191] See for example William Patry, *How to Fix Copyright* (OUP 2012) (arguing that exclusive rights should not prevent copying where undertaken as a necessary adjunct to a lawful act.)

[192] cf. Martin Senftleben, 'Compliance of national TDM rules with international copyright law: an overrated nonissue?' (2022) 53 IIC-International Review of Intellectual Property and Competition Law 10 1477 (arguing that in the light of the Diplomatic Conference in 1996 running up to the adoption of the WIPO Copyright Treaty 1996 (hereafter WCT) and WIPO Performances and Phonograms Treaty 1996 (hereafter WPPT), TDM provisions fall outside of the scope of the reproduction right enshrined in Art 9(1) of the Berne Convention. As a result, it is suggested WIPO Member States are free to enact legislation as they see appropriate with no reference to international treaties or the Three Step Test.)

framework at their disposal. In Europe, this means recourse to a copyright framework that explicitly regulates all forms of technical reproduction, including both transient and incidental copies hidden from the human eye.[193]

Broadly speaking, the aim of TDM exceptions is to support new forms of knowledge acquisition and data-driven innovation. At the macro-economic level, this translates into policy designed to stimulate and nurture competitiveness and innovation in digital markets.[194] To determine how successful Europe's AI policy is likely to be, this chapter will analyse the extent to which the legal response meets the goal of encouraging the greater adoption of AI technologies in Europe. Subsequent to outlining the restricted acts in regards to which exceptions have been created, and discussing the interpretation of the term text and data mining, analysis is divided into the following categories: i) beneficiaries of the exception, ii) the categories of works covered, iii) storage and security, iv) communication to the public and re-utilisation, v) the interplay of the exceptions with licences and private ordering, and vi) technological protection measures. Following this, attention is directed to the 1988 UK decision to extend copyright protection to autonomously generated computer-created works, and the extent to which it confers UK investors with an advantage over their foreign rivals. A high-level summary of the key TDM and AI provisions in UK and Union law is provided to aid with interpretation.

Section One

1. Limitations and Exceptions to Restricted Acts – UK and the EU

In 2014, the UK was the second country in the world after Japan[195] to introduce a specific limitation and exception for the purposes of computational text and data analysis.[196] Five years later the European Union notably adopted two mandatory exceptions for TDM under Title II of the Copyright in the Digital Single Market Directive (CDSMD).[197] Article 3 guarantees cultural heritage organisations (CHIs)[198] and not-for-profit / public interest research organisations the ability to

[193] Article 5 (1). Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L 167/10 (hereafter InfoSoc or Information Society Directive).
[194] See for example 'Proposal of The Netherlands, Belgium, Germany, Finland and the Slovak Republic for an additional and optional exception for Text- and Datamining' WK 12354/2017 INIT 31/10/2017 (sic) stating the purpose of Article 4 of the CDSMD is for "unleashing the potential of artificial intelligence, Big Data and innovation, which are important policy objectives of the European Union…".
[195] Article 47-7. Reproductions etc for the purpose of data analysis, Act No 53 of 19 August 2009.
[196] Section 29A. The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014 (SI 2014/1372) (1/6/2014).
[197] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L130/92 (hereafter DSM Directive or CDSM Directive).
[198] When Article 3 was first introduced by the Commission in 2016, research organisations were the sole beneficiary of the exception. Responding to calls from the European and international library community the final Article was subsequently broadened to incorporate cultural heritage institutions.

undertake machine learning under certain prescribed conditions. By contrast, Article 4 is open for all to enjoy, but unique to European copyright law it permits rightsholders to reserve their exclusive

	Beneficiary of Exception	**Purpose of Exception**	**Interplay with Contractual Terms**	**Interplay with Technological Protection Measures**	**Other Provisions Relating to Exceptions**	**References to Computer Generated Works**
Copyright in the Digital Single Market Directive 2019/790 (CDSMD)	Art 3. Research organisations and cultural heritage institutions. Art 4. Any legal or natural person.	Art 3. Scientific research (human and natural sciences) Art 4. No purpose limitation. However, a rightsholder can reserve their exclusive rights ("reservation of rights") negating the exercise of the exception.	Art 3. Contractual terms cannot override the exception. Art 4. Contractual terms can override the exception if expressed "in an appropriate manner".	Arts 3 & 7. Allows circumvention subject to provisions in national law. Arts 4 & 7. Allows circumvention subject to provisions in national law.	Art 7.	No explicit reference. Art 2 however provides no limitation in regards to the types of information generated.
Artificial Intelligence Act 2024/1689	N/A	N/A	N/A	N/A	Art 53.1 (c) Providers of general purpose models must put in place policies to comply with Union law, including the reservation of rights set forth in Art 4 of the CDSMD.	Recital 99 describes the kinds of outputs Generative AI produce. Recital 105 describes the reasoning behind the obligation set forth in Art 53.1 (c).
Copyright, Designs and Patents Act 1988	s.29A. Any legal or natural person	Non-commercial research	Contractual terms cannot override the exception	s.296ze(9)Authorisation from government possible for circumvention except when the works are subject to contractual terms and "members of the public may access them from a place and at a time individually chosen by them."		s.9 The author of a computer generated literary, dramatic, musical or artistic work is the "person by whom the arrangements necessary for the creation of the work are undertaken." s.12(7) copyright expires 50 years from the end of the calendar year in which the work was made.

Fig. 2-1. A Simplified Table of the Provisions in UK and EU Copyright Law Relating Explicitly to Machine Learning

rights and deny beneficiaries the legal means to enjoy the exception. Furthermore, whether Member States may provide for remuneration pursuant to Article 4 is open to question.[199]

[199] See 'Comment of the European Copyright Society Addressing Selected Aspects of the Implementation of Articles 3 to 7 of Directive (EU) 2019/790 on Copyright in the Digital Single Market.' (European Copyright Society 3/5/2022).

In terms of derogations to exclusive rights, both the UK and CDSMD provisions create an exception to the right of reproduction. Either explicitly, as is the case with the Directive, or by reliance on the pre-existing *sui generis* database exception in the UK, a derogation from the restricted acts of extraction and re-utilisation is also provided. Furthermore, both Articles 3 and 4 of the CDSM Directive contain an exception to the new press publication right which reinforces the rights of press publishers to control their works online.[200] In the case of Article 4, the Directive additionally authorises an exception to "the translation, adaptation, arrangement and any other alteration" of a computer program.[201]

In both UK and EU law, of particular note, is the omission of an exception to the right of communication to the public. Thus, on first blush, those individuals and organisations wishing to work in collaboration or share outputs from machine learning are left unsupported.

1.1 Is the Use of Expression "Text and Data Mining" in Legislation Problematic?

Witten, Frank and Hall define text and data mining as "the extraction of implicit, previously unknown, and potentially useful information from data."[202] The common use of the term text and data mining in statute is however less rooted in the technicalities of computer engineering than the spotlight policy makers chose to give academic research.[203] Beginning in 2008, the question of whether EU copyright law permitted data analytics was raised consistently by both libraries[204] and research organisations.[205] As a result of legislatures honing in on the specific issues associated with automated text extraction in academia, it became not only the term of choice in European copyright law, but served to eclipse discussions on the economic importance and commercial application of AI more broadly. This no doubt explains why, when the Commission first proposed an exception for TDM in 2016, it did so solely for the benefit of not-for-profit public interest research organisations.[206] Turning first to the UK, the term "text and data analysis" in the section title of the CDPA raises the question as to whether a strict interpretation by the UK courts could limit enjoyment of the exception to the specific field of TDM alone. The drafting of s.29A arguably reinforces this by stating that "a person who has lawful access to the work may carry out a *computational analysis* of anything

[200] Article 15 (1).
[201] Article 4 (1)(b), Computer Programs Directive.
[202] Ian Witten, Eibe Frank, Mark Hall, *Data Mining – Practical Machine Learning Tools and Techniques* (Elsevier 2011) xxi.
[203] See 46-7 from *Digital Opportunity: A Review of Intellectual Property and Growth. An Independent Report by Professor Ian Hargreaves* (May 2011).
[204] It appears the first reference to text and data mining in Europe in a copyright context came from the British Library in 2008 arguing that Article 5 (3) of the Information Society Directive should be amended to allow society to benefit from new technologies that allow the digitisation of library collections and aid with knowledge discovery. See Response from the British Library to the Green Paper on Copyright in the Knowledge Economy 2008. See also reference to copyright and TDM in the 2009 British Library paper Copyright and Research at https://blogs.ch.cam.ac.uk/pmr/2009/05/09/british-library-document-on-copyright/
[205] Library, research and education sector members dominated the European Union's 2013 Licences for Europe stakeholder dialogue working group "Text and Data Mining for scientific research purposes." See https://libereurope.eu/wp- content/uploads/2013/05 /Letter_of_withdrawalL4E_TDM_May-24_4.pdf
[206] Article 3. Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market. COM/2016/0593 final - 2016/0280 (COD).

recorded in the work."[207] Reflective of this, the accompanying UK Intellectual Property Office guidance evinces that the purpose of the exception is "the use of automated analytical techniques to *analyse* text and data *for* patterns, trends and other useful information."[208] Thus, where the primary objective is not the analysis of previously unknown knowledge, trends and patterns, but, for example, the output of generative works, the UK courts may adopt a narrow teleological interpretation, thereby excluding broader categories of machine learning from its scope.

Perhaps conscious of the interpretational vagaries of the UK approach, the CDSM Directive sets forth a broad definition for text and data mining: "*any* automated analytical technique aimed at analysing text and data in digital form in order *to generate* information *which includes but is not limited to* patterns, trends and correlations."[209] Furthermore, as pointed out by the Hamburg Regional Court in the LAION ruling,[210] the all-encompassing nature of the Directive beyond just TDM is confirmed by the AI Act. The Regulation setting forth that those parties using protected subject matter for training, and the operation of general purpose models, must comply with the reservation of rights in Article 4 of the CDSMD.[211]

2. Scope of UK and EU TDM Exceptions

2.1 Beneficiaries

A particularly striking feature of the UK and EU TDM exceptions is the limited support afforded commercial AI actors. The UK, at the time of writing, has no explicit exception allowing commercial text and data mining as s.29A only permits "non-commercial research".[212] Somewhat less restrictively, Article 3 of the CDSM Directive may be enjoyed by research organisations which are "not-for-profit", or which pursue a "public interest mission recognised by a Member State", alongside publicly accessible cultural heritage institutions.[213] In contrast, Article 4 allows individuals and entities of both a commercial and non-commercial nature to undertake machine learning on the proviso that the rights have not been "expressly reserved by their rightsholders."[214]

Whilst the approaches taken by the UK and European exceptions to non-commercial research are no doubt similar, they nevertheless diverge in some notable ways. The UK exception designates any

[207] s.29A, CDPA.
[208] *Exceptions to copyright: Research* (Intellectual Property Office 2014).
[209] Article 2 (2) (Emphasis added).
[210] *Robert Kneschke v LAION*, Hamburg District Court, 310 O.227/23 (27/09/2024) (hereafter *Kneschke v LAION*).
[211] Inter alia Recital 60 of the AI Act (Regulation (EU) No. 1689/2024) confirms that providers of general purpose AI models must comply with Article 4 of the CDSM Directive.
[212] Although beyond the scope of this thesis, the temporary copying exception in UK and EU law may allow certain forms of machine learning as explicitly acknowledged in Recital 9 of the CDSM Directive 2019/790.
[213] Article 2 (1) (a) (b).
[214] Article 4 (3).

natural or legal person "*who has lawful access to the work*" as the beneficiary, and requires that any reproduction or extraction of a work governed by exclusive rights must be "*for the sole purpose of research for a non-commercial purpose.*"[215] This outright prohibition on commercial uses can be explained by the obligation that the UK was under at the time[216] to comply with the non-commercial confines of the research exceptions provided by Information Society Directive[217] and the Database Directive.[218]

At the European level, the seismic divide between commercial and non-commercial research that proliferates across the copyright acquis creates an ever-present constraint for research-intensive industries. Conscious of this, under Article 3, the Commission eschewed the purposive approach taken by the UK legislature in favour of one that focuses on the nature of the organisation undertaking TDM. As outlined in the Impact Assessment, wishing to reduce the legal uncertainty inherent in "research projects carried out by public organisations with a possible commercial outcome"[219,] the founding principle of the Article is one based on organisational type. As a consequence, a priori, it therefore permits certain levels of commercial exploitation.

As mentioned above, Article 3 of the CDSM Directive confers special privileges on two distinct beneficiaries: cultural heritage institutions and research organisations. As a new category of organisation in the copyright acquis, the contours of a research organisation's activities are described in comparative detail. The Directive sets out that they are required to conduct educational activities and / or scientific research on either a not-for-profit basis or on the condition that all profits are reinvested in research.[220] Where this is not the case, their activities must be "pursuant to a public interest mission recognised by a Member State."[221] The recitals elaborate that this could be by virtue of national laws, public funding, or public contracts.[222]

Having once defined the nature of the beneficiaries, rather than leaving them free to commercialise AI, the Directive then creates what Geiger et al have described as a definitional "quagmire".[223] Despite the Commission's impact assessment articulating a desire to support projects "with a possible

[215] s.29A (1) (a) (Emphasis added).
[216] See Recital 32 of the Information Society Directive which provides for "an *exhaustive* enumeration of exceptions and limitations". (Emphasis added). Prohibitions on Member States introducing further exceptions in addition to those set out in Art 5 of the Information Society Directive has been further confirmed a number of times by the Court of Justice of the European Union. For example, Case C-301/15 *Marc Soulier and Sara Doke v Premier ministre and Ministre de la Culture et de la Communication* [2016] ECLI:EU:C:2016:878, 26 (hereinafter *Soulier and Doke*). Case C469/17 *Funke Medien NRW GmbH v Bundesrepublik Deutschland I* [2019] ECLI:EU:C: 2019:623, 56. Case C-161/17 *Land Nordrhein-Westfalen v Dirk Renckhoff* [2018] ECLI:EU:C:2018:634,16.
[217] Article 5 (3) (a).
[218] Article 6 (2) (b). See also *Soulier and Doke*, 26.
[219] 4.4.3. Impact assessment for the proposal for a directive of the European Parliament and of the Council on copyright in the digital single market (SWD 2016) 301 Final (Part 1/3).
[220] Article 2 (1) (a).
[221] Article 2 (1) (b).
[222] Recital 12.
[223] Christophe Geiger et al, 'Text and data mining: Articles 3 and 4 of the Directive 2019/790' (2019) Propiedad intelectual y mercado único digital europeo, 27.

commercial outcome", the Directive nevertheless stymies collaborations with Article 2.1 proclaiming that the results of the research "cannot be enjoyed on a preferential basis by a… [commercial] undertaking that exercises a decisive influence upon" a research organisation or CHI.[224] Recital 12 further clarifies that a decisive influence is one that arises due to "structural situations" which include but are not limited to the "quality of shareholder or member". Thus, it appears evident that the European legislature intended to prevent commercial entities from circumventing Article 4 by entering into contractual arrangements with the beneficiaries of Article 3. As a consequence, the restrictions on permissible economic activities are so extensive that they call into question whether the Directive leaves any meaningful scope for commercial outcomes.[225]

In sharp contrast to the exhaustive focus on research organisations taken by the Directive, the actions of cultural heritage institutions remain unelaborated upon. Consequently, whether an oversight or not, it may be logically inferred that the intention of the legislature is that the purely non-commercial stipulations laid out in InfoSoc regarding CHIs are to apply to the ML activities of this category of institution.

A further issue caused by the organisational approach to beneficiaries adopted by the Directive under Article 3, is its exclusion of public interest actors such as unaffiliated citizen scientists, journalists, and independent researchers. Moreover, due to the "primary goal" requirement for research or education under Article 2.1, important public interest bodies such as Think Tanks, media organisations, and certain categories of medical research organisation are left excluded.

By contrast, Article 4 does not impose any restrictions based on user type. Any individual or organisation can make use of the exception for commercial as well as non-commercial purposes. Nevertheless, as discussed in more detail below, it creates a major innovational drawback as it permits a reservation of rights, rendering its application ultimately dependent on the wishes and prerogatives of copyright owners.

2.2 Categories of Works

While the Information Society Directive (2001/29/EC) is notable for its enumerated list of broad and expansive exceptions, Articles 3 and 4 of the CDSM Directive and s.29A of the CDPA are tightly

[224] Article 2 (1) *in fine.*
[225] A priori one case where commercialisation may exist is in the case of an entirely independent charity which offers commercial products and services pursuant to Art 2 (1) (a). c.f. Martin Senftleben et al, 'Towards a European Research Freedom Act: A Reform Agenda for Research Exceptions in the EU Copyright Acquis' (2025) 18-19 http://dx.doi.org/10.2139/ssrn.5130069 (asserting that structural situation refers to the entire institution and not a particular research project.)

delineated. Be that as it may, nevertheless, many important questions linger as to the exact scope of the exceptions and precisely which activities are licit absent the authorisation of rightsholders.

2.2.1 Are Computer Programs Excluded from Article 3?

Examining the Directive, it can be observed that whereas machine learning on protected works governed by the copyright acquis and *sui generis* database right is generally permitted, an explicit derogation for the reproduction, extraction, and re-utilisation of computer programs is only explicitly provided for under Article 4. No explanation for this disparity can be found in either the impact assessment, or the CDSM Directive itself.

Ducato and Strowel posit that one reason for this anomaly is that Article 5(3) of the Software Directive (2009/24/EC) already allows for research, as it furnishes end-users with the right to "observe, study or test" a computer program.[226] However, this so-called "black box" analysis does not allow for the reproduction or extraction of the software and therefore, as intimated by the authors themselves, hardly provides a satisfactory answer.

Kretschmer and Margoni assert that Article 3(1)'s derogation from the Information Society Directive's reproduction right should be sufficient to allow copying of computer programs by scientific researchers. Applying hermeneutic principles, the authors argue that "when special law does not derogate then general law applies."[227] In other words, where the Software Directive does not derogate from InfoSoc, the reproduction right as outlined in the 2001 Directive is operative. This assertion seems however doubtful on a number of fronts.

In line with the rule of speciality, Recital 20 of the Information Society Directive explicitly states it operates "without prejudice to the provisions" of the Computer Programs Directive. Furthermore, irrespective of CJEU assertions that uniform interpretations should be given to terminologies used in Union law, it cannot be ignored that the act of reproduction under the Software Directive is not co-extensive with the right as set out in the Information Society Directive.[228] Thus, it can be concluded that general law does not apply as the Computer Programs Directive clearly derogates from InfoSoc.

Most strikingly, the treatment of adaptations differs between the two Directives. Whilst formally the adaptation right is not horizontally harmonised across the EU, in reality the broad interpretation given

[226] Rossana Ducato and Alain Strowel, 'Ensuring text and data mining: remaining issues with the EU copyright exceptions and possible ways out.' (2021) 435 EIPR 322, 328.
[227] Thomas Margoni, Martin Kretschmer, 'A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology' (2022) 71 GRUR International, 8 685, 686 n.8.
[228] Thomas Dreier, Bernt Hugenholtz eds, *Concise European Copyright Law* (Kluwer Law International 2006) ("the definition of 'reproduction' in the Computer Programs Directive does not go as far as in Art 2 of the Information Society Directive.") 221.

to InfoSoc's reproduction right denotes that numerous forms of alteration to a copyright work may, subject to Member State law, constitute an act of reproduction. The Information Society Directive itself proclaims that reproduction can be "by any means *and in any form*."[229] Underlining this, the enumerated list of InfoSoc's exceptions implicitly conjure up permutations of both facsimile copying as well as those that require adaptations and alterations. Furthermore, the CJEU has hardly lifted an eyebrow in categorising materially differing types of adaptations and alterations to copyright works as reproductions under Article 2 of the 2001 Directive.[230]

The extensive conceptualisation of the reproduction right under InfoSoc stands in contrast to the Computer Programs Directive. As *lex specialis* it explicitly derogates from InfoSoc by dividing reproduction[231] from the exclusive rights of "translation, adaptation, arrangement and *any other alteration*".[232] Moreover, a careful reading of the Explanatory Memorandum to the Software Directive makes clear that reproduction does not encompass acts that require some form of transformation of form or purpose.[233] Thus, machine learning activities such as normalisation, format shifting, dimensionality reduction, feature construction etc are not covered by the act of reproduction under Union law. Instead, they constitute an adaption and alteration pursuant to Article 4 (b) of the Computer Programs Directive.[234]

In the light of this discussion, why s.29A – which was enacted five years before the CDSM Directive – permits the mining of computer programs when the Computer Programs Directive 2009/24/EC sets out an exhaustive list of exceptions which TDM falls beyond, is therefore highly pertinent. Close examination of the Software Directive however reveals the UK government was at liberty to create an exception for the data analysis of software. This is supported by Recital 19 of the 2009 Directive which allows for derogations in national law "on points not covered in this Directive." Unsurprisingly, given that machine learning was predominantly confined to universities' computer engineering departments at this point, an exploration of the Directive's text informs that ML was not envisaged by the EU legislator at the time of its promulgation.[235] As a result, it can be contended that the UK government had the necessary leeway to introduce a derogation permitting the reproduction and adaptation of software for the purposes of data mining.

[229] Article 2, InfoSoc. (Emphasis added).
[230] For example, in Case C-419/13 *Art & Allposters International BV v Stichting Pictoright* [2014] ECLI:EU:C: 2014:2214 the Court held that the physical transferral of a paper poster onto canvas was a reproduction. 43. In *Infopaq 1* at each stage the digitisation, creation of TIFF files, OCRing and printing out of 11 words were each ruled to be subject to the exclusive act of reproduction.
[231] Article 4 (a).
[232] Article 4 (b) (Emphasis added).
[233] Compare paras 4.1 (a) and (b). Proposal for a Council Directive on the legal protection of computer programs. COM(88) 816 final – SYN 183. (89/C 91/05).
[234] Although the Computer Programs Directive differentiates between a technical reproduction and an adaptation, how the two acts are delineated in national law will depend on the jurisdiction.
[235] Supra 233. It is interesting to note that it is believed the first commercial deployment of AI algorithms took place in 1989. 'Business Technology. What's the best answer? It's survival of the fittest.' *New York Times* (29/8/1990).

As a consequence, the apparent exclusion of computer programs from Article 3 is perhaps less problematic than it first appears. Whereas long-term storage[236] and human study[237] are explicitly covered by the 2009 Directive, and therefore derogations for the reproduction of computer programs were arguably required for Articles 5[238] and 6[239] of the CDSM, the same cannot be said of text and data mining. Data analysis and machine learning, while guided by a data scientist, is not study undertaken by a human, as implied by Article 5.3 of the 1989 proposal.[240] Neither is it study in the traditional sense as it is an automated process, whose prime purpose is not the acquisition of knowledge per se, but rather the weighting of and / or prompting of calculations by the model in order to produce a tangible output.

In conclusion, therefore, despite the lack of an explicit derogation in Article 3, by dint of Recital 19 of the Software Directive, it is proposed that Member States have the discretion in their transposition to permit the automated mining of object and source code for the purposes of scientific research. This would throw light on why an explicit exception for reproduction of computer programs was deemed superfluous by the Commission at the time of the original 2016 proposal. Its presence only in Article 4, which was proposed by Council late in the day, may be explained by the hurly-burly of a trilogue very much preoccupied with Articles 15[241] and 17,[242] and as a consequence, only lightly focused on the legal technicalities of TDM.[243] Under pressure from the Romanian Presidency to conclude a highly controversial directive, it is posited that there was a failure to realise the mismatch between the two articles and the technical reason underpinning the absence of a derogation from the Software Directive under Article 3.

2.3 Storage and Security

Unlike the UK TDM exception, which is silent on storage and security measures, the Directive creates different obligations dependent on the category of beneficiary. No security measures are required of beneficiaries of Article 4 (nor it can be observed, cultural heritage organisations storing reproductions in line with Articles 6 and Articles 8 - 11 of the Directive). Article 3, however, establishes that TDM-related reproductions "shall be stored with an appropriate level of security".[244] Additionally, CHIs,

[236] Article 5 (2) (back-up copies).
[237] Article 5 (3) In addition to AI not being referenced in the original Proposal for the Software Directive, as made clear in Article 5 (3) study was envisaged to be that undertaken by a human. "It is important, given the need to encourage computer literacy in all sections of the Community, that libraries are able to offer computer programs for *study by the public* in the same way as they offer other literary works." (Emphasis added).
[238] Article 5. Use of works and other subject matter in digital and cross-border teaching activities.
[239] Article 6. Preservation of cultural heritage.
[240] Supra 237.
[241] Article 15. Protection of press publications concerning online uses.
[242] Article 17. Use of protected content by online content-sharing service providers.
[243] Jon Fingas, 'EU copyright laws face uncertain fate after 11 countries reject proposal' https://www.engadget.com/2019-01-19-eu-copyright-directive-rejected-at-council.html
[244] Article 3.2.

research organisations, and rightsholders are encouraged to work together to "define commonly agreed best practices."[245] However, Recital 15 stipulates that any measures should remain "proportionate and limited to what is needed for retaining the copies in a safe manner". Despite this proportionality requirement, it nevertheless implies further onerous obligations on beneficiaries, not least the appointment of "trusted bodies for the purpose of storing … copies".[246]

Although the secure storage of data can perhaps be assumed given the public nature of the organisations who benefit from Article 3, and therefore the provisions of the Directive do not necessarily work as a hindrance, where Member States goldplate these obligations they can create undue burdens. Irish copyright law, for example, is particularly onerous in this respect, requiring organisations to not only inform rightsholders as to the reproductions they are holding, but also engage in dialogue where "additional security measures" are requested.[247]

2.4 Communication to the Public and Reutilisation

From an innovational perspective, one problematic feature of both the UK TDM provision and the Directive is the absence of an exception providing for the communication to the public and reutilisation of protected subject matter. In Union Law, the sharing of a qualitatively significant[248] part of a copyright work may implicate the act of communication to the public – an exclusive right of the author to control the transmission of their work including online, as well as via satellite, cable, and other forms of wireless and non-wireless communication.[249] The equivalent to the communication to the public right in the case of the *sui generis* database right is the restricted act of "re-utilisation".[250]

The exact contours of the communication to the public right have been the subject of much scholarly commentary as well as judicial interpretation at the Court of Justice of the European Union (CJEU).[251] One particular strand of Court decisions has revolved around the core concept of what constitutes a "public" – something which despite extensive case law still remains unharmonised and lacking a transparent and easy to apply *de minimis* public / private threshold.[252] To this end, ALAI observes that "the definition of 'public' tends to shrink or expand depending on the nature of the right at stake."[253] Of particular note, the European Commission has expressed the view that the term can potentially

[245] Article 3.4.
[246] Recital 15.
[247] 53A (3C), Copyright and Related Rights Act 2000.
[248] Caterina Sganga, 'The Right of Reproduction' in E Rosati (ed), *Routledge Handbook of EU Copyright Law* (Routledge 2021) 139 (stressing the test adopted by the CJEU for partial reproduction of a work is a qualitative not quantitative one).
[249] Article 3, InfoSoc.
[250] Article 7 (2) (b), Database Directive.
[251] For example, Branka Marušić, *The Autonomous Legal Concept of Communication to the Public* (Elgar 2023), Justin Koo, *The Right of Communication to the Public in EU Copyright Law* (Hart 2019).
[252] *Report and Opinion on the Making available and communication to the public in the internet environment: focus on linking techniques on the internet* (Association Littéraire et Artistique Internationale 2014) 5.
[253] Ibid.

cover a very small number of people "consisting only of a few persons" beyond "the immediate family circle and the closest social acquaintances".[254]

Examination of relevant CJEU case law,[255] none of which has addressed research issues specifically, reveals a nuanced multi-factor approach to delineating between the public and private sphere that depends on the interplay of three criteria:

i) the number of people (case law suggesting that it requires a "fairly large number",[256] though this may take into account not just how many people have access at one time but by the "cumulative effect" of making the work available to successive individuals[257]);
ii) the character of the group (whether private, family, "an indeterminate number"[258] of potential recipients or a "new public" which was "not taken into account by the authors of the protected works within the framework of an authorisation given to another person"[259]);
iii) the nature of the communication (whether or not for financial profit[260,261,262] or whether the communicating party derives direct or indirect benefit.[263])

It goes without saying that whether the communication to the public right, opaque in its conceptualisation, is engaged when sharing protected training data or AI outputs, will depend on the specificities of the matters before the court as well as judicial interpretation in line with national law. For instance, based on the criteria set out above, a defined group of collaborators in the context of a not-for-profit project sharing information may not constitute communication to the public. Conversely, communication to a larger group of individuals for commercial research purposes may feasibly reach the required threshold. Be that as it may, the lack of guidance provided by Article 3 of InfoSoc combined with the absence of research relevant CJEU case law means that it will frequently not be

[254] See Commission staff working paper on the review of the EC legal framework in the field of copyright and related rights 11634/04 (23/7/2004).
[255] For example, Case C-135/10 *Società Consortile Fonografici (SCF) v Marco Del Corso* [2011] ECLI:EU:C: 2011:431 (hereinafter *Del Corso*); Case C-306/05 *Sociedad General de Auditores y Editores de España (SGAE) v Rafael Hoteles SL* [2006] ECLI:EU:C:2006:764 (hereinafter *Rafael Hoteles*).
[256] *Del Corso* 84, *Rafael Hoteles* 38.
[257] See *Rafael Hoteles* 39, Case C-607/11 *ITV Broadcasting Ltd and Others v TVCatchUp Ltd* [2013] ECLI:EU:C:2013:147 33, Case C-351/12 28 *Ochranný svaz autorský pro práva k dílům hudebním o.s. v Léčebné lázně Mariánské Lázně a.s.* [2014] ECLI:EU:C:2014:110 28, Case C-117/15 *Reha Training Gesellschaft für Sport- und Unfallrehabilitation mbH v Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte eV (GEMA)* [2016] ECLI:EU:C: 2016:379 44 (hereinafter *Reha v GEMA*).
[258] Case C-89/04 *Mediakabel BV v Commissariaat voor de Media* [2005] ECR I-4909 30, Case C-192/04 *Lagardère Active Broadcast, v Société pour la perception de la rémunération équitable (SPRE) et al* [2005] ECLI:EU:C:2005:475 31.
[259] Joined cases C-431/09 and C-432/09 *Airfield NV and Canal Digitaal BV v Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (Sabam) and Airfield NV v Agicoa Belgium BVBA* [2011] ECR I-09363 72.
[260] Joined Cases C-403/08 and C-429/08 *Football Association Premier League Ltd and Others v QC Leisure and Others and Karen Murphy v Media Protection Services Ltd* [2011] ECR I-9083 204 (hereinafter *FAPL*).
[261] See *Reha* v *GEMA* 49 "[A]lthough it is true that the profit-making nature of the broadcast of a protective work does not determine conclusively whether a transmission is to be categorised as a 'communication to the public' … it is not however irrelevant."
[262] *Joined Cases* C-775/21 *and* C-826/21 *Blue Air Aviation SA v UCMR – ADA Asociaţia pentru Drepturi de Autor a Compozitorilor and Uniunea Producătorilor de Fonograme din România (UPFR) v Societatea Naţională de Transport Feroviar de Călători (SNTFC) 'CFR Călători' SA* [2023] ECLI:EU:C:2023:307 71.
[263] *Del Corso* 97-99, *Rafael Hoteles* 44, *Reha v GEMA* 63.

possible to discern whether an organisation's dissemination of data to research partners engages the communication to the public right or not. Whilst inclusion in the Directive of an explicit derogation for communication to the public would therefore certainly not create a bright line for organisations wishing to share data, its absence contributes to legal fragmentation and uncertainty across the European Research Area for those individuals and organisations wishing to engage in both domestic and cross-border TDM collaborations.[264]

2.4.1 What Copyrightable Outputs Arise from the Processing of Protected Subject Matter?

To precisely determine the contours of an exception for communication to the public and its potential impact on data scientists, it is first necessary to identify those works involved in the AI lifecycle which constitute protected subject matter. Following this, this subsection shall examine a handful of Member States who have sought to address the omissions of the Directive before examining alternative exceptions that may in theory be invoked to communicate training data and / or the outputs from an AI.

Taking the broad notion of reproduction of protected subject matter operative under InfoSoc,[265] engagement of the restricted act can be identified in principle at four different stages:

i) reproductions created during the data preparation phase;
ii) a curated subset of i) which is ingested into the model for training, validation, and testing;
iii) (in the case of eager learning models[266]) the parts of the machine learning code which have been created by ii (i.e. potentially adaptations); and
iv) outputs from the model

As outlined in Chapter 1, the reproduction right is invoked as data is massaged and curated according to the question that the AI model is designed to answer. Once the data is ready, it is then ingested into the model.[267] As close in form to the original, where they constitute protected subject matter categories, (i) and (ii) will therefore unquestionably be regulated and governed by copyright or related rights.

[264] For example while the Italian (Article 70-ter (1)), German (§60d (4)(1)) Slovenian (Article 57b (5)) and Irish (s. 53A(5)) transpositions of Article 3 explicitly permit forms of communication to the public of protected matter, those of Spain (Article 67), Portugal (Article 75(2)(v), 76(4), (5) (6)) and Sweden (§15b) do not.
[265] See subsection supra 2.2.1.
[266] Eager learning models are machine learning models such as neural network models, SVMs, decision trees, logistic regression etc whose structure is in part based on the data it has used in the training phase.
[267] According to the CrowdFlower Report on average data preparation and handling takes 82% of a data scientist's time.

Category (iii) raises the distinctly knotty problem as to whether AI software trained and generalised on protected subject matter constitutes a reproduction or adaptation of the original work. This in turn raises the issue of whether a trained model can be shared with others without the authorisation of the rightsholders whose works it was trained on. Ultimately, the question of whether copyright works that shape the form of the AI code itself creates a derivative work will have to be clarified by the courts. However, it is posited that such a position is unlikely to succeed.[268] In a generalised model, as a rule, the training data remains only in the sense that it influences the parameters as to *how* a calculation takes place. The works protected by exclusive rights do not form in any sense part of the trained software model – all that remain are relationships and representations from the analysed data that as numerical values contribute to a particular model calculation.[269] These calculations do not equate to a corresponding datapoint in the original, and importantly from the perspective of copyright, neither can the parameter values be reverse-engineered to recreate a facsimile version of the original training data.[270]

The final category (iv) – outputs from an AI – has engendered much scholarly and public debate, particularly concerning cultural goods.[271] Moreover, cultural outputs have been the object of a growing number of lawsuits in both the US[272] and the UK.[273] Putting aside the thorny issue of determining the ownership of any copyright in the outputs from a model, assuming any such copyright exists, it goes without saying that whether an infringement has occurred will depend on the specific details of the case, which are likely to be highly technical and complex.

[268] For example, in dismissing many of the plaintiffs' claims in *Kadrey et al v Meta Platforms Inc* Case No. 23-cv-03417-VC (16/11/2023) the judge noting a finding of substantial similarity between the model and the works in the training corpus would be "implausible", 8. In *Getty Images and Others v Stability AI* [2025] EWHC 38 (Ch) 600-601, Smith J similarly ruled that weights were not an infringing copy because they do not contain a copy of the original work. Cf. *GEMA v OpenAI* (Case No. 42 O 14139/24) (where Regional Court of Munich I ruled that because pursuant to a prompt the mathematical formulae (parameters) had the potential to produce lyrics substantially similar to those used in the training phase the model software infringed the reproduction right.)

[269] cf. See discussion in Chapter 5 where auto-encoders (a type of neural network) are discussed. They reduce the data dimensionality to store in latent space the most relevant relationships extracted from the training data as code which facilitates the reconstruction of an approximation of the original data by a decoder. See G Hinton and R Salakhutdinov 'Reducing the Dimensionality of Data with Neural Networks' (28/7/2006) Science 313 5786 504.

[270] It is noteworthy that in model extraction attracts, with some prior knowledge about the model while it is possible to understand more about the data classes used or whether certain data was used or not in the training, it is not possible to recreate the training data precisely itself. See Giuseppe Ateniese et al, 'Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers.' (2015) International Journal of Security and Networks 10 3 137. Reza Shokri et al, 'Membership inference attacks against machine learning models.' (2017) IEEE symposium on security and privacy, IEEE 3.

[271] A few recent examples include Pamela Samuelson 'Generative AI meets copyright' (2023) Science 3816 654 158. Matthew Sag 'Copyright Safety for Generative AI' (2023) Houston Law Review 61 295. Andres Guadamuz, 'Do androids dream of electric copyright? Comparative analysis of originality in artificial intelligence generated works.' (5 / 6/ 2017) Intellectual property quarterly 169. Alain Strowel 'ChaptGPT and Generative AI Tools: Theft of Intellectual Labor?' (2023) IIC 54 4 491.

[272] See *Getty Images (US) Inc v Stability AI Inc* (1:23-cv-00135) District Court, Delaware (Feb 2023). *Andersen et al v Stability AI et al* (3:23-cv-00201) ND California (Jan 2023). *J Doe 1 et al v GitHub, Microsoft Corp, OpenAI et al* (4:2022cv06823) ND California. (Nov 2022) *Stephen Thaler v Shira Perlmutter Register of Copyrights and Director of the United States Copyright Office, et al*. 22-1564 (BAH) 18/8/2023. *Kadrey et al v Meta Platforms Inc*, No 23-cv-03417-VC (N D Cal June 25/6/2025) (hereafter *Kadrey v Meta* or *Meta*). *Bartz v Anthropic PBC*, No 3:24-cv-05417-WHA, 2025 WL (N D Cal 23/6/2025) (hereafter *Bartz v Anthropic* or *Anthropic*).

[273] *Getty images (US) Inc, Getty Images International UC, Getty Images (UK) Ltd, Getty Images DEVCO UK Ltd, iStockphoto LP, Thomas M. Barwick Inc, v Stability AI* IL-2023-000007 (15/1/2023).

A priori having discounted the model code as falling within the scope of a third party's exclusive right of communication to the public, the analysis now turns to the remaining categories of computer-generated outputs in greater detail. Broadly speaking, three types of results can be identified:

1. Firstly, basic informational results that do not contain or reach the threshold of protectable subject matter will raise no question of infringement. For example, expression such as a percentage of likelihood, a short identifier, chemical formula, numerals etc.[274] Similarly, data relating to a real-world phenomenon by an IoT device are not protectable.[275] Furthermore, information from the public domain, or derived from openly licensed content, is in many instances unlikely to infringe third-party rights.

2. A second output category are entire works such as an original copyrightable photograph or a picture replicated "as is" from the training data and shared in facsimile or close to facsimile form. Sharing training data with project partners or for the purposes of peer review, as well as communicating data with the competent authorities in line with the legislative obligations provided *inter alia* by the Digital Services Act[276] and AI Act[277], would also fall into this grouping.

3. The third type of AI output is, in the natural language sense of the word, an original model output derived from the training process. This category can be further divided into four sub-categories: i) those which resemble a copyright work as the model either overfits to a work (sometimes called "memorisation"), or the model is predominantly or exclusively trained on the work of that author and it competes in the same marketplace;[278] ii) those identical to the previous category except that the output does not compete in the same marketplace as the original; iii) outputs which resemble a pre-existing protected work but the model was not trained on works of the author, or the author's work is such a small contribution as not to have influenced the model; and iv) those which are entirely original with no characteristics or features discernibly derived from any pre-existing work.

Putting aside for the time being copyright questions relating to the training process to focus solely on the making available of the outputs, in the absence of authorisation or an exception, communication to the public of Category 2 and Category 3(i) are almost certainly an open and shut case of infringement.

---

[274] In *Authors Guild v HathiTrust* 755 F 3d 87 (2d Cir 2014) the visible outputs were keywords and frequency of their occurrence on the pages of a selected title.
[275] Article 43, Data Act.
[276] Article 72 (1).
[277] Annex VII, Article 4.3.
[278] See 'Opinion: Use of Copyright-Protected Content for Machine Learning' (18/12/2022) Israeli Ministry of Justice stating that "the safe harbor set forth in this Opinion would not apply to ML datasets that consists exclusively of works created by a single author in order to compete with this author in her existing markets." 3.

Category 3 iv) outputs, subject to any new jurisprudential developments, although trained on the works of others, may as Sag implies be non-infringing as they are "non-expressive works" lacking the expression of a third party author.[279] Categories 3 ii) and iii) do not lend themselves easily to general conclusions, and as a consequence whether the making available of these categories of output is an infringing act will no doubt turn on the matters at hand, the boundaries of communication to the public under national law, and any legal developments that may arise in the future.

2.4.2 Which Ancillary Permitted Acts May Allow Communication to the Public?

As mentioned above, despite the very real need to share training data and ML results for the purposes of collaboration, peer-review, or transparency,[280] neither UK law nor the CDSM Directive create an explicit TDM exception for communication to the public. Data scientists wishing to share copyright works are therefore required to rely on permitted acts provided elsewhere in the *acquis* and national legislation.

Importantly, a handful of the Information Society's exhaustive list of exceptions may facilitate communication to the public of protected subject matter for commercial and non-commercial purposes. These include the reporting of current events,[281] the use of AI by commercial entities involved in public security activities,[282] the use of public speeches and "similar works …to the extent justified by the informatory purpose",[283] as well as for the purposes of caricature, parody, and pastiche.[284, 285] Arguably, as discussed below in more detail, the quotation exception,[286] may offer most relief to commercial bodies wishing to share protected textual subject matter with other businesses as well as the public.

The picture is less bleak for EU researchers. Recital 15 of the CDSM Directive encourages Member States to utilise exception 5.3(a) of InfoSoc for non-commercial research purposes "such as scientific peer review and joint research" that involve dissemination of information. However, confusingly it simultaneously states that its application should apply in cases "other than text and data mining." In doing so, while difficult to interpret, the Directive appears to be encouraging the communication of

[279] See Matthew Sag, 'Copyright safety for generative AI' (2023) 61 Hous L Rev 295, 301 (arguing that AI outputs are only likely not to be fair use when the model overfits ("memorises") the training data.)
[280] See Chapter 1 and the obligation to share training data with competent authorities upon a reasoned request under the DSA, DMA and AI Act.
[281] Article 5 (3) (c).
[282] Article 5 (3) (e).
[283] Article 5 (3) (f).
[284] Article 5 (3) (k).
[285] Andres Guadamuz, 'A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs' (2024) 73 GRUR International 2 111, 127 (suggesting that AI outputs could be viewed as a pastiche as "one could convincingly argue that generative outputs are a form of evaluative reference of an original work that is also a modified version of the original.") See "§5 Pastiche" Stability AI defence submitted to the High Court on 27/2/24. Claim No IL-2023-000007.
[286] Article 5 (3) (d).

works not for TDM purposes directly – such as sharing training data or computer-generated works containing copyright – but for *ex post* activities that support or result from the ML process.

Article 5.3(a) of InfoSoc permits the communication to the public of reproductions "for the sole purpose of illustration for teaching or scientific research… to the extent justified by the non-commercial purpose to be achieved". Thus, despite the absence of an explicit exception permitting communication to the public, the EU 27 have the leeway in national law[287] to broaden the scientific TDM research exception as deemed necessary, including, as suggested by Recital 15, in regard to complimentary activities that support machine learning.[288]

Although not all countries have utilised the flexibilities provided by InfoSoc, as one example §60d (4) of the German Copyright Act permits the making available of reproductions to "a specifically delimited group of persons for the purpose of their joint scientific research" as well as to "individual third parties for the purpose of reviewing the quality of scientific research". Any communication to the public must however "be terminated as soon as the joint scientific research or the monitoring of the quality of the scientific research has been concluded."

By contrast, researchers in countries like the UK, Spain, and Sweden – countries which neither have a TDM exception permitting communication to the public nor a general research exception – may theoretically rely on the quotation exception as set forth in national law. Despite the explicit prohibition of onward dissemination of protected subject matter under s.29A of the Copyright, Designs and Patent Act, as the UK government itself pointed out in its 2014 implementational guidance, those wishing to communicate TDM inputs and outputs with others may be able to rely on the quotation exception to share AI outputs. In doing so, the guidance highlights that quoting "parts of a work" in the context of publishing scientific articles relating to TDM is a non-infringing act.[289]

Nevertheless, for EU based innovators, while there are theoretically no purpose-related barriers to quoting protected matter in the context of machine learning,[290] pursuant to *Pelham GmbH v Hütter* users must ensure that they "have the intention of entering into 'dialogue' with that work", and furthermore that the original work quoted is identifiable. [291] Certainly, in validating the accuracy of model predictions or sharing the model outputs with reference to the inputs, dialogue with the original work in the training set can arguably be seen to take place as envisioned by the CJEU. A criterion

[287] Article 25, CDSM Directive.
[288] Recital 15 *in fine*: "Uses for the purpose of scientific research, *other than text and data mining,* such as scientific peer review and joint research, should remain covered, where applicable, by the exception or limitation provided for in Article 5(3) (a) of Directive 2001/29/EC." (Emphasis added).
[289] *Exceptions to Copyright: Research* (Intellectual Property Office October 2014) 9.
[290] Stavroula Karapapa 'The Quotation Exception Under EU Copyright Law' 251 ff. in op. cit. *Routledge Handbook of EU Copyright Law*.
[291] Case C-476/17 *Pelham GmbH and Others v Ralf Hütter and Florian Schneider-Esleben* [2019] ECLI:EU:C:2019:624 70 (hereinafter *Pelham*) (see *in fine* penultimate paragraph).

which may prove significantly more problematic however is that the original work remains identifiable given the volume of works often involved upstream in AI training.[292]

Whereas flexibilities such as the quotation exception may provide some relief, it however misses the bigger innovational requirement – namely the sharing of the entire corpus of training data with commercial partners and fellow researchers. The conditions under which 5(3)(a) or the quotation exception in InfoSoc would provide grounds to share an entire work,[293,294] yet alone an entire corpus of works with partners, however, remain as a matter of law highly indeterminate.[295] The practice of lawfully quoting an entire work centres on the issue of intelligibility in relation to the quotation's intended purpose. Examples of potentially lawful uses of whole textual works include a slogan, title, or haiku.[296] Quoting only part of such works will render them incomprehensible, as is the case if only part of a drawing, painting, or photograph is quoted. Moreover, as stated above, pursuant to *Pelham* the CJEU has defined the need for dialogue surrounding the work which "illustrat[es] an assertion", "defend[s] an opinion", or facilitates an "intellectual comparison between that work and the assertions of that user".[297]

While disseminating training data for peer review purposes of AI outputs could theoretically fall under such categorisations, broader practices such as sharing an entire corpus with project partners in order to reduce bias and the likelihood of poor model predictions, or using a whole work in order for a model to communicate its prediction, evidently falls outside the quotation uses hitherto conceived by the European Court.

In conclusion, as will inevitably happen, when courts address communication to the public in the context of AI, a reliance on fundamental and constitutional rights such as freedom of expression, information, and science may provide a far stronger juridical rationale and underpinning than exceptions endogenous to copyright law. The advancement of science, and the societal importance of accurate and safe AI, are public interest objectives that sit beyond and above authorial entitlements and therefore would appear better addressed, not within a framework of statutory exceptions to the exclusive rights of authors, but as fundamental rights.

[292] Id. 71. In asserting that those quoting from a work must enter into dialogue with it, the CJEU observed the "essential characteristics of a quotation are the use, by a user other than the copyright holder, of a work or, more generally, of an extract from a work for the purposes of illustrating an assertion, of defending an opinion or of allowing an intellectual comparison between that work and the assertions of that user".
[293] In *Painer* (2012) the Advocate General found that Art 5(3)(d) can cover quotations of an entire work – in this case an entire photograph. In the final judgement this position was not questioned by the Court, and the ruling was evidently based upon this assessment.
[294] In *Pelham,* the court noted "the essential characteristics of a quotation are the use … of *a work* or, more generally, of an extract from a work." 71 (Emphasis added).
[295] For an in-depth exploration of the quotation exception, see 'Is it possible to share an entire work?' Lionel Bently and Tanya Aplin, *Global Mandatory Fair Use: The Nature and Scope of the Right to Quote Copyright Works* (CUP 2020) 104-8.
[296] Bently and Aplin, Id. 107.
[297] *Pelham,* 71.

2.4.3 The Exclusive Act of Reutilisation and TDM

While the UK and EU legislature failed to provide a derogation to the right of communication to the public, both Articles 3 and 4 of the CDSM Directive mandate Members States to implement an exception to the restricted act of re-utilisation.[298] Accordingly, in their "making available to the public" " distribution" "on-line… or [by] other forms of transmission" beneficiaries of the exceptions are permitted to communicate a substantial part, or the whole of a work, protected by the *sui generis* database right.[299] By contrast, the UK legislature declined to create an exception for the reutilisation of the *sui generis* right, only permitting under s.20 (1) (b) of the Copyright and Rights in Databases Regulations 1997 the extraction of the work for non-commercial research purposes.

As with Article 5.3(a), the implementation of the derogation to the exclusive right of reutilisation into national law is uneven and fragmented. For example, Sweden[300] and Germany[301] require that the sharing of a substantive part of a *sui generis* database is undertaken with the authorisation of the rightsholder. In contrast, Italy, Ireland, Spain, and Slovenia set forth markedly different provisions. Ireland[302] only permits the re-utilisation if the organisation is undertaking "research", thus potentially not extending to all uses permitted by Article 4. While in the case of the Italian[303] and Slovenian law,[304] the derogation is extended only to research organisations or cultural heritage organisations. By contrast, the Spanish transposition of Articles 3 and 4 permits the making available of a substantive part of a *sui generis* database for both commercial and non-commercial machine learning.[305]

2.4.4 Cross-Border Effect

Regardless of the political ambition of the European Union to create a single borderless European Research Area (ERA), the translation of these lofty ideals to support AI falls short, as the Directive fails to confer a cross border effect.[306] As with the operation of both public-private partnerships and public-public collaborations that require data sharing, the legal basis for cross-border communications of AI corpora and protectable outputs remains elusive.

298 Article 7(1), Database Directive.
299 Id. Article 7.2 (b).
300 §49 of the Swedish Copyright Act creates an exception for TDM, but as §15 (a) and (b) only create an exception to the reproduction right, it can be concluded re-utilisation is not a permitted act.
301 For example, §87c (1) (5) of the German Copyright Act only creates an exception to the reproduction right for *sui generis* databases.
302 Article 329, Copyright and Related Rights Act 2000.
303 Article 70-ter.1, Legislative Decree No 77 of 8 November 2021.
304 Article 57 b, Copyright and Related Rights Act (ZASP) 1995.
305 Article 67.7, Royal Decree 17910 24/2021 (2/11/2021).
306 Contrast Articles 3 and 4 with Articles 5 (3) and 9 of the CDSM Directive and Article 4 of the Directive 2012/28/EU on certain permitted uses of orphan works which create a cross-border effect.

As implied by Recital 15 of the CDSM Directive, while it may be assumed that the right of communication to the public under Article 5.3 (a) of InfoSoc could, in principle, be justified through the cumulative application of multiple exceptions, the absence of a cross-border unifying legislative provision leaves many legitimate questions unresolved. To illustrate this with a few examples: if a research organisation in Germany[307] were to share its training data or outputs with partners in Slovenia[308] and Spain, would sharing be lawful with Spanish partners as national law has not created a derogation for communication to the public? Furthermore, it remains unclear how the remuneration required for the making available of a work under German law is collected or even calculated when partners supply protected works into Germany from abroad.[309] If this hypothetical partnership were to exist, pursuant to Articles 3(2) and 3(3), which country or countries' security measures would apply? Furthermore, if the aim was to mine software, the presence of a German research institution complicates matters as unlike Spain[310] and Slovenia[311] German copyright law outlaws machine learning undertaken on code.[312]

Thus, while the Directive no doubt intends to support machine learning and collaboration across the EU, not least in the context of Horizon Europe, its execution leaves many important legal questions unresolved. Not only does a failure to create an explicit exception for communication to the public for commercial use stymy economic activities, the Directive's silence in the context of domestic and cross-border scientific research collaboration has resulted in confusion and a chilling lack of legal harmonisation across the European Research Area.

2.5 The Interplay of ML Exceptions with Contractual Terms and Conditions

2.5.1 Contract Override

While as shown above, many aspects of the implementation of TDM exceptions in Europe impose significant barriers to innovation, s.29A and the Directive's adept handling of the interplay between exceptions and contract law provides high levels of certainty for European innovators. Both the CDPA and Article 7 of the CDSM Directive hold that any contractual provisions contrary to s.29A and Article 3 shall be unenforceable.

This raises the interesting legal question whether those wishing to undertake data analysis now have a legal right to do so. Whereas traditionally the view that L&Es are no more than a defence against

[307] §60d (4), German Act on Copyright and Related Rights.
[308] Article 57 b (5).
[309] §60h (2)3, German Act on Copyright and Related Rights only exempts reproduction from equitable remuneration.
[310] Royal Decree 24/2021 Article 67 6 (b).
[311] Article 57 b (1).
[312] §69d (6), German Act on Copyright and Related Rights.

claims of infringement has long held sway,[313] the imperative non-contractual override provisions contained in the CDPA and Article 7 potentially tip them away from constituting mere defences into being full-blown user rights.[314] Such an interpretation is further bolstered in the case of Article 3 as, in addition to contracts, the exception is further shielded from override by technological protection measures.[315]

In legal theory, rendering proprietary entitlements unenforceable when they conflict with permitted acts affords these activities the status of *jus cogens*.[316] Most commonly associated with international law, the term denotes fundamental principles and values which cannot be contractually derogated from. Thus, both the Directive and the CDPA create an imperative and affirmative right in law for those wishing to undertake TDM that cannot be rendered null and void by the contractual practices of rightsholders.

From a public policy perspective, these normative developments are highly significant. As access to digital content is typically the result of unequal bargaining power,[317] rightsholders are largely unrestrained in their ability to create information monopolies via private ordering, often to the detriment of the public interest. Without regulators intervening to ensure that exceptions as embodiments of fundamental rights and inhibitors of market-failure are safeguarded from contractual (and technological) pre-emption, scientific and technological innovation such as AI will inevitably be stymied.

One aspect where users can be contractually excluded from undertaking TDM nevertheless remains.[318] Although according to Borghi and Karapapa it represents a departure from earlier cases, which sought to avoid overprotection of databases, in *Ryanair Ltd v PR Aviation BN,*[319] the CJEU has nevertheless held that where contents of a database reside in the public domain, licences may preclude acts that would otherwise be lawful were the work protected by copyright or *sui generis* database rights.[320] Thus, paradoxically, works unprotected by proprietary entitlements are afforded stronger contractual protections than works subject to copyright and similar forms of legal protection.

---

[313] See references to European Court cases which frame exceptions as defences: Martin Kretschmer, Estelle Derclaye, Marcella Favale, Richard Watt, *The Relationship Between Copyright and Contract Law* (SABIP 2010) n 162, 244.
[314] For an in-depth discussion of these developments, see Maurizio Borghi 'Exceptions as users' rights?' in op. cit. 248 *Routledge Handbook of EU Copyright Law*, 274.
[315] Regarding UK law, this assertion however is less certain as override by TPMs is still permitted where the material has been made available online "on agreed contractual terms". s.296ZE (9).
[316] §9.6.13. Michel Walter and Silke von Lewinski, *European Copyright Law: A Commentary* (OUP 2010).
[317] For example, between an individual and a business or between a library who must give access to books, databases etc.
[318] Highly relevant to this discussion is the lingering question regarding the status of materials regulated by contract but not subject to copyright and related rights. Pursuant to Case C-30/12 *Ryanair Ltd v PR Aviation BN* [2015] ECLI:EU:C:2015:10 (hereinafter *Ryanair v PV Aviation*), Geiger et al observe that "if a database is not protected either by copyright or by the sui generis right, the author is not precluded from laying down any contractual limitations on its use by third parties, apparently including also those preventing TDM." Geiger et al, supra 223, 19.
[319] Id. *Ryanair v PV Aviation.*
[320] Maurizio Borghi and Stavroula Karapapa, 'Contractual Restrictions on Lawful Use of Information: Sole-Source Databases Protected by the Back Door?' (2015) 37 EIPR 8 505.

Be that as it may, in lifting this significant AI innovation cost from the shoulders of research organisations and CHIs as far as possible, the European legislature has created a rare but important moment of clarity for users of exceptions. Symbolically, the issue of contract override has moved from the surrounding areas of the European copyright landscape into its heartlands – affording end-users with the *right* to mine all types of protected informational subject matter.[321] Not only will this important copyright development likely lead to a broader re-evaluation in Europe of the interplay between contracts and exceptions, but given the size of the Single Market and the influence it gives the EU in international IP fora, it also stands to change the nature of the debate globally around the prominent role that L&Es play in fostering innovation.[322]

2.5.2 Reservation of Rights: Article 4

Although the 2015 Digital Single Market Strategy proclaimed the need for a modern copyright law and the importance of the free flow of data to "maximis[e] the growth potential of the digital economy",[323] surprisingly, the initial proposal for the Directive lacked a TDM exception in support of commercial entities. The political compromise that was required in Council to propose Article 4 has however resulted in a distinctly mercurial exception that gives with the one hand while adeptly taking away with the other.

Not conditioned by a research purpose, Article 4 permits an individual or organisation to undertake commercial as well as non-commercial text and data mining without the authorisation of rightsholders. All the same, unique to Union law, it simultaneously provides the copyright owner with the authority to negate the exercise of the exception in its entirety. All that is needed for a rightsholder to prevent the automated analysis of a work is to "expressly reserve … [their exclusive rights] in *an appropriate manner*".[324] Even then, while Recital 18 declares that the reservation of rights should be "by the use of machine-readable means" rightsholders are not obliged to do this in an operationally efficient form through machine-readable protocols such as robots.txt.[325] Instead, as referenced as obiter dictum in *Kneschke v LAION,* potentially they may reserve rights in the "terms and conditions of a website", or in contractual agreements with the licensee as long as they are searchable electronically.[326] Thus, at their most basic a copyright owner making available a Word file or an OCRed PDF would seemingly comply with the Recital.

[321] Borghi, supra 314, 274 ff. (discussing limitations in the scope of the CDSM contractual override provisions).
[322] It is interesting to observe that post 2019 a Japanese government (METI) study on the issue of contractual override of exceptions made particular reference to Article 30-4 of the JCA which *inter alia* allows machine learning. Infra 663.
[323] A Digital Single Market Strategy for Europe - Analysis and Evidence. COM (2015) 192 final.
[324] Article 4 (3).
[325] Recital 18: "In the case of content that has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine-readable means, including metadata and terms and conditions of a website or a service. … In other cases, it can be appropriate to reserve the rights by other means, such as contractual agreements or a unilateral declaration."
[326] Ibid.

Although normatively the legislature *ab initio* is the arbiter of whether a copyright-relevant act undertaken without a rightsholder's authorisation is lawful or not, Article 4 ultimately cedes the authority to determine the applicability of the exception to the rightsholders themselves. Thus, unique in European law, the authorities have arguably denuded themselves of their own law-making powers (or more accurately: the zone of influence of those powers), situating the final right to the exercise of Article 4 in the hands of the copyright owner. Inverting the norms of limitations and exceptions where, for public policy reasons, certain acts are rendered permissible by the legislature, instead rightsholders are vested with the power to arbitrate whether or not to reserve their proprietary entitlements and allow the act of machine learning.

Conceptually, the closest type of derogation to this "waivable exception" are the so-called hybrid exceptions that combine permitted acts and licensing. For example, Articles 5 and Articles 8–11 of the CDSMD (and in the UK s.35, which this author believes the hybrid exceptions introduced by the CDSM Directive were modelled upon[327]), permit a user privilege to be removed by rightsholders on the proviso that a licence is readily available that provides, as a minimum, the same rights afforded by that exception. Although hybrid exceptions provide discretion to the author and their representatives in choosing the mechanism used to provide the permitted activities, Article 4 goes well beyond, as it provides absolute agency to the copyright owner to remove enjoyment of the statutory limitation in its entirety. In other words, the right to engage in a statutorily prescribed activity through either a licence or an exception is eliminated, leaving end-users entirely subject to the will of private actors.

The ability under Article 4 to reserve rights, and the countervailing pressures that arise to negotiate a licence, raise issues of both an economic and legal nature. Unlike in countries such as Japan, where commercial AI-related activities are likely protected from contract override,[328] an innovation burden is placed on European companies to negotiate licences for materials that they may well already have lawful access to. Thus, adding significant cost, delays, and the potential abandonment of projects to the many legal factors that European companies must contend with, which their international competitors do not.

Not only does truncating access in this way increase the chance of AI bias and uncompetitive algorithms,[329] with the advent of a more interventionist algorithmic regulation environment, *a priori* it creates issues of conflict of law. Any licence for machine learning is likely to have strict security

[327] s.35, CDPA. Recording by educational establishments of broadcasts.
[328] 『令和３年度産業経済研究委託事業（海外におけるデザイン・ブランド保護等新たな知財 制度上の課題に関する実態調査「別紙２」新たな知財制度上の課題に関する実態調査）調査報告書』株式会社エヌ・ティ・ティ・データ経営研究所 | 2021 Industrial and Economic Research Commissioned Project: A Survey of New Intellectual Property System Issues Relating to Design and Brand Protection Overseas – Appendix 2 (NTT Data Institute of Management Consulting, Inc 2021) 36.
[329] Amanda Levendowski, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem' (2018) 93 Washington Law Review 579.

prohibitions concerning the communication of data to third parties or contain obligations to destroy training data during or after the licence period has expired. Licensing constraints may therefore contradict obligations in bodies of law external to copyright such as the Digital Services Act,[330] Digital Markets Act[331], and AI Act[332] which require, upon a reasoned request, data sharing with competent authorities. Not only does this create legal issues over whether public or private law should take precedence, in terms of practice it may dissuade a potential innovator from licensing data in Europe in the first place, as they are concerned about liabilities arising from communicating data to the authorities in conflict with their contractual obligations.

2.5.3 Lawful Access

Lawful access is an explicit requirement of both the UK exception and the CDSM Directive. Undefined in both statutes, Recital 14 of the Directive nevertheless elaborates that lawful access for research organisations *includes* access to works as a result of open access policies, contractual agreements, as well as subject matter made "freely available online". In practice, within the digital realm, this will often require assent to terms of use, typically set out in a licence or other form of contractual arrangement.

How to interpret a lawful access requirement remains open to question. A data scientist is hardly likely to know the legal route that a particular work took in order for it to be accessible to them in a subscription database, let alone in a less structured environment like the open web. Given the frequently unknowable origin of a work, as highlighted by the European Copyright Society, if lawful access were to be viewed the same as a "lawful source," it would frustrate the purpose of the exception as it would make all but the most exhaustively researched TDM project potentially unlawful.[333]

Related issues are raised by other commentators in regard to Article 4. Taking the Panama Papers as an example, Geiger questions how a court may rule on an investigative journalist using contentious or sensitive materials, particularly when the rightsholder was unaware of the use and thus unable to object or withhold permission. Going further than the European Copyright Society, Geiger thus indicates not just a lawful access requirement, but a lawful source requirement may abut fundamental rights in European and international law.[334]

[330] Article 72, DSA.
[331] Article 21, DMA.
[332] Article 21.1, AI Act ("Providers of high-risk AI systems shall, upon a reasoned request by a competent authority, provide that authority *all the information and documentation* necessary to demonstrate the conformity of the high-risk AI system".) (Emphasis added).
[333] Supra 199.
[334] Christophe Geiger, 'The Missing Goal-Scorers in the Artificial Intelligence Team: Of Big Data, the Fundamental Right to Research and the failed Text and Data Mining Limitations in the CSDM Directive' (2021) 66 PIJIP/TLS Research Paper Series 2021 2, 6ff.

A further important question arising from the lawful access pre-condition is whether project partners may share protected works when not all participants have lawful access to the original. Although only a Directive's main text is legally binding, Recital 14 states that in the context of university subscriptions it will extend to those "persons attached thereto and covered by those subscriptions". This may well suggest that only researchers affiliated to institutions that have paid for legal access to the content they are analysing can undertake TDM for scientific research purposes. Such a strict interpretation would arguably run counter to the intent of Directive, as it would render many scientific projects unworkable; even in the era of the "Big Deal" universities will, as a matter of course, have vastly differing subscriptions. Furthermore, requiring a university or a PPP partner to establish the gaps in the collection and *ex-post* license access from a multitude of different content providers will, if even possible, significantly hinder any project.

2.6 Technological Protection Measures

Alongside contract override of statutory entitlements, another feature of the digital age has been the dominant and exclusionary status afforded technological protection measures (TPM). As set forth in the Information Society Directive, TPMs encompass "any technology, device or component that … is designed to prevent or restrict acts … which are not authorised by the rightholder."[335] They cover both access control mechanisms, which prevent access entirely,[336] as well as measures such as copy restriction technologies aimed at controlling the copying of a work to which the user has access.

Although in transposing TPM obligations from the WCT[337] and WPPT[338] some WIPO Members States have only designated circumvention of TPMs unlawful where the act does not interfere with the legitimate interests of the rightsholder – leaving users comparatively free to circumvent protection measures[339] – the 2001 Information Society Directive has a blanket prohibition on circumvention without further intermediation by the authorities.[340] Whereas copyright law renders numerous acts involving protected subject matter permissible, as with contract override of exceptions, the European legislator has hitherto been content to erode the exercise of statutory user privileges. Only if TPMs frustrate the exercise of a subset of InfoSoc's enumerated exceptions,[341] and "in the absence of

[335] Article 6 (3), InfoSoc.
[336] Article 6 (3) gives encryption and scrambling as examples of access control measures.
[337] Article 11, WCT.
[338] Article 18, WPPT.
[339] For example, see Article 39a (4) Swiss Federal Act on Copyright and Related Rights. Also in Japan, circumvention of copy control mechanisms is only unlawful for private copying, and circumvention of access control measures for research and technological development purposes linked to TPM technologies is lawful pursuant to Article 113 (3).
[340] See supra 228, *Concise European Copyright Law*, 391.
[341] Pursuant to Article 6 circumvention is only permitted for Articles 5(2) (a),(c), (d),(e), 5.3 (a), (b)(e) and under certain circumstances 5(2)(b).

voluntary measures" does the Directive require that Member States intervene to provide the means for beneficiaries to enjoy their statutory entitlements.[342]

Further truncation of the ability to enjoy L&Es may befall users of online digital media. Art 6 (4) subparagraph four holds that Member States cannot intervene when the work at issue "is made available to the public on agreed contractual terms in such a way that members of the public may access them from a place and at a time individually chose by them."[343] Given the growing obsolescence of off-line electronic media, copyright flexibilities are in most cases therefore now rendered null and void by technical measures.[344]

While these conditions remain unchanged in the UK despite its departure from the EU, notably the CDSMD has sought to reduce the scope of InfoSoc's broad prohibitions on circumvention. By dint of Article 7 nullifying subparagraph four of Article 6(4) of the 2001 Directive in relation to the operation of Articles 3-6, beneficiaries of these exceptions are in theory now furnished with an unqualified legal right to circumvent technological protection measures.[345] How a person blocked by TPMs asserts their right, and how long it will take to resolve, is not addressed, rendering important procedural questions unharmonised under Member State law.

It can be observed that current Member State mechanisms for facilitating circumvention are generally convoluted, with some countries leaving the law entirely silent on the remedial pathways available to beneficiaries.[346] For example, in Germany, the Copyright Act establishes no more than the right to "pursue" a rightsholder.[347] In Italy, in the light of Article 71-*quinquies* which was left unchanged at the time of the transposition of the CDSM Directive, it remains unspecified how pursuant to Art 70-*sexies* research organisations in Italy should interpret the process for lawful circumvention for TDM under Article 7.[348] In support of text and data mining, Slovenian and Irish law are more obliging. They require rightsholders to give access within 72 hours and 30 days respectively of becoming aware that TPMs are preventing beneficiaries from undertaking machine learning.[349] Where access is not

[342] Only Portugal and Slovakia allow beneficiaries to circumvent TPMs themselves. Brigitte Lindner and Ted Shapiro (eds), *Copyright in the Information Society* (Edward Elgar 2019) 763, 843 ff.
[343] Article 6 subparagraph 4.
[344] See the UK Libraries and Archives Copyright Alliance (LACA)'s unsuccessful attempt to petition the UK Government to allow circumvention of captcha technology under s296ze of the CDPA for the purposes of text and data mining. https://librariesarchivescopyright.files.wordpress.com/2020/09/f6b15-notice_of_complaint_to_the_secretary_of_state_-_test_case_1_0.pdf
[345] Article 7 (2). Surprisingly ten EU countries have explicitly extended subparagraph four of InfoSoc to Articles 3-6 of the CDSMD. See upcoming article by Rossana Ducato, 'The text and data mining exceptions in the national implementations: a comparative study.' (Working Title).
[346] Id. See TPM Table 2.
[347] German Copyright Act, §956(2).
[348] Deborah De Angelis 'Un'analisi della ricezione italiana delle eccezioni e imitazioni obbligatorie introdotte dalla direttiva 2019/790/ue nel campo delle attività digitali di insegnamento, ricerca e accesso al patrimonio culturale' (An analysis of the Italian reception of the mandatory exceptions and limitations introduced by Directive 2019/790/EU in the field of digital activities for teaching, research, and access to cultural heritage.) (2022) N.4 Il Diritto di Autore, 610, 616-7.
[349] Article 57 (b) (4), Copyright and Related Rights Act (ZASP) 1995 (SI). Article 374(4) Copyright and Related Rights Act 2000. (IE).

forthcoming, next steps are less time-sensitive, as in Ireland[350] a data scientist will have to apply for an order from the courts, while Slovenian law sets out that a beneficiary may request mediation.[351]

Technically, Article 7 permits circumvention of TPMs pursuant to both Articles 3 and 4. Nevertheless, in practice, the benefits that accrue from its application to Article 4 are somewhat moot. While in principle circumvention privileges are afforded its beneficiaries, it seems probable that if the intermediated circumvention process outlined in national laws were initiated,[352] the most common response from the copyright owner would be to *ex post* assert a reservation of rights. Thus, while the Directive seeks to remove TPMs as barrier to machine learning for commercial purposes, the primacy of the reservation of rights in many instances will likely render the intervention by the European legislature superfluous.

Section One: Summation

The failure by the UK and European legislator to maximise the flexibilities granted to them in international law to clearly facilitate AI seems all but guaranteed to exacerbate Europe's innovation paradox. The commercial and non-commercial fault line that runs through European copyright law looms large as a principal reason why the region's governments have accorded comparatively weak support to TDM and other forms of machine learning.

The reluctance of European regulators to facilitate commercial TDM speaks not only to a failure of policy join up across the institutions of government, but to deep-seated jurisprudential traditions.[353] Despite modern society being reliant on computers and computer networks, whose almost every action engages copyright, European policy makers continue to see its prime objective as one that offers protection to a discrete category of rightsholder hailing from the creative industries,[354] thus overlooking what it actually has become in the information society: a mechanism that determines how the majority of citizens and businesses may access, communicate, and derive value from information and data in all its forms.

One significant caveat has been the protection of machine learning exceptions from contractual pre-emption, and in the case of the EU, the explicit safeguarding of TDM from override by technological protection measures. In the digital world, where private ordering of copyright works is the norm, in

[350] Article 374 (4).
[351] Article 166c, ZASP.
[352] Lindner and Shapiro, supra 342. However, according to Ducato supra 345, Slovakia has not implemented Article 7 in relation to the TDM exceptions and therefore circumvention of materials made available subject to agreed contractual terms remains potentially unlawful.
[353] Dinusha Mendis, Benjamin White and Dukki Hong, *Copyright and Open Norms in Seven Jurisdictions: Benefits, Challenges & Policy Recommendations* (CIPPM 2024).
[354] See Recitals 9-10 of the Information Society and the strong emphasis on natural and legal persons from the creative sector.

ensuring that rightsholders cannot override research and education related exceptions with contracts or indefinitely frustrate access through the application of TPMs,[355] the European legislator has arguably changed the global discourse that exists around L&Es. By turning defences against infringement into fully-fledged user rights[356] that cannot be nullified, the historical narrative around proprietary entitlements being the paramount stimuli of creative expression has ineluctably shifted. For reasons of innovation and competitiveness, limitations and exceptions as sources of innovation and technological advancement are no longer simply proffered by the legislature but, within certain boundaries, actively defended.

Section Two: Computer-Generated Works (CGW)

Turning from copyright derogations to restricted acts, many decades prior to the widespread adoption of TDM exceptions in Europe, the UK government had invested its attention into another facet of AI: so-called "computer-generated works".[357] In a significant departure from the three governmental reports that laid the foundations for the 1988 copyright reforms,[358] the Act dispensed with all consideration of computer-assisted works.[359] Instead, it announced that s.9(3) of the CDPA shall apply to works "generated by computer in circumstances such that there is *no human author* of the work."[360] The rights of this suppositional non-human author are bestowed for a period of 50 years on "the person by whom the arrangements necessary for the creation of the work are undertaken."[361] No doubt influenced by this, the legislatures of Ireland, New Zealand, India, South Africa, and Hong Kong have similarly followed suit in introducing copyright provisions for autonomous computer-generated works.[362]

As highlighted by the Shadow Secretary of State for Industry at the time,[363] true artificial intelligence – where machines display their own independent volition and mental acuity – was not a reality of the time, and of course remains so to this day. Despite calls from the European Parliament as recently as

[355] Europe's TDM Exception for Research: Will It Be Undermined By Technical Blocking From Publishers? Ligue des Bibliothèques Européennes de Recherche). Explaining the results of its 2020 university survey with LACA that where TDM had been discovered "[a]ctions taken by publishers included 1) suspension of campus-wide access to paid for electronic subscriptions 2) threats to cut off access to content unless TDM was stopped 3) technically limiting downloads to one document only 4) a request for additional payments and 5) the introduction of CAPTCHA technology to frustrate TDM.
[356] cf. Borghi, supra 314.
[357] s.9 (3), s.12 (7), s.178, CDPA.
[358] Whitford Committee, *Copyright and Designs Law: Report of the Committee to consider the Law on Copyright and Designs* (March 1977) 132-3. Green Paper, 'Reform of the Law relating to Copyright, Designs and Performers' Protection: A Consultative Document' (1981) 35. White Paper, Secretary of State for Trade and Industry, 'Intellectual Property and Innovation' (April 1986) 51. Interestingly the White Paper concludes that in terms of computer-assisted works "no practical problems arise from the absence of specific authorship provisions in this area".
[359] Pressure from the British Computer Society (BCS) erroneously asserting that if the UK did not extend protection to computer-generated works originating from the UK, they may not be protected under US law may explain this *volte face.* RJ Hart, 'Copyright and computer generated works.' (1988) Aslib Proceedings, 40 6 173, 176-8.
[360] s.178, CDPA (Emphasis added).
[361] s.9 (3), CDPA.
[362] Copyright and Related Rights Act 2000, s.21 (No 28/2000) (IR); Copyright Act of 1994, s.5 (NZ); The Copyright Act of 1957, s.2 (d) (vi) (IN); Copyright Act 1 (2) (h) (98/1978) (SA); The Copyright Ordinance, s.11 (3) (Cap 528 / 1997) (HK).
[363] See House of Commons Debate 28 April 1988 vol 132 (Second Reading). §539. Bryan Gould MP stating "We are the first legislature to recognise the advent – not here yet but coming – of artificial intelligence."

2016 for "the Commission to elaborate criteria for an 'own intellectual creation' for copyrightable works produced by computers or robots",[364] as discussed in the previous chapter, the fact is that those AI models which generate text,[365] draw,[366] or compose music[367] are not autonomous. None of these creative acts would be possible without the "motive force" of human contributors.[368] It is the human author viz. data scientist in tandem with the model's end user who ultimately shapes the "choice as to the contents and presentation" [369] of a computer-generated work, and without whom copyright as an anthropocentric body of law would not arise.

From a jurisprudential perspective, while the use of most other assistive technological tools to create a copyright work is now undisputed,[370] the role of machine learning software as a production tool generates voluminous scholarly debate.[371] A standard commentary on computer-generated works takes, as a starting point, the autonomy of the machine and therefore the notional absence of the human author. Absent the actor upon which all copyright law is conceived and built, the discourse proceeds to adjudicate on which human proxy is most appropriate to be awarded the economic incentives that exclusive rights afford. The discussion is complicated by the fact that multiple actors may in reality contribute to a model output, and at differing levels of abstraction *both* humans and computer functions are involved in the construction of computer-generated works. Nevertheless, as observed by Cronin, the correct question – and the one that commentators are in point of fact discussing – is "not whether a computer can be considered an author, but rather what is the appropriate assignment or apportionment of copyright in computer-generated works between human programmers and human users of their software programs."[372]

Whilst there is no agreement in the literature as to who should be awarded copyright in computer-generated outputs, three differing doctrinal approaches to the question can be discerned. These pertain to: the fundamental legal tenets of authorship (legal principle arguments); those which focus on the who or what generated the output (causational arguments); and finally, those which are focused on economic policy (economic incentive arguments). In accordance with these three often overlapping categories, the following section will highlight the most common rationales presented by

[364] Draft report with recommendations to the commission on civil law rules on robotics (2015/2103 (INL)) Committee on Legal Affairs.
[365] e.g. Natural language processing models.
[366] e.g. General adversarial networks (GANs), neural style transfer models, diffusion models etc.
[367] e.g. Variational autoencoders (VAEs), convolutional neural networks (CNNs), recurrent neural networks (RNNs) etc.
[368] Bridy, supra 1 (motive force being a term deployed by Bridy to encapsulate the notion of originality as used in copyright law.)
[369] Jane Ginsburg 'The concept of authorship in comparative copyright law.' (2002) 52 DePaul L Rev1063, 1077 (a term used by Ginsburg to define authorial originality).
[370] Bridy, supra 1, 5 ff. (discussing the landmark US cases *Burrow-Giles Lithographic Co v Sarony* 111 US 53 (1884) and *Bleistein v Donaldson Lithographing Company* 188 US 239 (1903) that resulted in the photographer of a work being recognised as the copyright owner irrespective of their reliance on a machine as a tool.)
[371] It is interesting to note that this is not a new phenomenon. The first article in English on the issue of computer-generated works dates back to 1966. John Banzhaf III, 'When a Computer Needs a Lawyer.' (1966) 71 Dick L Rev 240.
[372] Charles Cronin 'Virtual Music Scores, Copyright and the Promotion of a Marginalized Technology.' (2004) 28 Colum JL & Arts 1 14.

commentators in making their determination as to whom, if anyone at all,[373] copyright should be awarded to in relation to an original output generated by a computer.

1. Legal Principle Arguments – The Case of Authorship

Arguments centred on jurisprudence pivot around the fundamental requirement of copyright law that an author is human. In focusing on the doctrinal requirement for personhood, the question whether a work authored by an autonomous computer with no human input can be accommodated by copyright law must be answered in the negative. As highlighted by Ricketson, owing to the absence of a human author, copyright laws such as those seen in the UK – that designate works produced by autonomous AI as protectable subject matter – are non-compliant with the international copyright treaties that govern WIPO Member States.[374]

Although the Berne Convention contains no explicit reference to a human author, its anthropocentric nature is very much in evidence. Most obviously, the duration of copyright as set forth in Article 7(1) lasts a minimum period of 50 years *post mortem auctoris*. Further, as Ricketson indicates, the requirement for personhood is underlined by the protection of moral rights in the Convention which self-evidently "make[s] no sense other than in relation to human authors."[375]

A mix of statute and case law at the national and supra-national level similarly confirms that human personhood is a prerequisite for the granting of authorial entitlements. In the US, the Copyright Act not only refers to the death of the author but the rights *post mortem* belonging to an "author's widow or widower" – evidently a state of affairs strictly the preserve of a living person.[376] Additionally, the US Copyright Office has declined to register computer-generated artwork on a number of occasions on the grounds they lack human authorship.[377] Perhaps most memorable of all, on appeal in *Naruto v Slater*[378] the Ninth Circuit denied authorship to a macaque that had taken a selfie, opining that non-humans "lack … statutory standing under the US Copyright Act".[379]

[373] See Jared Vasconcellos Grubow, 'O.K. Computer: The Devolution of Human Creativity and Granting Musical Copyrights to Artificially Intelligent Joint Authors' (2018) 40 Cardozo Law Review 387, 419 and Shlomit Yanisky-Ravid & Luis Velez-Hernandez, 'Copyrightability of Artworks Produced by Creative Robots and Originality: The Formality-Objective Model' (2018) 19 Minn J L Sci & Tech 1 (both arguing that authorship should be granted to the AI.)
[374] Ricketson characterises s.9 (3) as a breach of the Berne Convention due to its non-standard duration of 50 years and a failure to award moral rights. Sam Ricketson, 'The 1992 Horace S. Manges Lecture - People or Machines: The Bern Convention and the Changing Concept of Authorship' (1991) 16 Colum VLA JL & Arts 1, 30.
[375] Ibid.
[376] For example, US Copyright Act §101.
[377] A 2018 application from Imagination Engines Inc where the author of the artwork 'A Recent Entrance to Paradise' was given as the Creativity Machine was rejected by the Copyright Office a decision which was subsequently upheld by a federal court in *Stephen Thaler v Shira Perlmutter* Civil Action No. 22-1564 (BAH) District of Columbia (18/8/2023).
[378] *Naruto v Slater* No 16-15469 (9th Cir 2018) 4.
[379] Ibid.

As signatories of the international treaties, European copyright law ineluctably displays the same anthropocentrism. Both statute[380] and CJEU rulings[381] conform with the human personhood requirements of Berne. Arising from a slew of case law from the European Court, the author's own intellectual creation test is horizontally harmonised, and as Guadamuz observes, emphatically denies copyright protection to any other entity other than a human being.[382] Absent of any new and divergent ruling post-Brexit, the same test applies in the UK today, thus bringing into question s.9(3)'s compliance not only with the Berne Convention, but UK jurisprudence itself.[383]

Ramalho brings our attention to the cornerstone of copyright itself: the idea-expression dichotomy. Underlining the humanistic origins of ideas, she posits that "if the expression is created by an AIs [*sic*], it is unclear to what extent we can talk about an 'idea' that is being expressed."[384] Similarly, Mezei focuses on the anthropocentric nature of copyright jurisprudence, arguing that the lines of code of a software program are no substitute for human authorship and cannot replace creativity, rooted as it is in human achievements.[385] Furthermore, he questions how software can exercise those rights that the law treats as moral and personal, pondering a machine can decide that their right of integrity has been abused and when to disclose or withdraw their work. Gervais[386] points to other facets of personhood, notably the Benthamite view that rights correlate with and bring with them responsibilities.[387] According to correlativity theories, ownership cannot be separated from commitments. For example, due to a lack of legal personhood, machines are unable to litigate to defend their own proprietorial interests, and conversely cannot be litigated against. Thus, according to a Benthamite view of the world, unable to assume their legal responsibilities, machines should not be able to enjoy the rights afforded by the law.

This line of argumentation has particular resonance in countries like the UK who have introduced provisions for autonomous computer-generated works. Content to accept the bestowal of intellectual property rights, it can be asked whether endogenously and exogenously to copyright, the robot's human proxy is also willing to accept the legal liabilities that may arise from the actions of their autonomous AI. In other words, lacking legal personhood, while a notional autonomous robot which

[380] The author's own intellectual creation requirement is codified by Article 1(3) Software Directive, Article 3(1) Database Directive and Article 6 of the Term Directive.
[381] The author's own intellectual creation test has been applied by the CJEU inter alia in *Infopaq 1*, *Painer*, Case 393/09 *Bezpečnostní softwarová asociace - Svaz softwarové ochrany v Ministerstvo kultury* [2010] ECR I-13971 (hereinafter *BSA*), *Case* C-604/10 *Football Dataco Ltd and Others v Yahoo! UK Ltd and Others* [2012] ECLI:EU:C: 2012:115 (hereinafter *Football Dataco*).
[382] Guadamuz, supra 271, 178.
[383] Estelle Derclaye, 'Infopaq International A/S v Danske Dagblades Forening (C-5/08): wonderful or worrisome? The impact of the ECJ ruling in Infopaq on UK copyright law.' (2010) 32 EIPR 5 247, 250. Post Brexit the continuing operativity of the author's own intellectual creation has been confirmed in of *THJ Systems Ltd v Sheridan* [2023] EWCA Civ 1354.
[384] Ana Ramalho 'Will Robots Rule the (Artistic) World? A Proposed Model for the Legal Status of Creations by Artificial Intelligence Systems.' (2017) 21 Journal of Internet Law 1 12, 14.
[385] Péter Mezei, 'You AIn't seen nothing yet: Arguments against the protectability of AI-generated outputs by copyright law' in Maurizio Borghi and Roger Brownsword (eds), *Law, Regulation and Governance in the Information Society* (Routledge 2022) 126
[386] Daniel Gervais, 'The machine as author' (2019) 105 Iowa Law Review 2053.
[387] Jeremy Bentham, *The Works of Jeremy Bentham Volume III* (William Tait 1843).

has achieved singularity would not be liable under existing laws, would the human proxy who made the arrangements for them and their computer-generated work assume the legal obligations and potential liabilities of their mechanical ward?

In sum, doctrinal arguments in the literature focusing on personhood serve to underline that copyright law in its current form does not accommodate authorship without an attendant human creative touch. Thus, by extension, computer-generated laws where framed as a form of copyright are not only doctrinally ill-fitting, but as asserted by Ricketson, unequivocally in breach of international obligations.

2. Causational Arguments – The Case of Effort

Stripping away the layers of theoretical debate found in the literature, much of it divorced from the technical realities of machine learning, as mentioned above, the substantive legal question commentators are in essence asking is which human contributor should be afforded copyright in AI-*assisted* works. In common with the inherent complexities of determining ownership of copyright in works with multiple contributors, such as a film or a popular song, *who* causes the outputs from AI to come into being suffers not from a potential lack of human authorship, but from an abundance of authorial candidates.

The primary potential contenders for authorship include the copyright owner(s) of the training data, the originator(s) of the AI software, the data scientist who trained the model and /or their employer, and the user or licensee of the generative model. Thus, when rightsholders of the information used to train the models are included, it is potentially a cast of many thousands, and arguably more than the legal framework has the capacity to deal with.

At the heart of causative arguments lies the principle of transitivity. That is to say, that if a relationship holds between the first and the second elements of a relationship, as well as the second and third, it must exist between the first and third elements also. Doctrinally, the fact that a computer directly produces art, text, and sound whilst copyright ownership sits elsewhere matters little, if at all. Degrees of separation between the copyright owner and the entity that directly undertakes the delivery and execution of the work are common in copyright law.[388] In both civil and common law jurisdictions, "work for hire" constructions are well known,[389] as are examples of the separation of

[388] Jane Ginsburg and Luke Ali Budiardjo 'Authors and machines.' (2019) 34 Berkeley Tech LJ 343, 358 (discussing the conception – execution theory of authorship and amanuensis.)

[389] For example, §8 of the Dutch Copyright Act; Article 15 Japan Copyright Act; §101 US Copyright Act; Article 11 Copyright Law of the People's Republic of China; s.11 (2) CDPA; Article 2 (3) Computer Programs Directive; Article 4 (1) Database Directive etc.

ownership from direct execution and fixation, which arise as a matter of course in many cinematographic works,[390] phonograms,[391] and computer programs.[392]

Copyright is afforded not to the person who physically labours to produce the work, but to the motive force that provides the "intellectual labour".[393] In UK law, for example, since *Donoghue v Allied Newspapers*, it is settled as a matter of law that a person who merely transcribes the words of another will not be afforded copyright.[394] Similarly, a person employing an amanuensis to precisely follow their detailed instructions will remain the copyright owner.[395] Pursuant to a long line of case law, Hedrick points to what she calls the "mastermind doctrine", evoking the notion that it is not those who physically craft a work but the person who intellectually conceives it that is awarded copyright.[396] One particularly striking example of amanuensis is the case of *Lindsay v The Wrecked and Abandoned Vessel RMS Titanic*, in which the courts granted copyright to the film director who assiduously planned every scene in advance, despite the fact that they were not present when the underwater filming took place.[397]

Likewise, it is long settled in law that not only human intermediaries but also technological tools can mediate between the author and the copyrightable output.[398] This is evident not only in cases regarding computers[399] but longer established situations involving authors using cameras,[400] video recorders, as well as the plethora of different machines which capture sound recordings.[401]

Turning to the specific issue of the hominal *who* should be awarded copyright; although software 2.0 generative models are programmed in part by the data they ingest during the training phase, little serious academic consideration is given to the licensors or copyright owners of the data as potential grantees of authorship. Even if it is the data that informs the weights and other parameters of NLP models, VAEs, GANs, etc, it would be difficult to see how originators of the data, absent any other role, could be viewed as having a direct causational effect on the particular form and expression

---

[390] Article 5, Berne Convention.
[391] Article 5, International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organizations (1961).
[392] Article 2 (3), Computer Programs Directive.
[393] *Trade Mark Cases* 100 US (1879) 94.
[394] *Donoghue v Allied Newspapers Ltd* [1938] Ch 106 "A mere amanuensis does not, by taking down word for word the language of the author, become in any sense the owner of the copyright."
[395] For example, *Andrien v So. Ocean Cty Chamber of Commerce* 927 F 2d 132 (3d Cir 1991) where the court ruled that the plaintiff was the sole author as he had "directed the copy's preparation in specific detail." Ginsburg and Budiardjo, supra 388.
[396] Samantha Fink Hedrick, 'I think, therefore I create: claiming copyright in the outputs of algorithms.' (2018) NYU J Intell Prop & Ent L 8 324.
[397] *Lindsay v The Wrecked and Abandoned Vessel RMS Titanic* 52 USPQ 2d 1609 (SDNY 1999).
[398] For example, the use of Photoshop in *Temple Island Collections Ltd v New English Teas Ltd & another* [2012] EWPCC 1 to create the final artwork was of no concern to the courts.
[399] For example, prior to the introduction of a provision for computer-generated works in the UK in 1988, courts had already rejected the notion that computer-assisted works could not be protected by copyright. For example, *Jockey Club v Rahim* unreported 22/7/1983 (relating to computers producing computer-generated lists of runners and riders in horse races; unreported), *Express Newspapers plc v Liverpool Daily Post & Echo* (1985) 1 WLR 1089 (concerning computers generating grids).
[400] As early as 1862 in France the courts ruled in *Mayer & Pierson v Betbeder & Schwabbé*, that a photograph was copyright subject matter. The same year in the UK the Fine Arts Copyright Act extended copyright to "paintings, drawings, and photographs".
[401] Rights in sound recordings were first granted protection in the UK in 1911 and in international treaties, by the Rome Convention 1961.

contained in the output. In the words of Ginsburg and Budiardjo, neither do they "conceive" of, or "execute" the work produced by the model, thus any claim of authorship is unlikely to succeed.[402]

The majority of commentators who put forward causational arguments focus on two prime contenders for authorship either in isolation or as joint authors: the "programmer" on the one hand (*ergo* the data scientist, and by extension their employer) and the "user" of the AI on the other.

Turning first to the programmer, as outlined in Chapter 1, the role of the data scientist is central to the creation of the model. [403] As a consequence, they are a pivotal in shaping the general form and expressional typologies of the output. The fact software can write (compose or draw) at all is a result of the physical and intellectual labour and choices made by the programmer. The output, including where random factors are programmed in, are entirely determined by the model type, training data, the hyperparameters selected in the training phase, and the output instructions provided. Therefore, in contrast to traditional software, where having decided on the suitability of the software to the task at hand, the user entirely determines the nature of the work they produce, the expression contained in a model output is shaped by the calculations "written" by the training data,[404] and the hyperparameters and other output restrictions embodied in the code as defined by the data scientist. Underlining further the crucial role of the programmer, a model may produce the same expressional output irrespective of which human user generated it.

Be that as it may, as Samuleson and Denicola indicate, granting authorship to the programmer from a doctrinal perspective is not well aligned with the incentive rationale of copyright.[405] As author of the software, it is argued, copyright already provides incentives to programmers to create and then sell or license their works.[406] By contrast, affording authorship solely to the programmer weakens the incentive for the user to create and disseminate new works – the very purpose of copyright law.

Moving to the software user, in practice, there are a spectrum of production modalities, and their role will vary from substantive on the one hand through to no more than selecting certain pre-defined prompts presented in the GUI on the other (see figure below). Particularly in B2B markets, the respective roles of the programmer viz. organisational licensor vis a vis the licensee may take many different forms. For example, users can license standardised industry AI services located in the cloud

[402] Ginsburg and Budiardjo, supra 388, passim.
[403] See Timothy Butler, 'Can a Computer be an Author? Copyright Aspects of Artificial Intelligence.' (1981) 4 Comm Ent L J 707 and Andrew Wu, 'From Video Games to Artificial Intelligence: Assigning Copyright Ownership to Works Generated by Increasingly Sophisticated Computer Programs' (1997) 25 AIPLA Q J 25 131, 174 (arguing that copyright for purely AI-generated works should be granted to the creator (or owner) of the software.)
[404] Op. cit. Karpathy, supra 53.
[405] Robert Denicola 'Ex Machina: Copyright Protection for Computer Generated Works' (2016) 69 Rutgers UL Rev 251, 283 and Pamela Samuelson, 'Allocating Ownership Rights in Computer-Generated Works' (1986) 47 U Pitt L Rev 1185, 1207.
[406] See Chapter 5. Whereas this is almost certainly the case for traditional software types, whether copyright inheres in many machine learning models is in practice far less certain.

(e.g. SaaS), or deploy the model locally, commonly providing the licensee with more implementational flexibility and agency in the further development of the software. At one extreme, a putative "user" may upload data for data analysis or training purposes with the direct intention of simply extracting an answer to the business need they have. At the other, they may engage in hyperparameter tuning including potentially contributing their own code to the model.

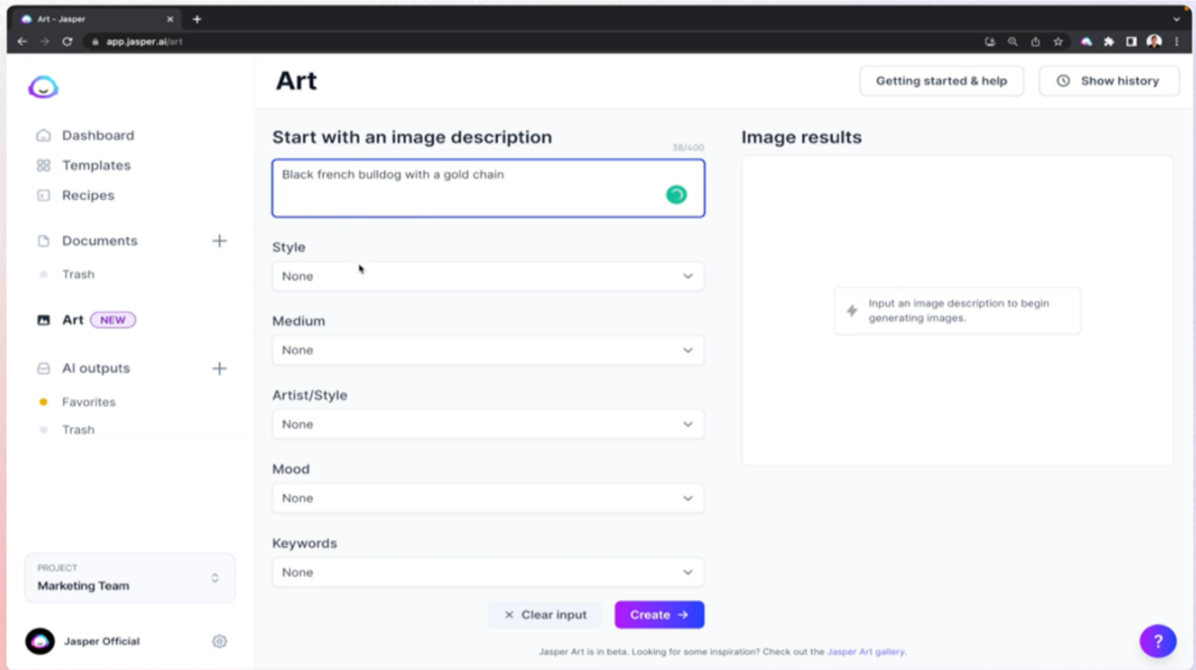

Fig. 2-2. Example of the Pre-defined Menus Provided by AI Online Generative Art Services [407] (jasper.ai)

In B2C markets, the picture is more uniform. As a rule, a provider will offer a service which allows users to produce an output such as a route, a translation, or a piece of art or music. This may involve little more than the selection of a pre-defined prompt or the input of a keyword or a pre-defined body of work. Based on the inputs, within the boundaries set by the model type, the AI creates and determines the final form and expression of the output.

Similar to other tool-reliant art forms, in line with the transitivity principle, the human user operates at a degree of separation from the work created. In terms of causation, at higher more abstract levels, the "contents and presentation"[408] produced by the AI are shaped by the decisions made by the programmer concerning the model type and the software configurations selected during the training phase. At a lower more immediate level, the user defines the precise instructions provided to the AI.

[407] <https://app.jasper.ai/art> accessed 1 May 2025
[408] Ginsburg, supra 369, ibid.

Ultimately, the exact form of the expression in the computer-generated work is nevertheless determined by calculations undertaken by the machine.
Thus, while the broad scope of the limits within which the AI can operate have been "human fixed", the precise shape and exact delivery of the work itself is determined by the trained model. Technically speaking, shaped by the training data, hyperparameter and parameters, in response to inputted instructions, the calculations that are undertaken across the hidden layers of a neural network determine the exact form of the output. In the language of copyright therefore, the AI is clearly a tool for human creativity, even if one that operates with some of its own leeway as defined by the programmer and user. Much like abstract artists engaged in drip painting, or Dadaists such as Duchamp[409] whose tools determine the precise form of their art, the intellectual labour in the framing of the final expression in the work is provided not by the machine, but the human author who instructs it.

Given the inevitable case-by-case nature of human causation and contribution, it is perhaps unsurprising different conclusions as to who the copyright should be awarded to in a generative work are put forth. Guadamuz recommends the CDPA's response, despite it offering little if any precision as to upon whom copyright should be conferred.[410] Bridy suggests that corporations should be given copyright under the work-for-hire doctrine where works are autonomously produced.[411] Wu,[412] however, concludes, in this author's view quite rightly, that there is no bright line answer due to the complex nexus of contributors and executions, meaning that any determination on copyright ownership will have to be undertaken on a case-by-case basis.

Returning to UK law, from a causational perspective answering the question who made the "arrangements necessary for the creation of the work" results in an elaborate array of contenders. Faced with the multiple pathways to create a CGW, s.9 (3) offers no guidance on how to differentiate and evaluate the respective contributions of the numerous actors involved in producing outputs from an AI model. As will be discussed in the next section, when courts in jurisdictions that have no such provision have addressed computer-generated works, although driven by the same economic rationales that underpin s.9 (3), they are evidently not disadvantaged by the lack of an explicit statutory framework. Faced with the realities at hand, as a matter of necessity, court decisions are based on the specificities of the respective contributions, in a way that no codified law could reasonably be expected to create *ex ante* rules for.

[409] Marcel Duchamp fired matches dipped in paint from a toy cannon for his painting "The Large Glass."
[410] Guadamuz, supra 271, 186.
[411] Annemarie Bridy 'The Evolution of Authorship: Work Made by Code' (2016) 39 Colum JL & Arts 395, 401.
[412] Wu, supra 403, 177.

3. Economic Incentive Arguments – The Case of Investment

Generative AI models are produced almost exclusively in a corporate environment. Consequently, courts and academic commentators alike are ineluctably drawn to investment incentive rationales. Denicola and Samuelson both contend that absent a contract stipulating ownership in the outputs, copyright incentives are desirable for the user to produce new works for the public. They argue that a programmer will already be rewarded with ownership over the AI software,[413] and practically speaking, will not be directly involved in the execution, distribution, and marketing of the generative output.[414] Accordingly, on pragmatic grounds alone copyright ownership should be afforded to the user of an AI model. Rewarding those who invest in bringing a work to market – the "market-maker" as Samuelson refers to them – has always been a crucial factor within the intellectual property system, and therefore remunerating those who take a risk of investing in the creation, promotion, and sale of their works supports economic incentives for investment and copyright rationales more broadly.[415]

Market norms, at least in B2C markets, accord with this position. Based on a small sample of web-based AI art generators, users appear to be overwhelmingly at liberty to do what they want with the work they produce.[416] Out of the twenty services analysed, ten asserted no title, interest, or copyright in the outputs.[417] Four services set out in their terms and conditions that the outputs are "copyright free"[418] or are in the public domain as subject to a CC0 licence.[419] Two services[420] (one of which involves the upload of photographs)[421] explicitly state that intellectual property in the outputs reside with the user. One was entirely silent.[422] While despite claiming no copyright in the output, three of the companies analysed did seek to control what users of the services could do with the generative art they produced.[423] For example, the terms of service limit the number of copies that can be produced,[424] or require all use of works to be stopped and physical copies to be deleted after the licence expires.[425] Another limitation placed on users was the disallowal of sale or sub-licensing of the user's own outputs.[426]

[413] Supra 406.
[414] Denicola, supra 405, 283ff. Samuelson, supra 405, 1226 ff.
[415] Ibid.
[416] The terms of use were analysed in combination with personal correspondence seeking clarification for twenty services, ten of which were recommended in https://www.unite.ai/10-best-ai-art-generators/
[417] Nightcafe, DALL-E2, Deep Dream Generator, Artbreeder, DeepAI, Starry AI, Runway ML, WOMBO Dream, Craiyon, Photosonic.
[418] Images.AI.
[419] Jasper Art, Stable Diffusion, Neural Love.
[420] NightCafe, neuralstyle.art.
[421] neuralstyle.art.
[422] Big Sleep (the lack of terms and conditions for output are no doubt due to the software having to be downloaded and installed by the user from GitHub as an open source software package rather than Big Sleep being a service.)
[423] Fotor, Shutterstock.com, Creative Fabrica.
[424] Shutterstock Standard Licence. 1.1 a.i.1.
[425] Fotor Terms of Use. 4.2.2.1 (1).
[426] Creative Fabrica. https://www.creativefabrica.com/subscription-license/

Based on this survey, AI service providers in consumer markets therefore may well have concluded that in order to maximise returns and uptake of their service, end-users need to be incentivised by leaving them free to reap the rewards of their investment in generating AI CGWs. Equally plausibly, in view of the uncertain copyright status of AI-generated works, as illustrated by *Getty Images and Others v Stability AI,*[427] companies have perhaps decided to relinquish legally questionable claims of ownership in an attempt to minimise their exposure to claims of copyright infringement.[428]

Whereas well-formulated laws can offer certainty for investors, in the face of much technical complexity – and multiple pathways to the creation of an output – it is proposed that a statute-based approach to CGWs is too simplistic and imprecise. It should not be forgotten that most generative works are either provided as part of an online service regulated by terms of use, or will be the result of private ordering where the question of ownership in the outputs is negotiated and reflected in the contract. Thus, the instances where copyright law alone directly regulates ownership in an AI output appear comparatively confined – most likely limited to cases where a work is used in contravention of the licence or in the rare cases where no contractual terms are in operation.

4. Case Law

Despite the widespread scholarly commentary on authorship of computer-generated works, litigation is in fact comparatively rare. This final subsection shall outline those relevant judgements that have been heard before the courts with a particular focus on their doctrinal underpinnings.

The first jurisdiction to be examined is that of China, where the courts' application of copyright's originality test is unsettled.[429] In the cases of *Feilin v Baidu* (2019)[430] and *Tencent Shenzhen v Shanghai Yingxin* (2019),[431] the materials under dispute were generated by NLP software designed to produce written articles and reports. The plaintiffs in both cases asserted copyright infringement arising from the unauthorised making available of their works online. In the course of *Feilin*, in order to confirm the plaintiff's assertion that its report had in part been created by the company's staff, the Beijing Internet Court created parallel test reports on the same topic as the original in question involving no human intervention other than the input of related keywords as instructed by the defendant. Comparing the different works in question, the Court concluded that while copyright inhered in the plaintiff's report owing to the fact it involved original human input in its arrangement

[427] *Getty Images and Others v Stability AI* [2025] EWHC 38 (Ch).
[428] See Stability AI defence submitted to the High Court on 27/2/24. §7.4, 87.3 – 5, 100.11 (asserting that the user creates the outputs and therefore by implication liability does not sit with the model creator.)
[429] Stephen McIntyre 'Trying to Agree on Three Articles of Law: The Idea/Expression Dichotomy in Chinese Copyright Law.' (2010) 1 Cybaris Intell Prop L Rev 62, 95 (describing the application of the originality principle demonstrating a "high degree of indeterminacy.")
[430] *Beijing Feilin Law Firm v Baidu Corporation* (26/4/2019*)* (2018)京 0491 民初 239.
[431] *Shenzhen Tencent Computer System Co Ltd v Shanghai Yingxun Technology Co Ltd* (2019)粤 0305 民初 14010 号.

and textual additions, its own test reports lacked originality as no material human authorial touch was in evidence.[432] In commenting on the latter, it is interesting to observe that the Court furthered an economic incentive rationale, opining that some form of intellectual property protection for keyword-prompted CGWs was desirable in order to stimulate users to invest in the creation of new markets.

More relevant to the question of the allocation of authorship in autonomously AI-generated works is *Tencent Shenzhen v Shanghai Yingxin*. While the plaintiff's article was independently written by Tencent Robot Dreamwriter software, nevertheless the judge concluded that human involvement in the arrangement and selection of data inputs, templates, and corpus styles significantly influenced the expression contained in the work. Reflecting causational arguments, although the article in question was entirely written by the software, owing to the human intellectual labour involved in its programming, it was concluded to constitute protectable subject matter.

A more recent Chinese case (*Li v Liu*)[433]involves the unauthorised republication of an image of a young woman created using Stable Diffusion. The plaintiff generated the image by inputting over 100 prompt words and selecting various hyperparametric and technical settings, including amending an intermediate image to produce the final work. As a result, the Court concluded that the final output was not merely an "automated intellectual achievement" but reflected the "personalised expression" of the plaintiff, thus constituting an original copyright work. In its summing up, the Court can be seen to extensively rely on incentive rationales for copyright law. Opining that AI generative software is just the latest in a long line of tools that authors can use to create original works, the Beijing Internet Court set out that "creation and AI technology can only prosper by properly applying the copyright system and using the legal means to encourage more people to use the latest tools to create. In such a context, as long as the AI-generated images reflect people's original intellectual investment, they should be recognised as works and protected by copyright law."

Much like the Chinese rulings, analogous US cases are steered implicitly by a combination of causational and incentive rationales. Although unlike the Chinese rulings neither involved ML software, the rationales adopted by the courts in these recent landmark cases are of note, as like *Tencent* they address copyrightability of outputs with due consideration of the upstream software authorship involved. In both *Rearden LLC v Walt Disney* (2018) and *Design Data Corp v Unigate Enterprise Inc* (2017),[434] the courts adopted a "lion's share test" first established in *Torah Soft.*[435] As a

[432] Lee Ju Yoen 'Artificial Intelligence Cases in China: Feilin v. Baidu and Tencent Shenzhen v. Shanghai Yingxin.' (2021) 7 China and WTO Review 1 211.
[433] *Li v Liu* (2023) 京 0491 民初 11279 号.
[434] *Rearden LLC v Walt Disney Co.* No. 17-cv-04006-JST (ND Cal 21/2/ 2018). *Design Data Corporation v Unigate Enterprise Inc* 847 F 3d 1169 (9th Cir 2017).
[435] *Torah Soft Ltd v Drosnin* 136 F Supp 2d 276, 283 (SDNY 2001) (concluding that the programmer had done the lion's share of the work as the user only inputted a single word or phrase into the program to produce the matrixes in question.)

consequence, copyright ownership was conferred on the party which undertook the majority of labour and intellectual creativity involved in the production of the final generated work. In *Rearden LLC,* the judge determined that the actors and directors contracted to Disney contributed the "lion's share of the creativity" in the 3D program files outputted by Rearden's software which, when combined with Disney's own CGI technologies, created animated characters.

The facts involved in *Design Data Corp* were comparatively less complex. Citing the district court in *Atari,*[436] the Ninth Circuit rejected the notion that Design Data Corp could extend its copyright beyond the CAD systems to include the design drawings produced by third parties using its software. The Court in this landmark ruling drew a distinction between the source code which operates a software program on the one hand, and the informational outputs produced by that program on the other. Subsequent to this, the Ninth Circuit then proceeded to adopt a "lion's share" test, contending as in *Rearden LLC*, that the plaintiff had failed to provide evidence that their contribution in the design of the software comprised the majority of the labour and intellectual creativity involved in the outputting of the files and images by third parties.

As shown, the rulings in all five cases discussed above are strongly aligned with both causational and economic incentive rationales. In each case, the labour and intellectual investment made by those involved in creating the work in question was at the heart of the courts' judgements. In China, the cases are particularly notable for the courts' extensive application of doctrinal rationales to the question of computer-generated works, situating their reasoning strongly in the need to support the growth of AI markets. By contrast, the focus of the US courts appears to be on more prosaic matters. Faced with conflicting claims of authorship, in adopting a lion's share test, the courts in these two cases were less concerned with the growth of marketplaces than procedurally determining the party who has contributed most to the conceptualisation and execution of the generated output. Be that as it may, within their respective jurisdictions, these cases will undoubtedly serve as a significant jurisprudential foundation for future generative AI litigation.

Conclusion

Copyright law as a prime legal regulator of TDM, computer-generated outputs, as well as the informational matter upon which machine learning models are trained upon, will unquestionably affect incentives to invest as well as market competition. As with any technology, it inspires both awe and fear, hype and opprobrium, something that policy makers themselves are not necessarily immune from. As explained, even though UK parliamentarians were well aware that computer-generated

[436] *Atari Games Corp v Nintendo of Am* 1993 WL 214886 (1993).

lacking a “human author”[437] were no more than a figment of the imagination, this did not prevent the codification of CGW into domestic statute. From an innovation perspective, while creating a whimsical legal standard such as this stokes much doctrinal debate, in practice it provides scant support investors seeking to navigate the intricacies of AI and copyright law.

Of far more relevance to the realpolitik of data-driven innovation than computer-generated works, are the varied attempts by policy makers to introduce limitations and exceptions to support TDM and machine learning more broadly. As discussed, regulators have failed to adequately support UK and EU-based companies and researchers engaged in AI. Whereas the non-override of exceptions by contract and technological protection measures sets a fundamentally important legal precedent, beyond this jurisprudential milestone, a lack of foresight and ineffective technical execution in the law appears to cement Europe’s ongoing position as underdog in global technology markets.

[437] s.178, CDPA.

Chapter 3: The United States and Japan – Text and Data Mining

Introduction

It is evident from the previous chapter that both the UK and the European Union have struggled to adapt and enhance their copyright laws to adequately support the needs of the AI economy. This chapter will analyse the current legal response to machine learning in Japan and the United States. These two jurisdictions are analysed not only as technology leaders in their respective AI fields, but because each has adopted distinct and innovative legal approaches to addressing the challenges posed by AI, which it is contended, have so far eluded the European authorities.

This chapter will proceed as follows: first it will outline the history surrounding the introduction of a codified open norm into Japanese copyright law. After this, it will evaluate the ambit of existing US case law and the extent to which it can be applied to differing forms of machine learning. Using the same analytical framework as the previous chapter, Part 2 will examine the scope of the legal response in Japan and the US as a means of establishing how far copyright law facilitates the deployment of AI. Part 3 turns to extra-copyright issues and the degree to which contracts and technological protection measures can erode the privileges afforded data scientists by law.

Section One

1. Japan

1.1 2009: The World's First TDM Exception

Japan is unique as its copyright laws are steered at the highest level by the Intellectual Property Strategy Headquarters (IPSHQ), a body chaired by the Prime Minister, comprising all members of the Cabinet and ten experts from the private and university sectors.[438] As such, it not only encapsulates the seriousness with which Japan takes IP as a tool of its industrial strategy, but its composition reflects the complex ecology of technology collaborations that exist between the government, corporations, and Japanese research institutions.

With such an explicit political expression of IP's importance as a regulator of science and technology markets, alongside its position as one of the world's leading technological powers, it is perhaps no surprise that Japan was the first country in the world in 2009 to codify a dedicated exception for text and data mining into its copyright law:

[438] See https://www.kantei.go.jp/jp/singi/titeki2/kousei.html <Accessed 25/11/2022>

Art. 47-7: Reproduction etc for Data Analysis[439]

To the extent that it is deemed necessary, it is permissible to make a recording or an adaptation of a work (including the making of a derivative work) if the purpose of doing so is data analysis (meaning the extraction, comparison, classification, or other statistical analysis of language, sound, or image data, or anything else contained within a large number of copyright works or a large volume of data) by means of a computer. Provided however, that this does not apply to database works compiled for use by persons who carry out data analysis.

Predating by five years the “Japan Revitalisation Strategy” – a cross-sectoral government action plan on Big Data[440] – relatively little is known about the specific policy origins of Article 47-7.[441] Analysis of publicly available governmental papers nevertheless reveals that the Ministry of Economy, Trade, and Industry (METI), and the Ministry of Education, Culture, Sports, Science, and Technology (MEXT), had announced plans to undertake research into an appropriate regulatory environment for data analytics markets as early as 2007.[442] One year later, the newly formed Expert Committee on Intellectual Property Systems in the Digital and Networked Era (IPSDNE),[443] announced its intention to investigate the copyright modalities of data analytics.[444] Highlighting the tension between the need for data scientists to accumulate large amounts of data which are often subject to the proprietorial entitlements of third parties, it observed “rights clearance is in reality complex, resulting in a considerable chilling effect upon research”.[445]

1.2 2018: The Long Road to 2018 and an Open Norm

Significantly, the introduction of the 2009 TDM exception came in the midst of an intense period of discussion between academics[446] and the government[447] concerning the economic and innovational

[439] Copyright Law of Japan. Law No 73. 10/7/2009 (hereafter JCA).
[440]『日本再興戦略』改訂 2014(Japan Revitalisation Strategy 2014) Cabinet Office (24/6/2014).
[441] This observation is based on desk-based research as well as a discussion between Professors Tatsuhiro Ueno and Yoshiyuki Tamura as part of a Waseda University webinar 5/2/2022.
[442]『知的財産推進計画 2007』(Intellectual Property Promotion Plan 2007) 94 ff.
[443] Established 13/3/2008 under Article 2 of the Order of the Intellectual Property Strategy Headquarters (Cabinet Order No 45 of 2003).
[444]『デジタル・ネット時代における知財制度の在り方について <検討の視点>』(On the Ideal Form of Intellectual Property Systems in a Digital and Networked Era - Perspectives for Consideration) IPSHQ (2008).
[445] Ibid. イ, 5.
[446] Tatsuhiro Ueno,『著作権法における権利制限規定の再検討、日本版、フェアユース の可能性』(Reconsidering Copyright Limitations and Exceptions in Copyright Law and the Possibility of a Japanese-Style Fair Use) (2007) Kopiraito 47 560 2. Keiji Sugiyama,「フェアユースと教育利用」森泉章ほか編『著作権法と民法の現代的課題: 半田正夫先生古稀記念論集』(‘Fair Use and Use for Educational Purposes’ in A Moriizumi (ed), *Contemporary Issues in Copyright Law and Civil Law: Essays in Honour of Professor Masao Handa on His Seventieth Birthday*) (Hōgakushoin 2003) 293. Hidetaka Aizawa,「著作権法のパラダイムへの小論」相澤英孝ほか 編『知的財産法の理論と現代的課題: 中山信弘先生還暦記念論文集』(An Essay on the Paradigm of Copyright Law’ in E Aizawa and others (eds) *Theory of Intellectual Property Law and Contemporary Issues: Commemorative Essays Honouring Professor Nobuhiro Nakayama on his 60th Birthday*) (Kōbundō 2005) 334.
[447] For example in 2008 in its final report『デジタル・ネット時代における知財制度のあり方について<報告>』(On the Ideal Form of Intellectual Property Systems in a Digital and Networked Era) the IPSHQ stated that “it would be appropriate to introduce a general exception –“Japanese fair use” – which permits fair uses overarchingly within prescribed limits”. 11 (27/11/2008.) This approach was subsequently endorsed in January 2011 in the Report of the Copyright Subcommittee of the Council for Cultural Affairs『文化審議会著作権

benefits of introducing fair use – a debate which first started in the early 2000s.[448] A common refrain at the time was that, due to legal policy inertia, Japan had ceded its domestic search engine market to foreign technology companies. This, it was asserted, was in part due to the rigid closed list of exceptions in the Japan Copyright Act (JCA), which rendered unlawful many of the activities necessary for the operation of computers and other network technologies. One often cited repercussion of this was that a significant number of Japanese internet companies had located their servers and related operations in the United States in order to avoid claims of copyright violation.[449]

In spite of the long-standing debate around "Japan fair use", including notably a 2011 recommendation from the Council for Cultural Affairs Copyright Subcommittee to introduce an open norm,[450] the government ultimately failed to enact any open-ended exceptions as part of the major suite of copyright amendments that took place in Japan between 2009 and 2012.[451] Nonetheless, with the benefit of hindsight it can be seen that the theoretical undergirdings for the 2018 revisions were set in place around this time. Of core normative importance was the Copyright Subcommittee's 2011 identification of three categories of activity that theoretically could fall under any prospective open norm. Of particular note for this study being Category C, which formed the theoretical basis of Article 30-4 which came into effect eight years later and *inter alia* supports AI:[452]

A. **Incidental uses.** Defined as "use of a work which arises in connection with other acts not primarily intended to make use of the work and which from the perspective of societal convention can be viewed as minor in terms in terms of quality or quantity."[453]

B. **Use of a work when exploitation is lawful.** Defined as "use of a work that is reasonably necessary in the course of attempting to achieve a lawful use of the work, and from the perspective of societal convention where the use can be assessed as qualitatively or quantitatively insignificant."[454]

C. **Exploitations where the intention of enjoying the expression in a work is absent.** Defined as "exploitation which, in light of the type and usage of the work and the purpose

---

分科会 報告書』which concluded that it would be appropriate to introduce some form of open norm into Japanese copyright law "based on a sufficient consideration of Japanese society and legal frameworks." 42, 61.

[448] For an overview of the history of Japan's discussions around fair use see Tatsuhiro Ueno,「権利制限の一般規定」中山信弘・金子敏哉編『しなやかな著作権制度に向けて』('General Provisions in Limitations and Exceptions' in Nakayama N and Kaneko T (eds), *Towards a Flexible Copyright System*) (Shinzansha 2017) 141.

[449] Nobuhiro Nakayama,『著作権法改正の潮流』(Trends in Copyright Law Reform) (2009/6) Kopiraito 578, 13.

[450] Supra 447.

[451] The JCA was amended on 10/7/2009, 3/12/2010, 22/6/2012 and 27/6/2012.

[452] Supra 447, 44-50.

[453] For example, use of a painting behind a person in a photograph or film. A prescriptive exception reflecting Category A, Article 30-2 was introduced in 2012.

[454] For example, reproducing copies of a protected work for a business to decide whether to apply for permission to use publicly. A prescriptive exception reflecting Category B, Article 30-3 was introduced in 2012.

and manner of its use, *cannot* be regarded as use for the enjoyment of the expression of the work via the perception of the senses."[455]

Undaunted by the triumph of legislative copyright orthodoxy a few years prior, in 2015 the IPSHQ returned yet again to the question of general exceptions. In doing so, it called for an investigation into "an (IP) system that can cater for the needs of the modern era through limitations and exceptions with high levels of flexibility … in order to … respond to the technological and social evolution of the internet era including the emergence of new business models, artificial intelligence and 3D printing."[456]

At the policy level, the re-emergence of the open norms debate so soon after the failure of the government to follow its own advice,[457] is indubitably interwoven with the 2014 Japan Revitalisation Strategy and the 5th Science and Technology Plan. As a country that prides itself on its cutting-edge technology, underpinning these new industrial strategies were concerns that Japan's competitivity was declining, intermingled with a keen awareness of the unique combination of social pressures Japanese society faces.[458] Its rapidly aging population, a low birth rate, limited immigration, and a tectonically active landmass provoked the Japanese government into looking at AI, IoT, and robotics as the key to addressing many of its societal and economic challenges. This shift in thinking coalesced into the concept of "Society 5.0" – the fifth social revolution. As a new blue-print for a "super-smart society", it is defined as "a human-centred society that balances economic advancement with the resolution of social problems by a system that highly integrates cyberspace and physical space".[459] Central to the concept of Society 5.0 are ecosystems of big data, AI, IoT, and robotics with markets and technologies being individualised to meet a person's specific needs.

In response to refreshed calls to investigate the modalities of a more flexible copyright system, the Agency for Cultural Affairs undertook public surveys and consultations with specialists in fields as diverse as constitutional law, civil and criminal law, legal sociology, and economics.[460] Arguably most consequential was a 2017 survey seeking the view of companies on the use of copyrighted works and their expectations of the law.[461] Based on 469 responses, 68.6% of organisations stated that an exception which explicitly permitted certain uses, accompanied by a provision making similar uses

[455] The example given in the Committee Report being reproductions and the making available of copyright works while developing new technologies which led to the introduction of Article 30-4 in 2012. Of importance to this study is that ultimately this became the basis for Article 30-4 which was introduced in 2018.
[456]『知的財産推進計画 2015』(Intellectual Property Promotion Plan) (July 2015) 42.
[457] Supra 447.
[458] Nakayama, supra 449, ibid.
[459] 'Society 5.0' Cabinet Office. https://www8.cao.go.jp/cstp/english/society5_0/index.html
[460] Isao Mizuta,『著作権行政をめぐる最新の動向について』(2017/11) (The Newest Developments in Copyright Administration) Kopiraito 679 57, 2 (The Director of the Japan Copyright Office outlining the consultation process undertaken by the Agency for Cultural Affairs.)
[461]『著作権法における権利制限規定の柔軟性が及ぼす効果と影響等に関する調査研究』(Research Study on the Effects and Impacts of Flexible Copyright Exceptions) (Aoyama Shachū Co Ltd, March 2017).

also lawful, would support their business development activities. Strong support from the business community for a general flexible exception was also evident with 63.8% of respondees agreeing with the statement that case law and flexible interpretation and application of copyright law would make it easier for their business to adapt with the times.[462] At the same time, 46.9% agreed that the introduction of an open norm could lead to more litigation imposing a burden on both users and rightsholders, and 48.4% concurred with the proposition that until a court ruling confirmed the legality of a use, the ambiguity could have a chilling effect on business.

In the light of such countervailing responses, the Agency for Cultural Affairs ultimately rejected the indeterminate and open-ended nature of fair use, justifying its response in terms of the conservative nature of Japanese business. In doing so, it concluded that due to "a high level of legal compliance and a desire to avoid litigation, introducing an exception with high levels of flexibility would unlikely have much effect."[463] No doubt facing continued pressure from the IPSHQ – and reflective of the unquestionably receptive nature of the survey responses – nevertheless, in April 2017 the Copyright Subcommittee of the Council for Cultural Affairs proposed the codification of "a multi-layered system of diverse limitations and exceptions offering differing levels of both clarity and flexibility".[464] Building on the 2011 categorisation detailed above, as part of this new-found normatively ambitious approach to copyright, the Subcommittee proposed a revised three tier stratification of exceptions calibrated according to their potential impact upon rightsholders:[465]

Level 1: Acts which would not normally be viewed as prejudicial to the interests of rightsholders. (e.g. functional reproductions related to the internal workings of computers and networks, AI, etc)

Level 2: Acts where any possible harm to the interests of rightsholders would be minor in effect. (e.g. search engine snippet services, information analytics services including location based services, plagiarism detection services, etc)

Level 3: Acts that may conflict with the market for copyrighted works, but are desirable in order to realise public policies goals. (e.g. quotation, press reporting, education and library exceptions, disability exceptions, etc)

[462] Ibid. Q1-15. 20.
[463] Mizuta, supra 460, 8.
[464]『文化審議会著作権分科会 報告書』(Report of the Copyright Subcommittee of the Council for Cultural Affairs) April 2017, 68.
[465]『著作権法の一部を改正する法律 概要説明資料 平成 30 年法律第 30 号』(An Overview of the Act for Partial Amendment of the Copyright Act 2018 Law No 30) (Agency for Cultural Affairs) 6.

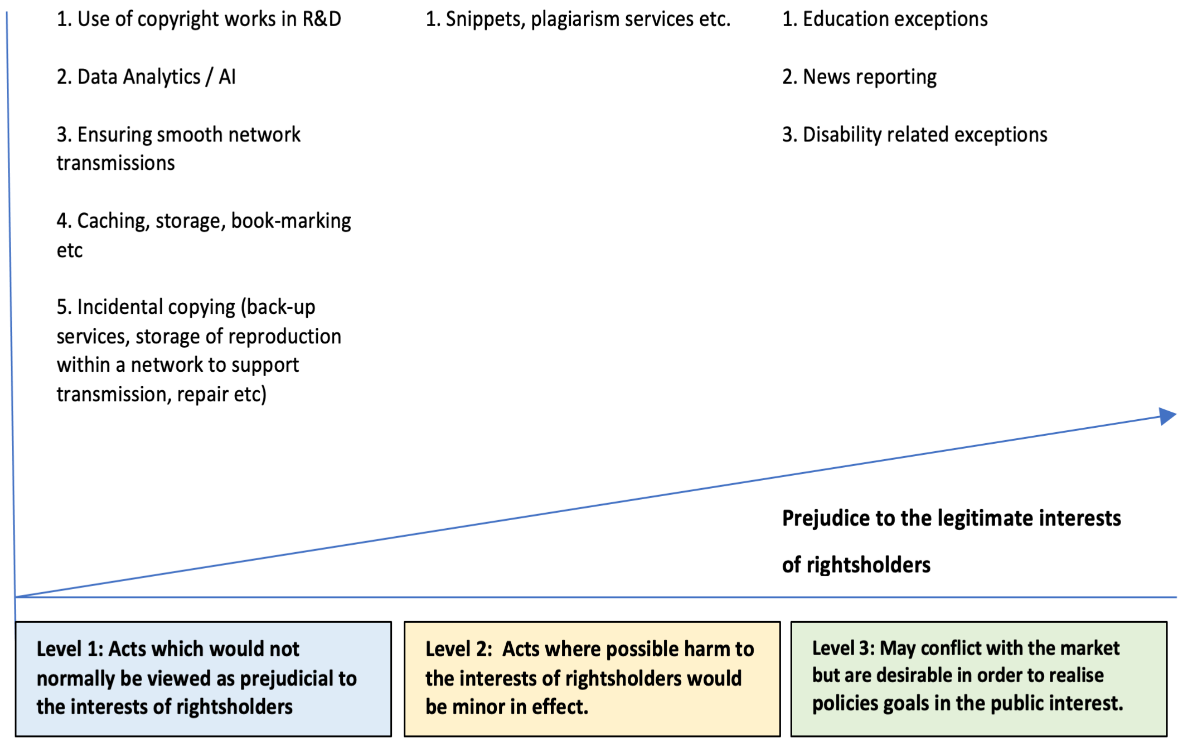

Fig. 3-1. 2018 Copyright Reform – A Recategorisation of Limitations and Exceptions [466]

In 2018, in culmination of nearly two decades of debate and numerous consultations by the government, the National Diet finally enacted three general exceptions evincing their own distinct purpose-limitations, modalities, and levels of flexibility.[467] Of main relevance to this thesis is Article 30-4 (Exploitation where the intention of enjoying the thoughts or sentiments in a work is absent)[468] which constitutes a Level 1 permitted act as designated by the government that, in principle, would not normally prejudice the legitimate interests of rightsholders:[469]

1.3 Art. 30-4: Exploitations Where the Intention of Enjoying the Thoughts or Sentiments in a Work is Absent

Art. 30-4: Provided the exploitation would not unreasonably prejudice the interests of the copyright owner in light of the manner of its exploitation, as well as the nature and use of the copyright work, it

[466] Examples sourced from Satoshi Ikemura,『柔軟な権利制限規定と実務への影響』(Flexible Limitations on Rights and Their Impact on Practice) (2018/9) Business Law Journal 126, 21.
[467] Article 30-4 (Exploitation where the intention of enjoying the thoughts or sentiments in a work is absent). Article 47-4 (Exploitation of Works by a Computer Incidental to the Exploitation of Works). Article 47-5 (Minor Exploitation by a Computer Incidental to Computerised Data Processing and the Provision of the Results Thereof). All three exceptions were passed by the Diet on 18 May 2018 and entered into force on the 1st of January 2019.
[468] For a general overview of Article 30-4 in English see Tatsuhiro Ueno, 'The Flexible Copyright Exception for 'Non-Enjoyment' Purposes – Recent Amendment in Japan and Its Implication' (2021) 70 GRUR International 2 145.
[469] It is noteworthy that Article 30-4's formulation is identical to the "Category C" exception as defined by the Council for Cultural Affairs Subcommittee in its January 2011 Report, supra 447, 48.

is permissible to exploit a copyright work to the extent deemed necessary irrespective of the means in any of the following cases, as well as any other cases where such exploitation is not for enjoying, or allowing another person to enjoy, the thoughts or sentiments in a work.

(i) Use in testing for the development, or practical application of technologies concerning audio or audiovisual recordings including other such exploitations of copyright works;

(ii) Use in data analysis (meaning the extraction, comparison, classification, or other analysis of information including but not limited to language, sound and image data, from a large number of copyright works or a large volume of data; the same shall apply to Article 47-5, paragraph (1), sub-section (ii));

(iii) in addition to the uses outlined in the preceding two sub-sections, exploitation of works in the course of data processing by a computer, or any other form of exploitation, that does not involve perception of the expressions in the copyright work by the human senses (with the exception of exploitation of computer programs by a computer which shall be excluded.)

Before proceeding to an evaluation of Article 30-4's doctrinal underpinnings, first an analysis of its scope and structure would prove instructive. Its most striking feature is that that whilst still providing for a dynamic interpretation due to its open-ended formulation, unlike fair use – which accommodates no purpose limitation at all – Article 30-4 circumscribes user privileges to those acts which do not invoke sensual or intellectual enjoyment on the part of the beneficiary. Accordingly, it excuses two categories of application: i) those which do not involve human perception, and ii) those acts which involve human perception but are absent any aim to enjoy the original expression contained in the work.

The enumerated non-exhaustive examples contained in Article 30-4 traverse both categories:data analytics (Article 30-4 ii),[470] "back end" computer network processing activities absent of human perceptions (Article 30-4 iii), as well as R&D that involves the use of copyrighted works for testing and evaluation purposes (Article 30-4 i).[471]

Less obvious permitted acts, in keeping with Article 30-4's intent, can also be imagined. For example, parody could be justified by the fact a sufficiently reformulated work is the object of the enjoyment and not the original work upon which it is based.[472] Prior to a public performance, where enjoyment

[470] Essentially replacing and augmenting the data analytics exception (Article 47-7) introduced into the JCA in 2009.
[471] A modified form of the rescinded 2012 testing and development exception (Article 30-4).
[472] See Nobuhiro Nakayama,『著作権法第 3 版』(Copyright Law 3rd Edition) (Yūhikaku 2020) 381-2. Moriyuki Kato, Tatsuhiro Ueno, Kōji Okumura,『柔軟な権利制限規定による著作物の利用拡大と,これからの課題 上』(Expansion of the Use of Copyrighted Works through

by a third party would be triggered, private music practice by an orchestra may similarly be excused as the intent of the musicians is honing their performance skills and not the enjoyment of the music by a third party.[473]

By setting forth a purpose limitation in its opening paragraph, complemented by a list of non-exhaustive examples, Article 30-4 thus seeks to balance and afford users of the exception with both flexibility and legal clarity.[474] As explained by the Director of the Copyright Office, the intention of the government was to maximise use of copyright exceptions while achieving an effective level of flexibility, something that the introduction of an unbridled, entirely open-ended norm like fair use was judged not to provide owing to the perceived risk averse nature of Japanese business culture.[475]

Turning to the structure of Article 30-4, it sets forth a preamble outlining the purpose limitation of the exception side-by-side with three factors to be assessed in determining if the usage of a copyright work is lawful. Namely:

i) the manner of exploitation;
ii) nature of the work; and
iii) usage of the work.

In addition to these three factors, the preamble also simultaneously establishes that the exception is "irrespective of the means" undertaken, thus explicitly excusing the reproduction, communication to the public, and other restricted acts reserved by statute. As to the protection of rightsholders, distinct from fair use and arguably redundant due to the operativity of the Berne Convention, internal to the exception itself, explicit reference is provided for in wording derived from the third step of the three-step test. Overarching the three assessment factors contained in Article 30-4, the preamble sets forth that no use of a work may "unreasonably prejudice the interests of the copyright owner", thus delineating the outer boundaries of permissible uses and the ultimate test against which claims of putative infringement will be based.

Although structurally, and in terms of its formulation, Article 30-4 diverges significantly from fair use, commonalities emerge on closer observation. Most obviously, the exception's three assessment tests can be seen to overlap broadly with the four fair use factors. The "nature of the work" appears indistinguishable from the second fair use factor. Thus, the Japanese courts may give greater leeway

---

Flexible Limitation Provisions and Future Challenges (Part I)) (2019/4/1) New Business Law, 12, 13 (discussing a number of examples of potentially non-infringing acts).
473 Nakayama id. 83. Kato et al, ibid.
474 Moriyuki Kato,『著作権逐条講義（七訂新版）』(An Article by Article Commentary on Japanese Copyright Law) (7th Edition, Copyright Research Information Centre 2021) 279.
475 Mizuta, supra 463, ibid.

to the use of informational goods than works of entertainment or as discussed later in the chapter, AI custom-made databases.[476] Although the exact distinction between the "usage of the work" and "the manner of exploitation" is uncertain, they can be seen to align with the first and fourth fair use factors respectively. Moreover, as with fair use, the syntax and formulation of the Japanese text clearly reveals that all three factors of Article 30-4 are to be evaluated in turn. In the absence of case law, at this point in time, it is too early to say whether each prong will be given equal weight or, as in the US, certain factors will be afforded comparative deterministic weight and prominence.

Taxonomically, according to Flynn and Palmedo, open norms display "generality" when they apply "a single test to a group of uses or purposes".[477] Thus, while Article 30-4 displays both generality and openness (defined as a typology of exception which lacks a "defined (aka closed), list of purposes, uses, works or users"),[478] as a consequence of its purpose limitation, it nevertheless has a more restricted mode of application compared to fair use which is characterised by an entirely open-ended scope of application that theoretically may apply to any fair utilisation of a copyrighted work. To differentiate fair use from the more restricted and purpose-delineated nature of flexible exceptions found in Japanese law, such as Article 30-4,[479] this author has therefore proposed that L&Es that display generality subject to a purpose limitation be referred to as *qualified* general exceptions.[480]

Moving from structural characteristics to Article 30-4's doctrinal underpinnings, as is well known, copyright is designed to protect the original authorial expression which inheres in informational and cultural goods. Thus, historically, leaving the emotional response it engenders, as well as other non-reproductive acts beyond the control of the author.[481] Accordingly, as outlined by the Japanese Copyright Subcommittee in its rationale for Article 30-4, it can be inferred that the grant of copyright is intended to safeguard the rights of authors in instances specifically limited to those where original expression engenders human emotion.[482]

Whereas conceptually and historically the enjoyment of a work has lain outside the boundaries of privileges afforded to authors, as has been extensively covered in the literature, with the advent of digital technologies copyright law has indelibly intertwined the ability to enjoy a work with its copying. Automated mechanical reproductions constitute an unavoidable preliminary step before accessing and using a digital work, such that enjoyment is no longer independent of, or divisible from,

[476] Infra subsection 3.2.
[477] Sean Flynn and Michael Palmedo, 'The User Rights Database: Measuring the Impact of Copyright Balance' (2019) Working Papers 42, 8.
[478] Ibid.
[479] See also Articles 47-4 and 47-5 of the JCA which demonstrates generality within the bounds of a purpose limitation.
[480] Mendis, White and Hong, supra 353, 70.
[481] For example, see Mark Lemley and Bryan Casey, 'Fair Learning' (2021) 99 Tex L Rev 743 (situating unregulated acts such as reading, listening and exhaustion in the idea-expression dichotomy as important unregulated legal incentives to allow others to create further works.)
[482] 『文化審議会著作権 分科会報告書 平成 29 年 4 月』(Report of the Copyright Subcommittee of the Council for Cultural Affairs April 2017) 41.

the restricted act of reproduction. Activities such as reading and listening, or extracting facts from a book with a paper and pen, that hitherto were unregulated in the analogue environment, are in the digital era now constrained by copyright where they are undertaken with the aid of a "copy-reliant technology".[483] In allowing copyright to expand beyond its original goals and become a general regulator of technology markets, it now governs many aspects of our lives for which it was never designed and for which it is in many senses manifestly ill-suited.

Faced with the chilling effect copyright reportedly had on Japan's internet economy and wishing not to repeat the same mistakes again, the Japanese government in doctrinal terms has sought to ameliorate the harmful effects of copyright law on technology markets by returning to first principles. Relying on copyright jurisprudence in its rationale for the enactment of Article 30-4, the Agency for Cultural Affairs sets out that "[t]he economic value of a work is predicated on payment by a person who wishes to satisfy his or her intellectual and spiritual needs by enjoying the thoughts or sentiments expressed therein. Thus, it can be assumed that acts that do not trade on the enjoyment of the thoughts or feelings expressed in a work do not impair the opportunity for rightsholders to receive remuneration from those who wish to enjoy the thoughts or feelings expressed in a work."[484] Contrary to Japan's civil law traditions that have commonly found justifications for copyright in natural rights theories, only those reproductions that impede the economic and commercial purpose for which copyright exists in the first place shall be protected.[485] As a result, post the reforms of 2018, only those copies which "satisfy an intellectual or emotional demand through their enjoyment"[486] constitute protectable subject matter.

As spelled out by the Japanese government, the doctrinal justification for Article 30-4 finds strong support in the legal literature. In particular, that of a number of American scholars who have argued that "non-expressive" and "non-consumptive" uses, including but not limited to TDM, are paradigmatic fair uses.[487] As defined by Sag, these terms refer to acts of reproduction "not intended to enable human enjoyment, appreciation, or comprehension of the copied expression as expression."[488] Echoing in many ways the position set out by the Japanese Copyright Office, Sag argues that many non-expressive computational uses are entirely consistent with the fundamental goals of the idea-

[483] Matthew Sag, 'Copyright and Copy-Reliant Technology' (2009) 103 Nw U L Rev 1607 (arguing for the systematic application of fair use to uses where expressive works are copied for non-expressive purposes by copy-reliant technologies.)
[484] 『平成 30 年著作権法改正の概要』(2018) L&T 81 47 (explaining the rationale for the introduction of the exceptions and the doctrinal theory underpinning copyright.) See also testimony of Tsukasa Nakaoka Deputy Director of the Agency for Cultural Affairs to the Education and Science Committee of the Upper House. 17/5/2018.
[485] See William Fisher, 'Theories of Intellectual Property Rights' in Stephen Munzer (ed), *New Essays in the Legal and Political Theory of Property* (CUP 2001) (outlining four theories of IP including "natural property rights" that find their justifications in the writings of John Locke, and similar natural law theories revolving around human personhood arguments derived loosely from Kant and Hegel".)
[486] Mizuta, supra 460, 41.
[487] Lemley and Casey, supra 481. James Grimmelmann, 'Copyright for Literate Robots' (2016) 101 IOWA L Rev 657. Sag, supra 483. Michael Carroll, 'Copyright and the Progress of Science: Why Text and Data Mining Is Lawful' (2019) 53 UC Davis L Rev 893.
[488] Matthew Sag, 'The New Legal Landscape for Text Mining and Machine Learning' (2019) 66 J Copyright Soc'y USA 291, 301.

expression dichotomy as they do not communicate or hinder “human appreciation of the expressive qualities of [the original copyright work.]”[489]

By turning copyright’s application away from mere procedural reproductions and communications of authorial expression, to focus on acts which require human enjoyment, the Japanese authorities have explicitly relied on the idea-expression dichotomy to minimise the economic and social harms caused by copyright’s over-extension in the digital realm. With a clear eye on its industrial strategy, Article 30-4 creates a breathing space where Japanese companies and researchers can operate and innovate free to consume copyright works on the proviso that authorisation is only required where human beings derive sensual or intellectual pleasure from the use of a work. As a consequence, automatic and mechanical acts of reproduction of a functional or operational nature that are essential for the use of computers and other network technologies are rendered beyond the monopolistic control of rightsholders. Further, as the Japanese legislature has defined such acts as being beyond the purpose for which copyright exists as a market intervention, not only copying, but the communication of copyrighted works are afforded a non-infringing status, on the proviso that the user does not derive any enjoyment from the author’s original expression.

Thus, rather than rely on external legal frameworks – such as fundamental rights in the European Union or constitutional law in the United States – the Japanese legislature has sought to define the role and function of copyright endogenous to the law itself. By statutorily determining the telos of copyright to be the enjoyment of authorial expression, the government has within the limits set by international treaties significantly recalibrated the spheres of influence of both public and private rights. Since the early 1990s, the technology-induced extension of copyright has created negative externalities for researchers and businesses alike, which by virtue of Article 30-4, the Japanese government has sought to ameliorate. In doing so, not only will Japanese copyright law potentially bolster data-driven innovation but from a jurisprudential perspective it may have the positive and enriching effect of re-focusing copyright back to first principles.[490] Whereas before Japanese courts have lacked the clear guiding vision of purpose for copyright that, for example, the US constitution provides, Article 30-4 now furnishes the authorities with the necessary tools to develop new conceptual frameworks to balance the legitimate needs of rightsholders with those of users of copyright works, to the ultimate benefit of the public interest.

Section Two

[489] Ibid.

[490] Roya Ghafele and Benjamin Gibert, ‘A Counterfactual Impact Analysis of Fair Use Policy on Copyright Related Industries in Singapore.’ (2014) 3 Laws 2 327. (finding a positive correlation between the introduction of flexible copyright norms and the faster growth of information technology businesses). See also Mendis et al, supra 353.

2. United States

2.1 Fair Use and Landmark TDM Judgements

Owing to their traditionally rigid copyright frameworks, while Japan and Europe have had to introduce new limitations and exceptions to support AI, the US – as so many times before –has relied on the dynamic judge-led doctrine of fair use.[491] Enacted in 1976 and codifying well over a 100 years of case law, §107 of the US Copyright Act provides a balancing test of a minimum of four factors to determine if an unauthorised use of a copyright work is infringing. Described by Goldstein as "the great white whale of American copyright law", it is hard to overstate its importance to US copyright jurisprudence.[492] Frequently providing the heavy-lifting for the US tech industry, fair use rulings have rendered lawful many different forms of technological reproduction that it is no exaggeration to say form the life-blood of the digital age. These include: the systematic processing of content by search engines,[493] caching,[494] reverse engineering,[495] and even extending to the legitimisation of plagiarism software.[496] Importantly, there is also general agreement amongst scholars that US case law can be relied on to support certain forms of machine learning – most obviously those text and data mining techniques whose goal is the identification, extraction, and sharing of data derived or extracted from the corpus upon which the AI model has trained and / or analysed. Moreover, a small but growing body of case law indicates that generative AI models also constitute fair use.[497]

Before proceeding any further, however, as a starting point it may prove useful to outline some of the foremost TDM / data analytics case law in order to contextualise and frame the discussions and conclusions that follow. Additionally, it is worth underlining that these cases focused on the application of algorithms to analyse and create new insights. As addressed in greater detail later in the chapter, this stands in contrast to much of the current litigation, which centres on generative AI:

Authors Guild Inc v Google[498]

In December 2003, in collaboration with publishers, the service Google Print was launched providing online access to small textual snippets derived from books that Google had licensed access to.[499] In

---

[491] Copyright Law of the United States (Title 17) §107.
[492] Paul Goldstein, 'Fair Use in Context' (2007) Colum JL & Arts 31 433.
[493] *Kelly v Arriba Soft Corp,* 336 F 3d 811 (9th Cir 2003) (finding a search engine's copying of images, creation and provision of a thumb nail was lawful). *Perfect 10 Inc v Amazon com Inc,* 508 F 3d 1146 (9th Cir 2007) (ruling as in *Arriba Soft* that the reproduction and making available of thumbnails is transformative).
[494] *Field v Google Inc*, 412 F Supp 2d 1106 (D Nev 2006) (finding caching of links by Google were transformative acts under fair use).
[495] *Sega Enterprises Ltd v Accolade Inc,* 977 F 2d 1510 (9th Cir 1992), *Sony Computer Entertainment v Connectix Corporation*, 203 F 3d 596 (9th Cir 2000).
[496] *A V ex rel Vanderhye v iParadigms, LLC* 62 F 3d 630 (4th Cir 2009) (concluding that copying of students' papers for detecting plagiarism using machine learning software was transformative.)
[497] For example, *Kadrey v Meta*, *Bartz v Anthropic*.
[498] *Authors Guild Inc v Google Inc,* 804 F 3d 202 (2d Cir 2015).
[499] https://www.searchenginewatch.com/2003/12/17/google-introduces-book-searches/

December 2004, Google further announced the Google Library Project and its intention to digitise books from a handful of large research libraries in North America[500] and the University of Oxford. The original press release declared that the library partnerships would allow users "to search across library collections including out of print books and titles that weren't previously available anywhere but on a library shelf". [501] In terms of the legal basis for the service, no more was said than the "[l]ibrary content will be displayed in keeping with copyright law". This ambiguous phrasing, in fact, referred to Google's well-laid plans to digitise books and maps based on a fair use defence.

Within a year of the launch of the Library Project, five US publishers[502] in collaboration with the Authors Guild commenced proceedings against Google for presumptive violation of copyright law.[503] Subsequent to this, it took one rejected settlement agreement[504] and the US Supreme Court rejecting a writ of certiorari from the Author's Guild, to definitively hold that Google's book digitisation, supply of surrogates to partner libraries, and making available of snippets in response to a search query was fair use.[505] Importantly, in ruling that the digitisation and reproduction of copyright protected works for the purposes of information extraction (and by extrapolation computational data analysis more broadly) was fair use, the courts confirmed as they have done many times before in technology-related cases that Google's actions were transformative.[506]

<u>Authors Guild Inc v HathiTrust</u>[507]

Although TDM technologies were not explicitly discussed by the court, this second precedential case, also brought by the Authors Guild, concerns the HathiTrust Digital Library (HDL). Established in October 2008 by thirteen US universities, HDL aimed to archive and give access to the digital book surrogates scanned by Google and sourced from the university libraries of Cornell, Michigan, Indiana, and Berkeley. One strand of the *HathiTrust* case concerned an online service which returns the results of a search on a keyword and supplies the term's frequency across the title, including the page numbers upon which it appears. Unlike Google Books, no snippets of text are displayed unless the work is in the public domain or permission has been granted by the copyright owner. Following in the footsteps of other technology cases,[508] as well as the prior ruling by the District Court in *Authors*

---

[500] Harvard, Stanford, Michigan and the New York Public Library.
[501] http://googlepress.blogspot.com/2004/12/google-checks-out-library-books.html (6/12/2004).
[502] McGraw-Hill, Pearson, Penguin, Simon & Schuster, and John Wiley & Sons.
[503] *Authors Guild v Google Inc,* No 05 Civ 8136 (SDNY 20/9/2005). *McGraw-Hill Cos Inc v Google Inc,* No 05 Civ 8881 (SDNY 19/10/2005).
[504] *Authors Guild v Google Inc,* No 05 Civ 8136 (DC) 770 F Supp 2d 666 (SDNY 22/3/2011).
[505] *Authors Guild Inc v Google Inc,* 804 F 3d 202 (2d Cir 2015).
[506] Ibid. 225.
[507] *Authors Guild Inc v HathiTrust,* 755 F 3d 87 (2d Cir 2014).
[508] Supra 493-494.

Guild v Google,[509] the appeals court concluded that HDL's "creation of a full-text searchable database is a quintessentially transformative use"[510] and as a consequence Hathi's actions to be fair.

<u>A.V. ex rel Vanderhye v iParadigms</u>
In many ways, displaying more sophisticated forms of data analysis than either *Google Books* or *Hathi Trust*, the *iParadigms* case involved the commercial plagiarism detection service Turnitin. The plaintiffs alleged that absent permission the archiving of their essays for comparison with other works as part of Turnitin's licensed service violated their copyright. Nevertheless, in line with a long line of non-expressive use cases,[511] including the two precedential TDM cases mentioned above, the US Court of Appeals for the Fourth Circuit ruled that iParadigm's use of the students' essays to detect plagiarism was highly transformative and as a consequence constituted a fair use.[512]

2.2 Transformative Use – *Primus Inter Pares*

As illustrated, the reliance by the US courts on transformative use in these three landmark US cases forms the core doctrinal justification for TDM, and by implication machine learning more broadly being admissible as a fair use. Before analysing in subsections 2 and 3 the extent to which §107 supports AI innovation, it is nevertheless apposite to touch upon how transformative use as a judicial doctrine arose, and the general rationale for its application to TDM cases. Building on this, the analysis also considers the extent to which existing case law, in conjunction with well-established US copyright doctrines, might support a finding of fair use not only for TDM but also for generative AI technologies.

As many commentators[513] and indeed the courts have observed,[514] transformative use as a deterministic tool in establishing fair use has assumed the status of *primus inter pares*. Critically, its presence is often not only decisive in assessing the purpose and character of the use,[515] but as all four factors are to be weighed hand-in-hand, it permeates the evaluation of the remaining three factors also. Lai's impressive study of all cases citing transformative use between 1991 and 2016 shows the decisive nature of transformative ruling on a fair use defence. It demonstrates empirically that a finding of transformative use overrides market harm considerations under the first and fourth factors,

---

509 *Authors Guild Inc v Google Inc,* 954 F Supp 2d 282 (SDNY 2013) 291.
510 Supra 507, 97.
511 Supra 493-494.
512 Supra 496, 639-40.
513 Sag, supra 488, 315. C Asay et al, 'Is transformative use eating the world?' (2020) BCL Rev 61 905. §13.05[B][6] Melville B Nimmer and David Nimmer, *Nimmer on Copyright* (Matthew Bender 2019). Neil Netanel 'Making sense of fair use' (2011) Lewis & Clark Law Review 15 715.
514 See for example *A V ex rel Vanderhye v iParadigms,* 638 quoting *Campbell v Acuff-Rose,* 510 US 569 (1994) 579.
515 §107.1.

renders the type of work irrelevant (the second factor), and in terms of the amount and substantiality of the portion to be used (third factor) will routinely permit the copying of an entire work.[516]

Originating from a landmark ruling concerning a rap parody of Ray Orbison's "Pretty Woman", the US Supreme Court has defined a transformative use as one which adds "something new, with a further purpose or different character, altering the first [work] with new expression, meaning, or message".[517] Thus, where the original work is added to, either because it has been altered in some substantial way or where the purpose to which it is put is different to that of original work, the authorities will frequently utilise the transformative use test. For this study, of crucial importance is that the transformative use doctrine has been applied to safeguard a wide array of technological uses from proprietary control. In common with the *Google Book* cases and *iParadigms*, other important examples where copy-reliant technologies have been found to be transformative include *Field v Google*[518](caching and indexing by search engines), *Kelly*[519] (use of thumbnail images), and *Perfect 10*[520] (inter alia hyperlinking).

2.2.1 Are all forms of AI transformative?

In the light of the *Google Books* cases and *iParadigms,* it seems relatively settled as a matter of law that the training of machine learning models on protected subject matter, analysis, and the sharing of limited amounts of data derived from a corpus of copyright works is fair use. While not a term of art generally used by the courts, these types of models are quintessential "non-expressive' or "non-consumptive" uses.[521] As defined by Sag, these terms refer to "any act of reproduction that is not intended to enable human enjoyment, appreciation, or comprehension of the copied expression as expression."[522] As further elaborated upon in the following section, the three landmark TDM cases illustrate that the transformative nature and different purposes served by an AI undertaking data analysis to that of the original work is central to a determination of fair use. Thus, applications such as the use of a digitised book to determine its relevance to a person's research, or uses of essays to detect plagiarism, may be excused as fair.

516 Jiarui Lai 'An Empirical Study of Transformative Use in Copyright Law' (2019) 22 Stanford Technology Law Review 163.
517 *Campbell v Acuff-Rose Music Inc* (1994) 579.
518 Supra 493.
519 Supra 493.
520 Supra 494.
521 While the term non-expressive use predates discussions relating to TDM, the term non-consumptive use began to gain currency subsequent to the Amended Settlement Agreement that was ultimately rejected by the courts in *Authors Guild v Google Inc* (22/3/2011).
522 Sag, supra 488, 301.

Nevertheless, whether other forms of AI whose prime purpose is not data analysis can be excused in the same way is less certain.[523] In particular, those generative techniques which produce works entirely dependent on creative expression, such as generative art, songs, text, etc.[524] Broadly speaking, as models capable of producing outputs similar to the data they were trained on, they have the ability to supplant the works of original creators, and thus a transformative use defence may not succeed.[525] This brings into question whether or in what circumstance such technologies are able to enjoy a fair use defence.

In the mainstream literature, although the outputs of generative models are referred to as generative works, this study shall sometimes use the term "media AI works". This is for two reasons: firstly, because the technical literature defines a vast array of probabilistic models as generative that do not produce works of an artistic or aesthetic nature.[526] Secondly, nearly all AI models generate some form of informational output, and therefore to avoid confusion between generic AI and those models that produce outputs of a predominantly creative and artistic nature, more precise language is desirable.

Importantly, at the time of writing, none of the foremost precedential US judgements to date have dealt directly with media AI works and therefore it cannot be said with any exactitude whether a fair use defence is likely to succeed. Before examining the likelihood of media AI works being rendered fair, and in what circumstances, it is posited that a more in-depth evaluation of the *Google Books* cases and *iParadigms* will first prove instructive.

2.3 Different Purpose and Character - First Fair Use Factor

At the heart of judgements relating to transformative use lies the question of purpose, and whether in the words of the Supreme Court the use of a copyrighted work supplants or "supersedes the objects [or purposes] of the original creation".[527] Borrowing from the work of the renowned judge Pierre Leval, in cases where a transformative use has been found, courts have often defined a transformative use as one which "employ[s] the quoted matter in a *different manner* or for a *different purpose* from the original."[528] Thus, a usage divergent in character and intent to that of the original work is required.

---

[523] For example, see Lemley & Casey, supra 481. Benjamin Sobel, 'Artificial intelligence's fair use crisis' (2017) 41 Colum JL & Arts 45. Matthew Sag, 'Copyright safety for generative AI' (2023) 61 Hous L Rev 295.
[524] Examples of models which create generative AI works include GANs which undertake style transfer on images, LSTM and RNN models which predict the next word or note in a sequence, and MuseGAN generators which can generate music.
[525] See *Thomson Reuters Enterprise Centre GmbH v ROSS Intelligence In*c (1:20-cv-00613) District Court Delaware (a partial summary judgement ruling that a generative AI that utilised Westlaw headnotes, annotations etc in its training and which produced relevant court rulings in response to a prompt, was not transformative.)
[526] David Foster, *Generative Deep Learning* (O'Reilly 2019) 1.
[527] *Authors Guild v HathiTrust*, 97 (quoting from *Campbell* at 579).
[528] Pierre N Leval, 'Toward a Fair Use Standard' (1990) 103 Harv L Rev 1105, 1111 (Emphasis added).

In justifying its finding of fair use in *Hathi Trust*, the Second Circuit drew attention to the different purpose of Hathi Trust's search tool. In opining "there is no evidence that Authors write with the purpose of enabling text searches",[529] it defined the purpose for writing in a manner that excludes not only search, but by implication the many other AI outputs and actions that do not substitute the purpose for which the original work was created in the first place. In similar vein, Judge Chin in *Authors Guild v Google,* ruled "Google Books does not supersede or supplant books because it is not a tool to be used to read books."[530] Thus, in contrast to works created by their authors for them to be intellectually or aesthetically enjoyed by a human audience, applied computational uses such as model training, knowledge acquisition, classification, prediction are likely in many cases to constitute a fair use under US law as they display a different purpose to that for which the work which was originally created. In the same way, in *iParadigms* the court found the use of students' essays for plagiarism detection not only served a different purpose to the original, but created a public good through its wide adoption in academia, thus weighing in favour of a positive finding under the first factor.[531]

2.4 The Second and Third Fair Use Factors

According to Patry, the second factor (the nature of the copyrighted works), has seldom played a significant role in determining fair use.[532] In both *Google Books* and *HathiTrust,* the courts found that the use of both fiction and non-fiction titles was of no consequence as search did not "provide a meaningful substitute for the original".[533] Thus, in effect abrogating the second factor to the primacy of a transformative use finding. Similarly, in *iParadigms*, the transformative use – namely detecting plagiarism – was the determining factor for the court, and not the nature of the students' work being digitised.

In terms of the third factor, when it comes to computational uses the amount and substantiality of the work copied is evidently less consequential than whether it is indispensable and proportionate to the purpose.[534] Thus, it can be seen that the third factor does not operate in isolation, with an evaluation being undertaken in the light of the purported justification under the first factor.[535] Courts have across different technologies consistently held that the wholesale coping of the work can be fair.[536] Not only in the landmark TDM cases, but in other paradigmatic non-expressive use cases such as *Kelly* and *Field v Google,* the courts have frequently found that an entire copy of the original work made and

529 *Authors Guild, Inc v HathiTrust***,** 97.
530 *Authors Guild v Google,* 954 F Supp 2d 282 (SDNY 2013) 291.
531 *A V ex rel Vanderhye v iParadigms,* 638.
532 William F Patry, *Patry on Fair Use* (Thomson Reuters 2015) §4.1. See also Lai, supra 516.
533 *Authors Guild v Google* (2015) 220-1.
534 *Campbell v Acuff-Rose Music Inc,* 510 US 569 (1994) 586.
535 Leval, supra 528, 1123.
536 Amanda Levendowski, 'How copyright law can fix artificial intelligence's implicit bias problem' (2018) 93 Wash L Rev 579 627.

stored on the defendants' servers was a justifiable fair use. The rationale of the court being that wholesale copying was indispensable to the purpose, namely, a transformative act fundamental to the workings of the internet.[537]

2.5 Market Harm, Substitutability and the Fourth Fair Use Factor

As shown by Lai, courts have frequently determined that the more transformative the use, the less significant other considerations become.[538] In terms of a test for market harm, therefore, the less a work serves the same purpose and use as the original, the less likely economic damage to authorial interests will arise owing to its inherently non-competitive nature. In terms of the core question: how to measure economic impact on rightsholders, significantly the Second Circuit in *Campbell* has stated that the sole benchmark for evaluation is whether the secondary use acts as substitute for the original work.[539] Thus, as many types of AI output "do not affect the market for the original",[540] namely a market where works are to be viewed, heard, or read by people, it can be concluded instances of unreasonable market harm are likely to be comparatively few.

A further important criterion of market harm is whether the copied works are made accessible to the public. As Grimmelman has argued, it appears largely settled as a matter of law that automated so-called "non-display uses", where no access by the pubic occurs, are excused as fair as they create no cognisable market harms.[541] In *HathiTrust,* while the entire digital corpus was made available on highly restrictive terms to the universities' computer engineers,[542] no copyrightable materials appeared in the search results made available to the public. Similarly, the exclusive right of making available was not a feature of *iParadigms* as although student essays and journal articles were copied by the software, no public access to those works from the Turnitin systems are possible.

By contrast, Google Books provides verbatim copies of short extracts from the original to the American public in response to a query. All the same claims of market harm from the Authors Guild were dismissed by the Second Circuit, as the snippets were not only limited in length, but it was not possible to recreate the original book through multiple searches due to the functional restrictions of the service. In conclusion, Judge Leval opined Google's use of verbatim snippets was non-infringing as no "meaningful substitute" of the expressional content of the work was made.[543] That is to say, the usage did not supplant the reading of the book by the user because the use of the snippets was for a

[537] See Grimmelmann, supra 487, 658 (arguing that "robotic readers get a free pass" under copyright law to copy entire works.)
[538] See Lai, supra 516.
[539] *HathiTrust* at 99, quoting *Campbell* at 591.
[540] *Campbell*. Ibid.
[541] Grimmelmann, supra 487.
[542] *HathiTrust* at 100. The Court found no issue with Google's communication of the works to the HathiTrust and the subsequent secure storage by IT staff.
[543] 225, *Authors Guild v Google* (2015).

different purpose – namely "to help the searcher learn whether the book's use of that term will be of interest to her".[544]

2.6 Is the Creation of a Media AI Work a Fair Use?

Having outlined transformative use as the primary fair use rationale for TDM and other analogous automated knowledge discovery tools, this subpart will turn its attention to media AI works and whether, or the extent to which, existing fair use case law and jurisprudence could be applied to certain categories of generative AI works. As a usually original expressive output derived from a non-expressive use, the thorny issues presented by generative AI are, at the time of writing, subject to ongoing litigation. Thus, while a number of partial summary judgements in the lower courts have been handed down, it is too early to say a definitive line of legal interpretation has emerged.[545]

Accordingly, all arguments advanced below lie between the partially substantiated and logical conjecture. For the purpose of analysis, media AI works may be divided into two distinct categories: the first being those where the intention or unintended result is to mimic the work of a particular author and to compete in the same marketplace (mimetic AI works). [546] The second category are non-mimetic works which, despite being trained on copyright works, and while perhaps displaying similarities at the non-protected ideational level, are from a copyright perspective distinct and different from the works they have trained on in their precise expression.[547] To put another way, they are expressionally original. As discussed in the previous chapter, displaying similarity of expression to the works upon which they are based and therefore, inverting the legal nomenclature, the first category are both expressional and consumptive uses. As a consequence, it seems uncontroversial to conclude that any claims of infringement can be addressed within the scope of existing copyright frameworks as a case of substantial similarity.

Mimetic AI works, however, are the exception and not the rule. The majority of AI media works will, in the natural language sense of the word, be original as they lack duplicative expression substantially similar to that of another work.[548] As with any human author, although the models have learnt from observing the works of others by extracting latent features and associations within the data, the

[544] Id. 227.
[545] For example, *Getty Images (US) Inc v Stability AI Inc* (1:23-cv-00135) District Court, Delaware (Feb 2023). *Andersen et al v Stability AI et al* (3:23-cv-00201) ND California (Jan 2023). *J Doe 1 et al v GitHub, Microsoft Corp, OpenAI et al* (4:2022cv06823) ND California. (Nov 2022), *New York Times v Microsoft et al*, No. 1:23-cv-11195 (SDNY).
[546] For example, by over training the model which in the case of generative AI can lead to overfitting, also known as memorisation.
[547] See Sag, supra 523 (asserting that the fact AI models may produce expressive works "does not necessarily negate the application of the non-expressive use principle" because what is deterministic is whether "the original expression of the authors of works in the training data is communicated to a new public.") 308 ff.
[548] It is interesting to observe that in *J Doe 1 v GitHub et al* the plaintiff's refer to a GitHub report where only 1% of softwares produced by Codex and Copilot of over 150 characters in length resembled software it had been trained on. See also n.563 (showing how in *Kadrey et al v Meta Platforms In*c, *Bartz v Anthropic, Universal Music Group et al v Suno Inc* and *Mike Huckabee v Meta Platforms Inc etc* the plaintiffs did not contest that the AI models in question were producing works displaying substantial similarity to those in the training corpora.)

expression in the AI work itself is in the overwhelming majority of cases original.[549] In common with other copy-reliant technologies, the decisive normative question therefore is not so much one that turns on the act of communication to the public, as owing to the original nature of the output *prima facie* no infringing act can be said to take place, but counter-intuitively one that pivots on the scope of the reproduction right.[550] More precisely, whether the storage of the training data, and where generalisation occurs, the act of mechanical copying in order for the model to extract latent patterns and features *about* a work as a pre-cursor to an original expressional output, constitutes an infringement. Importantly, the programs do not extract wholesale expression.[551] Instead they "learn" to identify and replicate common micro-sequences and combinations of chords, pixels, and words as governed by syntax and other pre-determined exogenous compositional frameworks. For example, the curve or straight line in a car's bodywork, or common chord sequences associated with a genre of music. Thus, ultimately, the courts are faced with the question of whether the expressive nature of the output, notwithstanding its originality, renders the reproduction and extraction of compositional micro-elements *qua* non-protectable facts and data subsumed within a copyrighted work, an infringing act.

Before discussing the legal arguments concerning these two classes of outputs, it is worth noting they both originate from reproductions in the preparation and training phase. Of import to this discussion is that in two significant 2025 US rulings brought by authors against the developers of generative AI models, the creation of a corpora of copyright materials from lawfully acquired books was found to be fair use.[552] In line with the three paradigmatic TDM cases discussed above, the courts in both cases recognised that where there is no evidence of infringing outputs, the creation of a digital library of works for analysis and *inter alia* the subsequent training of an AI model was highly transformative under the first factor. While still early days, coming off the back of a long line of TDM and related non-expressive rulings, these admittedly early generative AI decisions suggest that courts will likely continue to exclude data collection and training from the scope of authorial control.

As these two recent rulings underline, crucial to the finding of fair use is the nature of the output. Prior to the summary judgements in *Kadrey et al v Meta Platforms Inc* and *Bartz v Anthropic* , only

[549] Sag, supra 523 arguing that while rare, memorisation (i.e. overfitting) can occur where contingent on a particular prompt the training data may be recreated. Eschewing a discussion of procedural short fallings in the training process, Sag argues it is most likely to occur "when models are trained on many duplicates of the same work, images are associated with unique text descriptions, and the ratio of the size of the model to the training data is relatively large.") See also Nicholas Carlini et al, 'Quantifying memorization across neural language models' arXiv:2202.07646, 1 (referencing studies that find between 0.00000015% and 1% of training data may be memorised.)
[550] Sag, supra 483 (arguing that copyright's primary function is to protect the author from expressive substitution and therefore the act of communication to the public "is the touchstone of copyright infringement" and acts of reproduction with no access should be exempted from copyright protection.) 1628.
[551] See Sag, supra 523, passim.
[552] *Kadrey et al v Meta Platforms Inc*, No 23-cv-03417-VC (N D Cal June 25/6/2025) (although raising questions of market dilution, finding the training of Meta's LLM (LLaMA) on in-copyright books to be "highly transformative.") *Bartz v Anthropic PBC*, No 3:24-cv-05417-WHA, 2025 WL (N D Cal 23/6/2025). (finding that "the purpose and character of using copyrighted books to train LLMs to generate new text was quintessentially transformative." However, notably, the Court rejected the use of pirated copies for building a digital AI corpus arguing that a fair use finding in such an instance "would destroy the [entire] publishing market.") 30.

Google Books involved any substantive expressional output, and in that instance the different purpose and usage (i.e using a snippet to decide whether the book was of relevance to the user) clearly sets it apart from media AI works whose objective is, like the original work in the training set, to be appreciated and enjoyed by a human audience. Thus, while predicting how a future court may rule is therefore not without significant challenge, as we see in the *Meta* and *Anthropic* rulings, it is posited that a combination of different legal doctrines may lead the courts to conclude that generative AI is fair use where there is no substantial similarity between the inputs of the model and their outputs.[553]

The rest of this subpart will outline how existing case law may shine a light on how the fair use doctrine could be applied to media AI works.

In terms of the fundamental tenets of copyright law, and as a number of leading commentators have asserted, copyright is ultimately designed to protect less the mere reproduction of a copyright work than its communication of original authorial expression.[554] Thus, where no original protected expression is communicated to the public, as was the case *inter alia* in *Hathi Trust, iParadigms* and *Anthropic,* and because AI models as a rule do not contain the training data, predicating infringement on the act of reproduction alone will run contrary to nearly two decades of US case law.[555]

At the technical level, in most neural networks, all that is held is a learnt generalisation of the core relationships and features from the training data in the form of mathematical computations. In the case of creative works, having analysed the micro-elements and relationships which inhere in the syntax of the language, the combination of pixels, or the rhythm, melody, and harmony of music, the relationships that have been identified – from (in most instances) anywhere from hundreds to billions of examples in the training data – are then represented, not as facsimile replicable data, but in the form of executable mathematical calculations within the model. Namely, the model parameters and other integers that operate in tandem with the model's programs. Ultimately no verbatim reproduction of the original copyright work remains. Instead, what has been created is a "mental model"[556] of all the concepts within the whole training dataset which, upon the receipt of an instruction, allows the model to reconstruct the relationships encoded therein. These mathematical predictions may produce an output which, dependent on the training, nature of the programs, and instructions received can

[553] Cf. Lemley & Casey supra 481, 777. Sag, supra 488, 322 ff. (positing that generative artwork is a less clear fair use). Sobel, supra 523. (arguing that style transfer techniques are not fair use). Note Sag supra 523, taking a position (that "machine learning models still qualify as non-expressive use so long as the outputs are not substantially similar to any particular original expression in the training data.") 309.
[554] See Sag, supra 483, 1608 ff. Drassinower characterises copyright as "an exclusive right of public presentation." Abraham Drassinower, 'Authorship as Public Address: On the Specificity of Copyright Vis-a-vis Patent and Trade-Mark' (2008) 199 Mich St L Rev 204.
[555] For paradigmatic non-expressive use rulings see supra 493 and 494.
[556] See Theodoros Chiou, 'Copyright lessons on machine learning: what impact on algorithmic art' (2019) 10 J Intell Prop Info Tech & Elec Com L 398 (deploying the term "mental model" to underscore that most AI models will not contain the data they were trained on, but develop the ability to create abstractions of the works they trained on.) 399.

range from entirely original to something which overfits and as a consequence is similar to the work it has been trained on.[557]

Without doubt, courts evaluating original media AI works will examine if any new work created from the training corpora exhibits substantial similarity, or in the case of transformative uses, substitutability with the original copyright work. At the level of the AI model itself, derived from the training corpus, it comprises executable software instructions, which as pointed out by Anthropic's defence clearly bear no resemblance to the copyright works upon which they are trained.[558] As such, it performs a different purpose and function to the original and therefore cannot as code "supersede or supplant" the copyright works from which it has in part been derived.[559] As with the human brain, that stores shadows of information in imperceptible form within the synaptic nerves, so it is with most neural networks: no copy of the original works remain. What has been created is, in the words of Judge Leval, "new information"[560] – a mechanism that like any other computer software, can, if instructed, produce copyrightable works, but which in and of itself as a creative tool does not substitute for the creative works it was trained on.

Moving to the question of original model outputs, a court may well consider whether a novel AI media work constitutes a derivative work. As defined in §101 of the US Copyright Act, a derivative work is one which is "based upon one or more preexisting works" presented in a form which "may be recast, transformed or adapted." Thus, as highlighted by Samuelson et al, in order for a claim to succeed there must be proven substantial similarity in expression.[561] As Goldstein sets forth, in order to constitute a derivative a "work must borrow *original and expressive content* from another work."[562] As outlined above, however, where novel, the AI outputs do not resemble any individual work. Instead, akin to a pastiche, they only utilise unprotected ideational and generic forms of expression from the original and therefore evidently do not constitute a derivative work.

Given the transformative nature of the creation of both the training corpora and the resulting model, the attention of the court will by necessity not just fix on the nature of the outputs, but also the existential question of the economic impact of the existence of the AI model as a tool on (human) incentives to create. That is to say, where there is no substantial similarity between the inputs and the outputs of the model – as can be observed in many of the cases currently progressing currently

[557] For example, over-fitting can be the result of poor tuning of the loss function. For a more detailed explanation of how overfitting is avoided in the context of an image generator, see Sag, supra 523, 323-5.
[558] *Bartz v Anthropic PBC*, No 3:24-cv-05417-WHA, 2025 WL Doc 122 22 10 (27/3/25) (stating it "is undisputed that Plaintiffs' books cannot write code".)
[559] *Authors Guild v Google* (2013) paraphrasing *Folsom v Marsh,* at 348. *Harper & Row v Nation Enterprises,* 471 US 539 (1985) 562.
[560] Leval, supra 528, 1111.
[561] See cases cited at n.98 in Pamela Samuelson et al, 'Comments in response to the copyright office's notice of inquiry on artificial intelligence and copyright' (30/10/2024) http://dx.doi.org/10.2139/ssrn.4976391
[562] Paul Goldstein, *Goldstein on Copyright* (Wolters Kluwer 2007) §2.16.2.

through the US courts[563] – the question arises whether the dissemination of an entirely original media AI work (produced by a ML model trained on transitory reproductions of protected subject matter) indirectly creates negative externalities on the potential market for, or value of, copyright works.

As we see in a number of the recent generative AI rulings, broader concerns of market harm and dilution arising from generative AI models feature prominently.[564] In terms of precedent, it is entirely plausible that courts considering questions of market harm may turn to the landmark case of *Campbell v Acuff-Rose.* A careful reading of this ruling reveals that the Supreme Court ruled that the only market encroaching activities that courts should contemplate are those that "serve … as a substitute for *the* original work."[565] Thus, where a generative work is either substantively similar to another work or an adaptation, save those where something "different and new"[566] has been added to the original or it fails to reach a threshold of substitutability,[567] it will be found to infringe. Likewise, in *NXIVM Corp v Ross Institute,* the Court of Appeals of the Second Circuit ruled that the only copyright relevant acts taken into consideration are those that impact on the "market for [the] *plaintiff's* expression … not on the effect of defendants' work as a whole."[568] Novel AI works absent of any original expression other than their own, evidently do not serve as a substitute for individual works in the training data and as a result would therefore inevitably fail a strict interpretation of market harm as outlined in *Campbell* and *NXIVM*.

Central to this unarguably pro-competition approach taken by the US higher courts are cases involving copying as an intermediate step in the production of a new and original work.[569] Whilst the existence of new or competing products in the marketplace as a result of liminal copying may well have an adverse effect on incentives for rightsholders, US courts have nevertheless repeatedly determined "economic loss … as a result of … competition does not compel a finding of no fair use."[570] Salient to this discussion are the landmark reverse engineering cases *Sega* and *Sony.* In both

[563] For example, *Kadrey et al v Meta Platforms In*c, *Bartz v Anthropic, Universal Music Group et al v Suno Inc* 1:24-cv-11611 50 (N D Mass 24/6/2025) ("Plaintiffs are not . . . alleging that these outputs themselves infringe the Copyrighted Recordings"). *Mike Huckabee v Meta Platforms Inc etc al* ("Bloomberg") 1:23-cv-09152-MMG (SDNY 22/03/2024) (The plaintiffs stating that the purpose was not to reproduce their works or to create competing works, but to "teach the model grammar, vocabulary, context, and various language patterns.")

[564] In *Anthropic,*at 28, Judge Alsup rejects a market dilution argument stating that the purpose of copyright law is to promote authorship and not protect existing authors from "an explosion of competing works". By contrast, in *Kadrey v Meta* whereas no evidence was presented and therefore the court could not make a determination, Judge Chhabria was evidently receptive to market dilution arguments.

[565] See *Campbell* at 587 quoting Leval, supra 528, 1123. Emphasis added to highlight the phraseology indicates a specific copyright work and not copyright works in general of the same category.

[566] *Cariou v Prince,* 714 F 3d 694 (2d Cir 2013) 706 (where the majority of an appropriation artist's paintings and collages incorporating the plaintiff's photographs were held to qualify as fair use). See *Campbell* at 579.

[567] *Authors Guild v Google* (2015) 229.

[568] *NXIVM Corp v The Ross Institute,* 364 F 3d 471 (2d Cir 2004) 482.

[569] *Walker v University Books,* 602 F 2d 859 (9th Cir 1979) (finding the reproduction of cards to create the defendant's own cards was a non-infringing preliminary step). *See v Durang,* 711 F 2d 141 (9th Cir 1983) (ruling that drafts were insufficient to support an infringement claim as the final work did not resemble the plaintiff's script). *Sega Enterprises Ltd v Accolade Inc,* 977 F 2d 1510 (9th Cir 1992) 1522 (finding that irrespective of the ultimate intent being to produce unlicensed games that could be played on Sega's console, the reproduction of the plaintiff's object code was for the *direct* purpose of research and therefore fair use.) *Sony Computer Entertainment v Connectix* at 607 (finding that intermediate copying of the plaintiff's object code to create a competing game platform that was able to run Sony PlayStation games was fair use.)

[570] *Sony Computer Entertainment v Connectix,* 607.

cases the courts found intermediate copying intended to enable novel downstream products to compete in the same market as the original to be fair. In taking such a position, the judicial authorities sought to distinguish infringing acts which usurp the expression that inheres in a specific work of copyright, from those broader activities which, as a result of new competition, simply "supress … demand".[571]

Critically, as articulated by the Ninth Circuit in the *Sega* decision, any "attempt to monopolize the market by making it impossible for others to compete runs counter to the statutory purpose of promoting creative expression and cannot constitute a strong equitable basis for resisting the invocation of the fair use doctrine."[572] Thus, from an economic perspective, while the exercise of proprietary entitlements in the United States rightly operates to protect unique authorial expression, at the market level it may refuse to endorse anti-competitive legal interventions that bar new entrants whose products are novel and original in nature.

More recently, similar thinking on the chilling effect on market competition of an overly broad copyright regime has been evident in the Supreme Court's reasoning in *Google v Oracle.* Concerned that a rightsholder's monopoly control over software would restrict parallel as well as down-stream innovation, in support of its fair use determination the court opined that Google's unlicensed use of Oracle's APIs "creates new opportunities for [a] whole market to grow".[573] Thus, also relevant to the fourth factor, are the wider "public benefits the copying will likely produce."[574]

As in the landmark reverse engineering cases Sega and *Sony* before it, the Supreme Court's *Google v Oracle* decision strongly evidences the concern the higher courts have for safeguarding competition and new markets from the excessive monopolistic effect that copyright can give rise to. In no small measure, this awareness of anti-competitive behaviours is driven by the constitutionally prescribed purpose of copyright in the United States. As has frequently been observed by the courts, while private incentives and rewards should accrue to authors "the ultimate aim [of copyright] is … to stimulate artistic creativity for the general public good."[575] Given the ubiquity of valuable AI applications in both business-to-business and consumer markets, it seems questionable whether a higher US court would seek to harm the very real innovational and creative benefits that accompany the widespread adoption of this new general purpose technology. Not only is AI a capacious market

---

[571] *Campbell* 592 quoting *Fisher* v *Dees,* 794 F 2d, 438.
[572] *Sega v Accolade Inc* (9th Cir 1992), 1523-4.
[573] *Google LLC v Oracle America Inc,* 141 S Ct 1183 (2021) 27.
[574] Id. 35.
[575] *Twentieth Century Music Corp v Aiken,* 422 US 151, 156 (1975). See also *Sony Corp v Universal City Studios* quoting from *Fox Film Corp* v *Doyal,* 286 US 123, 127 ("The sole interest of the United States and the primary object in conferring the monopoly … lie in the general benefits derived by the public from the labors of authors.")

for technology companies spawning numerous public welfare benefits,[576] but perhaps crucially for any future ruling it is also creating new opportunities for artists and the creative industries who are adopting it in large numbers as a tool for their own creative endeavours.[577] Thus, as a reductive tale of free-riding may ultimately fail, because AI has also become a potent and very real tool for creators themselves.

A further challenge for any court considering media AI works is that while an output is original in expression, at the same time, stylistically it will commonly bear a generic resemblance to the works upon which it was trained.[578] Loss functions and other routines in the model are able to ensure outputs "in the style of" the data seen in the training phase, and whereas some may display greater originality, others will no doubt be more derivative in expression. Decades of jurisprudence concerning expressional similarity in the context of parody,[579] pastiche, art,[580] music composition,[581] and *scènes à faire* nevertheless inform that style, genre, and stock characterisation lie beyond the scope of copyright protection.[582] Thus, as discussed in detail in Chapter 4, in copyright law the appropriation of style and common themes from the work of others may, depending on the particular circumstances, fall not copyrightable expression but as a non-protectable idea. As highlighted by Henderson, Lemley et al: where the expression in a media AI work is original, irrespective of stylistic commonalities with the works it has been trained on, courts guided by copyright doctrine may determine such outputs to be non-infringing.[583] Such determinations are nevertheless highly complex and, given the subjective nature of the determination, prone to being arbitrary in nature.[584] Certainly, the recent rejection of fair use by the Supreme Court in the context of the appropriation art case *Warhol* v *Goldsmith,* in which Andy Warhol created a highly-colourised silkscreen illustration of a photograph of Prince, proves much food for thought for those wishing to build stylistically on the work of others. Not only could the ruling deter some companies from launching generative AI products, the *Warhol* case may form a potentially significant backdrop to the ongoing generative AI cases in the US, particularly where the works of others are so directly built upon.

[576] See *Generative AI Market and Region Forecast, 2022 – 2030* (Acumen Research and Consulting 2023) (predicting that the generative AI market will be worth $110.8 billion by 2030 growing at 34.3% per annum.)
[577] Ben Rogerson, 'Almost 60% of musicians are already using AI in their productions, study suggests' (*musicradar,* 21/4/23).
[578] For example, NovelAI Diffusion Anime is designed to create text to image creation of anime.
[579] *Campbell v Acuff-Rose.*
[580] *Cariou v Prince,* 714 F 3d 694 (2d Cir 2013).
[581] For example, see *Griffin v Sheeran,* 17 Civ 5221 (LLS) (SDNY Apr 5, 2023) (finding that the chord progression in the Sheeran song 'Thinking Out Loud' was commonplace.)
[582] See Richard Posner, 'When Is Parody Fair Use' (1992) 21 J Legal Stud 1 67, 68 (discussing style and genre as non-copyrightable in the context of parody). Joanna Demers, 'Sound-Alikes, Law, and Style' (2014) 83 UMKC L Rev 303 (concluding that sound-alike music where styles and genres are recycled produce "semblances [which] are perfectly legal.")
[583] Peter Henderson, Mark Lemley, Xuechen Li et al, 'Foundation Models and Fair Use' (2023) Stanford Law and Economics Olin Working Paper No 584, 14.
[584] cf. See Kagan J and Roberts CJ's strong dissent in *Andy Warhol Foundation for the Visual Arts Inc v Lynn Goldsmith et al*, 21–869 (18/5/2023) (referring to the Supreme Court's analysis of the first fair use factor as a "shambles".) 2.

Away from purely economic arguments, an observant court may take into consideration the negative public welfare impact on the accuracy of AI algorithms that a finding against fair use will unquestionably engender. As a generalisation, a reliance on licensing will inevitably reduce the volume of data a model is exposed to in the training phase. In turn, depending on the significance (veracity) and volume of the data denied the model, the prediction may well be negatively impacted upon. Not only will this reduce the standing of the AI software in the marketplace, but from a public welfare perspective increase the potential for severe negative externalities that can arise from bias, false positives, and other types of incorrect predictions. As Levendowski has argued, given that as a rule more data equates to higher predictive standards, copyright has the "power to bias AI systems" dangerously where exclusive rights curtail and limit access to training data.[585]

In conclusion, whereas by virtue of a trio of TDM cases ML models that produce outputs of a non-substitutive nature for the works they have been trained on can be said with high levels of certainty to be fair use, the same conclusion cannot necessarily be drawn for original media AI works. Nevertheless, drawing a clear blue line between these two different forms of AI may be more complex than it first appears. Differing strands of case law concerning intermediate and transformative copying, together with the public interest inherent in protecting market competition from the monopoly effects of copyright, may collectively support a finding that the creation of an original AI-generated work constitutes fair use. Further, the constitutional underpinnings of US copyright law which create a nuanced judicial approach to market dilution may ultimately serve to determine that many forms of AI, beyond just those which undertake purely text and data mining type analysis, are legitimate non-infringing uses of copyright works.

Section Three

3. Scope of the Japanese and US Exceptions

This section explores the parameters of fair use and Article 30-4 to compare the distinct approaches taken by each jurisdiction. The first part examines the scope of the respective exceptions from the perspective of the economic rights provided by copyright law. After this, the interplay with contract law and technological protection measures is evaluated. Finally, the analysis concludes by determining which jurisdiction provides a more robust legal framework for innovators engaged in machine learning.

[585] See Levendowski, supra 536, 630.

3.1 Analysis of Scope

In contrast to the European approach to AI, the most striking feature is the open and flexible nature of the legal response. Reflective of their respective legal traditions, whilst the UK and EU have introduced a set of rigid and prescriptive exceptions, by contrast the US and Japan rely on the flexibility provided by open copyright norms to support AI and other forms of data-drive innovation. More precisely, it is well established that both exceptions allow forms of statistical and non-statistical data analysis,[586] TDM, machine learning, etc,[587] where the output is broadly non-expressive in its form.[588] While the precise legal position of media AI works is subject to ongoing litigation on a number of fronts in the US,[589] the Copyright Subcommittee of the Agency for Cultural Affairs in Japan has issued an opinion stating that the use of style in generative works, absent it being associated with an author's own particular expression, falls within the scope of Article 30-4.[590]
In terms of the analysis in the rest of the chapter, owing to the open-ended nature of both Article 30-4 and §107 and the limited case law,[591] hereinafter presumptions of differing levels of exactitude are presented in examining the extent to which AI is facilitated in these two jurisdictions.

3.2 Prohibitions on Types of Works

Although in the EU legitimate questions remain over whether the data analysis of computer programs is permitted under Article 3, no prohibitions against use of a particular statutory category of protected work is apparent in either the laws of Japan or the United States. In the US, although the second fair use factor requires an evaluation by the courts of the nature of the copyrighted works, historically the use of highly creative works has tended not to act as an obstacle to transformative uses when it comes to copy-reliant technologies.[592] Thus, as illustrated by the *Google Books* cases and *Hathi Trust*, whereas US copyright law makes in principle a distinction between different categories of works, it rarely proves dispositive.[593]

The same conclusion can, with one exception, be reached in regards to Japan. Whereas Article 30-4's preamble sets out that the "nature of the work" is a determining factor, ultimately infringement is predicated on whether use of a copyrighted work "would unreasonably prejudice the interests of the

[586] For Japan, see subsection 3.2.
[587] For US, see subsection 2.2.1.
[588] In the US, non-expressive forms of AI are so well-entrenched that they currently comprise one of the lawful acts of technological protection circumvention pursuant to §1201(a) (1)(c) of the US Copyright Act.
[589] Supra 545.
[590] 『AI と著作権に関する考え方について』(Copyright Subcommittee, Agency for Cultural Affairs 15/3/2024) Hereinafter "Generative AI Report" (stating "commonality of style itself does not constitute copyright infringement." The Report goes on to outline that where a limited number of copyright works are analysed with the intention of extracting all or part of the creative expression in an output where that expression can be determined it will fall beyond the scope of Article 30-4.) 21, 23.
[591] As far as this author is aware there is currently no case law in Japan regarding Article 30-4. See also Mendis et al, supra 353, 18.
[592] See *Kelly, Google Books*, *Google v Oracle, HathiTrust*, *Perfect 10* etc which involve different types of "creative" literary and artistic works including photographs, software, fiction and non-fiction.
[593] See *Authors Guild v Google* (2015) 220 quoting William F Patry, *Patry on Fair Use* (Thomson Reuters 2015) §4.1.

copyright owner." This broad test derived from the Berne Three Step Test seeks to resolve the work category and activity-related restrictions on AI imposed by the rescinded 2009 exception.[594] In doing so, it has removed the previous prohibitions on the use of techniques other than those purely statistical in nature.[595] Additionally, reference to TDM by sole means of a computer has been deleted, thus permitting not only purely analogue forms of copying or recording but alongside this the digitisation of text, sound, and film.[596] Thus, in rendering Article 30-4 technology-neutral, the government has sought to future-proof the legislation, making reproductions by any means possible including any future activities that may not be defined as computer-based.[597]

Although not explicit from the black letter of the law, Ueno however proposes that one limitation on scope remains: namely, databases specifically developed for the purpose of data mining.[598] Importantly, the 2009 provision did not apply to "database works compiled for use by persons who carry out data analysis". While the computational analysis of such databases does not invoke any enjoyment on the part of the data scientist and therefore falls within Article 30-4's purpose limitation, based on a supplementary resolution adopted by the Upper House's Education and Science Committee,[599] Ueno observes that mining custom-made TDM databases would likely unreasonably prejudice the interests of the rightsholder.[600] This position being based on a Committee resolution that states that only "acts envisaged as being the subject of limitations and exceptions in the extant law shall continue to have effect". Thus, Ueno asserts, the 2009 prohibition on machine analysis of tailor-made AI databases conceivably continues under Article 30-4.

3.3 Communication to the Public

As touched on above, this study agrees with the position formulated by a number of commentators that material harm to rightsholders will in reality be precipitated by the communication of original expression in a copyright work, and not from those acts of reproduction routinely undertaken by computers.[601] This includes, of particular relevance to this study, non-expressive acts such as model training and acts of automated data analysis. Assuming a legitimate source, other than hypothetical harms that may occur in putative licensing markets for internal copying by computers, it is only when works are made available to the public that cognisable harms can be said to materialise.

[594] Article 47-7 (Rescinded 2018).
[595] Non-statistical forms of machine learning include bio-inspired, neuromorphic, neurosymbolic, knowledge-based approaches etc.
[596] See New Business Law, supra 472, 17.
[597] It is noteworthy that the CDSM Directive, like the JCA, foregoes any mention of the technologies whereas s.29A of the CDPA refers to "computational analysis."
[598] Tatsuhiro Ueno『平成 30 年著作権法改正について』(On the 2018 Amendment to the Copyright Act) in Intellectual Property Law Annual Report 2018-19 (Hyōronsha 2018) 4.
[599] Education and Science Committee, House of Councillors hearing 17/5/2018.
[600] Ueno, supra 598, 4.
[601] Sag and Drassinower, supra 554.

In part because TDM is an area of the law that has attracted much attention from lobbyists, the failure by the EU and the UK to explicitly address the collaborative nature of much machine learning in statute has left European organisations wishing to share training data and results in an uncertain position.[602] In stark contrast, framed by the industrial strategy context of Article 30-4, legislation in Japan creates a broad and flexible exception to the right of communication to the public.

Expanding the authorised activities permitted by the 2009 exception, Article 30-4's preamble legitimises exploitation of protected subject matter "*irrespective of the means* … where such exploitation is not for enjoying, or allowing another person to enjoy, the thoughts or sentiments in a work" (emphasis added). Therefore, as long as a third party also intends to use the work for "non-enjoyment" purposes, it can be construed to mean the communication of protected subject matter in any form and by any means is a non-infringing act. Pursuant to the Japanese legal tradition of authentic interpretation, this position has been given legal standing in the National Diet by senior government officials in public testimony to the Education and Science Committee.[603] Thus, providing "the exploitation would not unreasonably prejudice the interests of the copyright owner", the communication of a copyright work for a wide array of commercial and non-commercial sharing activities are presumptively non-infringing. Illustratively, the creation of training datasets for third parties as a service, remote networked access to processed and pre-processed data amongst companies working together,[604] and the making available of AI results where they contain copyright protected materials are in theory therefore lawful.

In the US, where there has been some opportunity to explore these issues, objection by the plaintiffs to the sharing of protected subject matter between different organisations has not been upheld. In *Google Books,* the use of the digital surrogates by the HathiTrust members for differing purposes including preservation, catalogue search, and format shifting for the visually impaired were deemed to be "activities that advance the arts and sciences" and thereby non-infringing.[605] As a consequence, the upstream supply by Google in order for Hathi Trust to undertake their activities was as a dependent action also explicitly excused by the court as fair.[606]

Whilst the act of providing reproductions of copyright works between partner institutions appears not to have troubled the courts, the sharing of protected subject matter to a wider public will

[602] Lobbying by the creative industries is particularly apparent: https://www.stm- assoc.org/2012_03_15_STM_ Summary_Statement_Text_Data_Mining_final.pdf ; https://www.stm-assoc.org/2016_11_28_STM_TDM_Infographic_A3.pdf https://libereurope.eu/the-hague-declaration/ https://www.ukmusic.org/news/uk-music-chief-outlines-threat-to-music-of-text-and-data-mining-to-musicians/ https://www.prsformusic.com/press/2022/uk-music-urges-government-to-protect-music-industry

[603] Tsukasa Nakaoka, Deputy Director of the Agency for Cultural Affairs, confirmed that the sharing of training data with third parties would fall under Article 30-4 in response to questioning on this point from Diet member Shigeki Kobayashi. The House of Representatives, Education and Science Committee. No 5. 6/4/2018.

[604] Ueno, supra 598, ibid.

[605] *Authors Guild v Google* (2013), 294.

[606] Ibid.

unquestionably attract more critical scrutiny. Where it does, whether the communicated work is qualitatively or quantitatively significant in addition to the purpose and character of the use, will be pivotal to any determination of fair use. In *Google Books*, the verbatim communication of snippets to the public was deemed transformative as their purpose was not conveying their expressional content but to establish whether the book was of interest to a user of the service.[607] Thus, where an activity is deemed transformative, as discussed above, the communication of AI outputs will be deemed non-infringing, as they neither supplant nor substitute the works from which they were derived.[608] More broadly, as Levendowski sets out, the ability to freely use protected subject matter to improve the probabilistic outcomes of AI, and thus reduce bias, will incontrovertibly play a role in the safeguarding of public welfare.[609] It seems highly probable, therefore, that the proportionate making available of protected works for the purpose of reducing negative social and economic outcomes may find favour with the courts, possibly couched in language that rests on the copyright clause in the US constitution.

In Japan, it remains to be seen whether authentic interpretation by senior government officials on the wide scope of communication to the public will be endorsed by the courts. Whereas sharing in the context of TDM and other knowledge retrieval techniques may well be non-infringing under Article 30-4, the situation surrounding media AI works is far less certain. Whereas a mimetic work displaying substantial similarity to copyrighted materials in the training set will almost certainly be infringing, it is more difficult to predict how the courts will approach the knotty issue of original generative AI. At stake will be whether the authorities conceptually view there to be an uninterrupted causational link between the quintessentially "non-enjoyment" computational analysis undertaken by the computer and the generative outputs designed to be enjoyed by a human audience.[610] In other words, whether the mechanical processing of the copyright works to set the parameters of the model – a quintessential case of "non-enjoyment" – is superior to, or integrally coupled with, the fact that *ex-post* entirely novel and original outputs are as a consequence enjoyed by a public. Thus, a court could consider that the overarching intent of the computational analysis is to elicit sensual or intellectual enjoyment in the outputs and as a result conclude the final motive renders the automated and mechanical reproduction null and void, even though at the liminal stage, it evokes no enjoyment still less perception by a human being.

It is axiomatic that the industrial policy goal of Article 30-4, and the precise form of its enactment, may however suggest another outcome is equally plausible. Article 30-4 (iii) expressly renders the

[607] *Authors Guild v Google* (2015), 227.
[608] See subsection 2.2 and 2.6.
[609] Levendowski, supra 536.
[610] See n.590 where the authorities have stated that "commonality of style [in generative AI outputs] itself does not constitute copyright infringement."

"*data processing by a computer, or any other form of exploitation, that does not involve perception of the expressions in the copyright work by the human senses*" as a non-infringing act. As model training is imperceptible to the human senses, a court may struggle to contradict such a clearly enumerated permitted activity where the generated output is either entirely original or fails to reach a threshold of protectability. A court alive to the strategic policy goals of Article 30-4, not least its link to the government's Society 5.0 strategy, may excuse the communication to the public of original media AI works justified by the broader political and economic context of the exception. Without doubt, a restrictive ruling that situates original outputs from machine learning as infringing would foreclose much new innovation beyond that of just the creative sector. As a consequence, it would run counter to the very clear industry-oriented copyright objectives of the IPSHQ as validated and enacted by the National Diet.

3.4 Lawful Access

Whereas the European exceptions require lawful access to a work as a precondition to undertaking TDM, Article 30-4 remains silent on the issue. Consequently, it remains uncertain whether the use of lawfully acquired sources is a prerequisite under Japanese law. The doctrinal justification for the introduction of Article 30-4 indicates that, with the potential exception of databases specifically created for AI,[611] lawful access is not a legal prerequisite for enjoying the exception. That is to say that as data processing and other non-enjoyment acts fall outside of the doctrinal purpose assigned copyright by the Japanese legislature, lawful access may not be a requirement as it will not cause harm to the legitimate interests of rightsholders. As attested to by the Deputy Director of the Agency for Cultural Affairs before the Diet, as the act of data analysis does not trade on the enjoyment of the thoughts or feelings expressed in a work, it does "not impair the opportunity for rightsholders to receive remuneration … and therefore as a rule does impact on the interests of the rightsholder that the Copyright Act is intended to protect."[612] Thus, paradoxically, a requirement for legal access would reward copyright owners for an activity that the Japanese authorities have explicitly defined as being unrelated to the purpose for which an author's works are granted legal protection. In conclusion, it may then be plausibly inferred that neither lawful access nor a lawful source to a work is a necessary requirement of undertaking ML, as it does not trespass on the privileges afforded copyright owners by the law.[613]

---

[611] See subsection 3.2.
[612] Nakaoka, supra 603.
[613] See supra 590. It is interesting to observe that no particular view in the Copyright Sub-Committee Report is expressed on the use of pirated content in the training phase. The Report's references to the use of such content by developers in the context of generative AI refer only to potential liability when data is knowingly collected from pirated sites *and* it results in outputs of a creative expressive nature in the "generation and use phase". 27-9.

Whether a similar argument can be advanced in the United States is a less certain matter of law. In the landmark case *Harper & Row*, the judicious quoting from President Ford's memoir in a magazine was found to be infringing, the Supreme Court opining that the character of the first factor necessarily "presupposes good faith and fair dealing".[614] Likewise in *Atari Games Corp v Nintendo of America Inc,* the Court of Appeals of the Federal Circuit denied fair use by asserting bad faith on the behalf of the plaintiff. In the view of the court, Atari had lied to the Copyright Office to obtain the source code, and relying on *Harper & Row* the judge concluded that "[t]o invoke the fair use exception, an individual must possess an authorized copy of a literary work."[615] The recent summary judgements in *Anthropic* and *Meta* take a similar position, concluding that only the use of a lawfully acquired work complies with the fourth fair use factor.[616]

Nevertheless, the *Anthropic* summary judgement notwithstanding, much recent case law has reversed the applicability of the clean hands doctrine to fair use. In *NXIVM Corp v Ross Institute et al,* the Second Circuit Court of Appeal found the authorities in *Atari* to have misread *Harper and Row*, concluding that bad faith was not injurious to a fair use defence because the defendants' use of copyrighted materials was transformative.[617] Accordingly, Carroll contests "[w]hether good faith is, or should be, a legally cognizable factor in the fair use analysis".[618] In doing so, he points out that in *Campbell* the Supreme Court cast doubt on the validity of *Harper & Row* making it clear that "[i]f the use is otherwise fair, then no permission need be sought or granted".[619] Similarly, the influential Second Circuit judge Pierre N Leval asserts that a good faith requirement would run contrary to both the constitution as well as ultimately the goal of US copyright law itself, namely "the enrichment of public knowledge."[620, 621] Like Carroll, Leval maintains that *Campbell* corrected *Harper & Row*'s "misstep" on a good-faith requirement.[622] Furthermore, he contends that a good faith requirement would be detrimental to freedom of expression as it would impose huge burden and risk on publishers and authors engaged in investigative journalism, biographies and the like, who otherwise would be reliant on the goodwill and sanction of the author in order to do their work. Moreover, according to Leval, courts would be in an invidious position if they had to interpret good faith as they would have

[614] *Harper & Row v Nation Enterprises*, 471 US 539 (1985). Citing John Schulman, 'Fair use and the revision of the copyright act' (1967) Iowa L Rev 53 832.
[615] *Atari Games Corp v Nintendo of America Inc,* 975 F 2d 832, 843 (Fed Cir 1992).
[616] Supra 552.
[617] *NXIVM Corp v The Ross Institute*, 364 F 3d 471 (2d Cir 2004) (finding that posting online sections from a manuscript that was subject to a non-disclosure agreement was transformative as it involved criticism of the original work.)
[618] Carroll, supra 487, 954.
[619] *Campbell.* n.18 Citing *Fisher v Dees* 794 F2d 432, 437 (CA9 1986) (relating to the unauthorised use of a song to create a parodic version.).
[620] Pierre N Leval, 'Campbell as fair use blueprint' (2015) Wash L Rev 90 597, 612.
[621] See Carroll, supra 487, 957 ff. (arguing inter alia that a requirement for lawful access would deprive "society of the benefits of the secondary use as a punishment for a prior infringing act" and that a data scientist would be denied the constitutional right to free speech based on their own findings from processing data.)
[622] Leval, supra 620, 614.

to make a "moralistic evaluation" as to the nature of the dealings between the author and their sources.[623]

It is posited that case law pertaining to search engine caching can also be used to support the view amongst leading commentators that lawful access to content is not an absolute requirement of the US copyright framework. In *Field v Google* the Court's decision that the activity of search engines was *inter alia* fair use implicitly recognises the legally complex realities of search engine caching. In copying, website terms of use may well be violated and a certain proportion of copyright works available on the internet and reproduced by search engines in their cache will inevitably be infringing.[624]

In sum, therefore, subsequent to *Harper & Row,* some of the more recent precedent suggests both explicitly and implicitly that lawful access is not necessarily seen by the courts as dispositive against a fair use determination. Although the recent summary judgment in *Anthropic* strongly diverges from this line of rulings, the jurisprudential landscape concerning generative AI and copyright is in a formative stage. As such, despite its intuitively meritorious stance on piracy, it is too premature to conclude whether this interpretation will emerge as the prevailing judicial approach in machine learning cases.

3.5 Storage and Security

As with lawful access, storage and security requirements are similarly not a feature of the JCA. On that account, the long-term preservation of data by the beneficiaries of Article 30-4 would in all probability be deemed non-infringing, as when directly required by an otherwise permitted act, such as TDM, it would not "prejudice the interests of the copyright owner". Likewise, many digital preservation and dark archiving processes are imperceptible to the human senses and therefore would be explicitly permitted under Article 30-4 (iii) which excuses all non-display uses from the scope of exclusive rights.

§107 of the US Copyright Act similarly contains no stipulations governing storage and security. All the same security and the harms that may befall authors from unauthorised access to their works featured prominently in both *HathiTrust* and *Google Books*. In the former case, in response to assertions by the plaintiffs that security breaches may result in market harm, upon extensive examination and in the absence of substantiating evidence, the court nevertheless rejected the claim.

[623] Id. 613.
[624] Mark Lemley, 'Intellectual Property and Shrinkwrap Licenses' (1995) 68 So Cal L R 1239 (stating that in some cases browse-wrap licences may be enforceable against users and therefore by implication that the indexing of websites with standard terms of use may not always be lawful.)

Citing *Sony v Universal*, Judge Parker opined that in order to succeed any security-related claims needed to be not merely "speculative"[625] but probative that harms are "certainly impending".[626] In *Google Books,* while ultimately ruling Google's security measures to be sufficient, nevertheless the court did not reject the plaintiffs' security concerns out of hand, finding them to be "theoretically sound".[627] Leading to the conclusion that entities wishing to enjoy fair use will also have to "carry [their] burden" in protecting the security interests of authors.[628] Thus, lax security measures that would contribute to a plausible chance of imminent breach may potentially negate a fair use defence pursuant to the fourth factor.[629] Whereas this is certainly equitable, the judgements themselves give no indication of the standard of security measures required by those undertaking machine learning. This author's opinion therefore aligns with that of Sag, that in operation any security obligations or damages awarded should be proportionate to the harm the breach is likely to cause to rightsholders.[630]

3.6 Acknowledgement

As a general observation, L&Es not only create user freedoms but are frequently accompanied by obligations on beneficiaries also. As one example, grounded in the Berne Convention,[631] the moral right of attribution is something common to most copyright systems and furthermore has force in principle irrespective of the exercise of limitations and exceptions.[632] In view of this, it is therefore perhaps surprising that the UK is the only jurisdiction in this study that has sought to create explicit obligations for machine learners to provide acknowledgement.[633]

Beyond the narrow confines of the law, it can be observed in practice that the citing of authors whose works have been used in the training or analysis phase of TDM appears to be uncommon.[634] While from a legal perspective this may be explained by differing moral rights regimes internationally, in the main, this is probably owing to practical business reasons, including competitive advantage, the sheer

---

[625] *Authors Guild v Hathi Trust* at 101. Citing *Sony Corp of America v Universal City Studios Inc,* 464 US 417 (1984).
[626] Id. *Authors Guild v Google* (2015), 227. Citing *Clapper v Amnesty Int'l USA* 568 US 398 (2013).
[627] *Authors Guild v Google* (2015), 227.
[628] Id. 228.
[629] For an in-depth exploration of security issues and fair use, see David Nimmer, 'Security Measures and the Future of Fair Use' in Shyamkrishna Balganesh (eds) et al, *The Cambridge Handbook of Limitations and Exceptions* (CUP 2021) (arguing for a more explicit basis for courts' fair use decisions concerning security measures.)
[630] See Sag, supra 488, 355.
[631] Berne Convention for the Protection of Literary and Artistic Works, Article 6 bis.
[632] See Main Committee I Report of the WIPO Stockholm Conference 1971: "as in the case of *all uses of the work*, the rights granted to the author under Article 6 bis (moral rights) are reserved." II Record [205]. See Peter Jaszi 'Public interest exceptions in copyright: A comparative and international perspective' in *Correcting Course: Rebalancing Copyright for Libraries in the National and International Arenas, Columbia University, New York, May* 6 (2005) 10. Note however that fair use in US law has been rendered an express limitation on the right of attribution and integrity granted to visual artists by §106A (a): ("*Subject to* section 107 …"). See also Tanya Aplin and Lionel Bently, *Global Mandatory Fair Use* (CUP 2020) 57 ff. (discussing the operativity of Article 6 bis vis-à-vis Article 10 (3)).
[633] s.29A1.b.
[634] Rita Matulionyte, 'Can AI Infringe Moral Rights of Authors and Should We Do Anything About It: An Australian Perspective' (2023) 15 Law, Innovation and Technology 1 124, 132.

number of works involved in most data analysis and training, high costs,[635] as well as low awareness amongst businesses of the relevance and application of copyright law to AI.[636]

The issue of attribution is comparatively unelaborated on in the literature. Where it has been discussed, divergent opinions are apparent. Matulionyte maintains that a *prima facie* case of infringement can be established against AI companies for not attributing authors whose works have been used in the training phase even where this reaches the level of hundreds of millions of potential contributors.[637] By contrast, in the context of TDM, Geiger proposes that the principle of *impossibilium nulla est* should apply to the citing of copyright works – a position which this author endorses given the practical challenges associated with having to attribute such a large number of copyright owners.[638] From the standpoint of competition, the high computational costs of instance attribution or the listing rightsholders whose works have been used in the training or analysis phase irrespective of their contribution to the model or its output, may serve to reinforce market dominance by incumbents as only the largest corporations will be able to afford the necessary expenditure.[639]

This question, however, is no longer a matter of legal theory alone. In November 2022, a class-action suit was filed against GitHub, Microsoft, and OpenAI for failure to attribute the authors of software used in the training of the assistive programming products Codex and Copilot. The plaintiffs claim *inter alia* that, in developing the AI software, it was trained on a large corpus of publicly accessible software code made available on GitHub and subject to eleven different open source licences, all of which mandate some form attribution.[640] As such, it is asserted in this ongoing case that not only do Codex and Copilot violate the licence terms, but the programs also infringe the US Copyright Act which prohibits the removal of copyright management information (CMI) including the name of the author[641] and the copyright owner.[642]

It would be foolhardy to predict how the case against GitHub might progress, not least because of the complex interaction of federal copyright law and contracts in the US.[643] Furthermore, the applicability of a fair use defence to the removal of CMI remains unsettled.[644] If the case proceeds and the defendants were to claim a fair use defence, empirical analysis however indicates that

[635] See Henderson, Lemley et al, supra 583, 23 (outlining the very high cost of instance attribution).
[636] See *Economic Impact of IP on investment in AI* (Mathys & Squire Consulting Ltd 15/3/2022) 7 ff. (showing that awareness of copyright as it relates to AI is limited beyond the creative industries sector.)
[637] Matulionyte, supra 634, 132 (discussing Clearview's facial recognition technology which was trained on 3 billion images.)
[638] Christophe Geiger, Giancarlo Frosio, Oleksandr Bulayenko 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU' (2019) Centre for International Intellectual Property Studies (CEIPI) Research Paper No. 2019-08, 14.
[639] See Pouya Pezeshkpour, Sarthak Jain et al, 'An Empirical Comparison of Instance Attribution Methods for NLP' (2021) arXiv:2104.04128 [cs.CL]. Instance attribution is the process that leads to the retrieval of "training instances that (may have) led to a particular prediction" and as the authors note "is computationally expensive, to the degree that may be prohibitive in many cases."
[640] *J Doe 1 v GitHub* n.4.
[641] §1202 (c) (2).
[642] §1202 (c) (3).
[643] See subsection 3.7 of this chapter.
[644] See Henderson, Lemley et al, supra 583, 16.

attribution is rarely a requirement. Elkin-Koren and Netanel's study of 185 fair use rulings reported on between 2008 and 2018 reveals that only two rulings determined that a failure to attribute may weigh against a positive finding, while confusingly two rulings explicitly stated the exact opposite.[645]

If the adjudicating court were to rule that the citation of the authors whose works were used in the training or analysis data is indeed necessary, to avoid confusion in the marketplace a nuanced approach to attribution should be sought. For those privileged to have access to the training data, contingent *inter alia* on the model type, its interrogation will reveal some information about the contribution of a specific individual's work to the model as well as its output.[646] In contrast, with the exception of models trained on a limited number of works or where the software allows intentionally or otherwise[647] a mimetic recreation of an input, a blanket public acknowledgement of all the contributors to the training data will inevitably mislead. A blanket approach to acknowledgement will indicate a link to an author's work that may be entirely absent as there may have been no influence on the model at all. Most machine learning models which generate media AI works do so having learnt a generalised representation of the works they are trained on and, as already discussed, do not resemble the work of a particular individual.[648] Thus, from a moral rights perspective, a requirement for attribution of training data as discussed by Matulionyte[649] not only suggests a paternal bond between an author and the AI output (which in many instances is purely theoretical) but, depending on the jurisdiction and the contextual presentation, could paradoxically be viewed as misattribution and as a consequence impinge on the author's right of paternity.

Fisk identifies four principal functions of attribution which broadly intertwine with intellectual justifications for both copyright and moral rights.[650] First, is the "reward function", which refers to the reputation and human capital gained by an author being acknowledged for the work they have invested in creating. This finds some parallels in labour theories of copyright law that justify the reward of copyright in terms of human labour.[651] Furthermore, it is the desire to compensate the "bond between authors and their works"[652] that underpins the right of attribution and the right of integrity as provided by the Berne Convention.[653] Second, Fisk identifies the "discipline function", which in modern times deters fraud, and encompasses the policing of plagiarism, libel, and copyright infringement. Third, is branding and the ability to create a commercially recognised brand associated

645 Niva Elkin-Koren and Neil Netanel, 'Transplanting Fair Use Across the Globe: A Case Study Testing the Credibility of US Opposition' (2021) 72 Hastings LJ 1121, 1177.
646 Some AI models like decision trees and linear regression are inherently transparent. For more complex models like neural networks software tools to ensure explainability are now common. Examples include, LIME, SHAP, DeepLift.
647 Pezeshkpour et al, supra 639.
648 See subsection 2.4.1, Chapter 2.
649 Matulionyte, supra 634.
650 Catherine L Fisk, 'Credit Where It's Due. The Law and Norms of Attribution' (2006) 95 Geo LJ 49.
651 Labour theories of copyright stem from the work of John Locke. e.g. Second Treatise. See Fisher, supra 485.
652 Cyrill Rigamonti, 'Deconstructing moral rights' (2006) 47 Harv Int'l LJ 353, 355.
653 Article 6 bis (1).

with quality underpinned by the ability to enforce attribution in both copyright and trademark law. Fourth is what Fisk calls the "humanising function", which corporations exploit for reasons ranging from anti-piracy to brand loyalty "to shore up the legitimacy of their property rights" in the eyes of consumers.

As identified by Fisk, few if any of the reasons for attribution appear directly relevant to an AI output when it is automatically derived from a vast array of different sources. Citing merely to indicate reproduction, which is required by a number of civil law jurisdictions irrespective of the contribution of an author's work to the output, could cause confusion and provides little reward to an author wishing for their own paternity and authorship of their work to be recognised.[654] Similarly, the mere selection of a work for inclusion in a training corpus does not enhance the reputation of an author, and neither is the humanising function relevant, as the work is self-evidently computer-generated. Moreover, where the media AI work is unlikely to attract copyright in the first place, attribution does not serve to disincentivise fraudulent behaviours as the work is in the public domain.[655]

Whereas there are legitimate public interest reasons for needing to understand more about the data a model has been trained on,[656] from a moral rights perspective – where there is no substantial similarity – attribution devoid from access to the training data doctrinally makes little sense. In sum, blanket attribution of a myriad of data providers whose works are not tangible in the outputs or retained in the model seems antithetical to much of the moral rights theory that underpins a legal obligation to attribute rightsholders.
From an economic rights perspective, the justification for more publicly available information about copyrighted training data is however doctrinally more straightforward. As seen in the run up to the implementation in the AI Act,[657] the push by creators, concerned that their works are being monetised by technology companies with no chance to exercise or benefit financially from their exclusive rights raises both copyright issues,[658] as well as wider existential concerns of "robots" replacing creators and what this means not just for an artist's livelihood but creativity more broadly. The economic and social issues raised by AI from different perspectives including copyright, privacy, market stability, and safety means understanding more about the works an AI have been trained upon is vital. Thus, if – or more likely when – these issues are addressed by the courts, the obligation to acknowledge

[654] Some common law countries like the UK (s.77) and Australia (s.194) link the right of attribution to public access to the work or its derivative. In jurisdictions such as Germany, Switzerland or France the mere reproduction of a work in any medium in theory however requires attribution.
[655] See summary judgement denied in *Stephen Thaler v Shira Perlmutter Register of Copyrights and Director of the United States Copyright Office et al,* 22-1564 (BAH) 18/8/2023 (confirming the Copyright Office opinion that entirely computer-generated works are not protectable under copyright law.)
[656] See Fotios Fitsilis, *Imposing Regulation on Advanced Algorithms* (Springer 2019) (discussing the need for stringent levels of accountability and transparency given the many concerns that exist around algorithmic accountability.)
[657] Leading to the final text of Article 52c 1 (d) ("make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model.")
[658] See Rita Matulionyte, 'Australian Copyright Law Impedes the Development of Artificial Intelligence: What Are the Options?' (2021) 52 IIC 4 417 (discussing how AI impinges on the economic rights of authors.)

authors whose works have been used in the training and analysis phases requires adept and informed handling. As shown, a blanket requirement to acknowledge – independent of any significant role played by the work of the authors – not only sets to confuse, but degrade the principle of paternity upon which moral rights are based.

3.7 Contract Override of Exceptions

As discussed in Chapter 2, indelibly intertwined with the scope of TDM exceptions is the question of contract override. Where the legal framework allows for private ordering to nullify user privileges, public policy goals are placed at the volition and whim of an individual rightsholder. Furthermore, where the interplay of contract and exceptions has not been affirmatively addressed in statute, a contradictory situation arises whereby actions that are otherwise licit are either rendered explicitly unlawful or at best unclear as to whether they remain enforceable.

A neoclassicist approach to the issue regards contractual pre-emption positively as something "essential to the efficient functioning of markets."[659] As highlighted by Netanel, however, such a narrow view upends the ultimately democratic purpose of copyright law which is to incentivise the production of new and original works by anyone who so wishes, including those rooted in the work of others.[660] Thus an over-expansion via contract of a copyright owner's monopoly not only serves to undermine wider societal innovation, and in the case of AI a strategically important industry, but inevitably threatens the public and private interest balancing function of copyright law itself.[661]
As shown in the previous chapter, while s.29A of the CDPA and Article 3 of the CDSM Directive are of great jurisprudential significance – as they expressly protect machine learning from contractual override – the situation in the US and Japan remains more opaque.

The Japanese Copyright Act, as enacted, remains silent on the question of contract pre-emption. Furthermore, the issue of contract override has not been tested in the courts.[662] The Agency for Cultural Affairs, as a body with the right of authentic interpretation, has however voiced its opinion a number of times on the matter.[663] In 2007, the Agency's Copyright Subcommittee stated that while in theory contractual pre-emption of statutory user privileges is permissible, "owing to the fact that exceptions are rooted in public interest and pure necessity, the cases where contract override is likely

[659] Neil W Netanel 'Copyright and a Democratic Society' (1996) 106 Yale L J 283, 383 (quoting from *ProCD Inc v Zeidenberg* 86 F 3d 1447 (7th Cir 1996), 1455).
[660] Id. 288.
[661] For example, see Netanel, supra 659. Derclaye & Favale infra 667. Jacques De Werra, 'Moving beyond the Conflict between Freedom of Contract and Copyright Policies: In Search of a New Global Policy for Online Information Licensing Transactions – A Comparative Analysis between US Law and European Law' (2001) 25 Colum JL & Arts 239. Christina Bohannan, 'Copyright Preemption of Contracts' (2007) 67 Md L Rev 616. Lemley, supra 624.
[662] See Masashi Sawada『契約による権利制限規定のオーバーライド』(The Override of L&Es by Contract) (2021) Kopiraito (2021) 727 61.
[663] For a consolidated overview of this issue see『新たな知財制度上の課題に関する研究会 報告書』 (Ministry of Economy Trade and Industry Feb 2022) Annex 2, 36-45.

to remain valid will be limited."[664] In evaluating in what circumstances L&Es negating contractual provisions would be lawful, the Committee inferred principles of proportionality should apply, recommending that in the light of the nature of the copyright owner "it is necessary to … *inter alia* consider the extent and form of the exception, its reasonableness, as well as the purpose of the relevant laws and regulations." [665]

Whilst providing little in the form of concrete guidance, in regards to AI a 2021 report from the Ministry of Economy, Trade, and Industry (METI) proves more instructive. Pointing to Article 30-4's background as a key pillar of Japan's industrial policy, it underlines the public interest inherent in protecting technological innovation from contract override. As a consequence, it concludes that the subversion of the government's economic and industrial policy goals by private contract is unlikely to be upheld by the courts.[666] Thus, although not explicit in statute, and subject to any future determination by the courts, organisations using copyright works for the purpose of AI training and analysis are, in all likelihood, entitled to disregard restrictive provisions contained in any private agreement to which they are party.

In the United States, the question as to whether rightsholders can use contract law to undermine statutory limitations and exceptions is a uniquely complex one. Whereas in common with other legal systems the principle of freedom of contract is important, and arguably pre-eminently so, one particular feature that distinguishes the US is the unique interplay that exists between state law and federal copyright law.[667] Core to this is §301 of the US Copyright Act.[668] It dictates that "all legal or equitable rights that are equivalent to any of the exclusive rights" relating to a copyright work "are governed exclusively" by the Act. As the 1976 Act itself is silent on the issue of contract override, the *prima facie* position directs an inquirer to court rulings based on federal copyright law to establish whether the pre-emption of fair use and other exceptions is permissible.

A close investigation reveals that the dynamics between federal and state law have however created an uncertain state of affairs, that turn on whether or not the courts view the issues in question as being equivalent to those governed by the US Copyright Act. At the heart of this tension lies the "extra element test", whereby state law is not pre-empted if an extra element in the relationship between the parties is discernible beyond that which is regulated by copyright law alone. Where an extra element is in evidence, it renders the dispute as non-equivalent and, as such, entirely subject to state law.

[664] Report of the Copyright Subcommittee of the Council for Cultural Affairs (January 2007) Chapter 3 2.1.
[665] Ibid.
[666] See METI, supra 663 id. Annex 2, 51.
[667] Estelle Derclaye & Marcella Favale, 'Copyright and Contract Law: Regulating User Contracts: The State of the Art and a Research Agenda' (2010) 18 J Intell Prop L 65,74 (outlining both constitutional and case law grounds as to why freedom of contract is paramount in the US.)
[668] For an in-depth discussion of copyright pre-emption in regard to the legal enforceability of AI model licences, see Peter Henderson and Mark Lemley, 'The Mirage of Artificial Intelligence Terms of Use Restrictions' (2024) pre-print arXiv:2412.07066, 39-54.

Conceptually, an extra element "changes the nature of the action so that it is *qualitatively* different from a copyright infringement claim",[669] and as a consequence in state law, as empirical analysis reveals, freedom of contract will prevail in most instances.[670]

Rub's analysis of 279 different court opinions (hereinafter the "pre-emption study") reveals two distinct positions taken by circuit courts on the issue of copyright equivalency. Firstly, in what Rub calls the "no pre-emption approach", the mere existence of a contract between plaintiff and defendant is often treated as the extra element. That is to say, circuit courts (where such an approach dominates) determine that contracts and licences as regulated by state law are not equivalent to copyright as they are *in personam* rights. Such relationships involve volitional commitment on behalf of the parties and, as a consequence, create individual liabilities that, exogenous to copyright law, arise from breach of a specific contractual promise. Thus, where an agreement is entered into, state law pre-empts the federal Copyright Act and, as a result, exceptions such as fair use may be contracted away. Such an approach, Rub concludes, is dominant in the Fifth, Seventh, Ninth, Eleventh, as well as the Federal Circuits.[671] In contrast to the no pre-emption position is the "facts-specific approach". Where courts follow this this line of reasoning, they do not take the presence of a contract to be a *de facto* extra element, and are required to evaluate the agreement in question to determine whether it displays a qualitatively different character to that of copyright law. Such cases, which are particularly prominent in the Sixth Circuit, result in diverse opinions: some leaving the contracts unhindered and in force under state law,[672] while others permit pre-emption by federal copyright law.[673]

On the particular issue of limitations and exceptions, while cases in the pre-emption study that dealt with fair use with no further complicating factors were rare,[674] those that were referenced resulted in pre-emption by state law irrespective of the circuit and in each instance contractual override of users' copyright privileges.[675] Focusing on the landmark reverse engineering cases, while Rub's study states that "[c]reating temporary copies … in order to achieve interoperability is typically considered fair

[669] *Computer Assocs Int'l v Altai Inc,* 982 F 2d 693, 716 (2d Cir 1992) (Quotation marks omitted.) (quoting *Mayer v Josiah Wedgwood & Sons Ltd,* 601 F Supp 1523, 1535, SDNY 1985.)
[670] Guy Rub 'Copyright Survives: Rethinking the Copyright-Contract Conflict' (2017) 103 Virginia Law Review 1141.
[671] Id. 1147.
[672] For example, *Wrench LLC v Taco Bell Corp,* 256 F 3d 446, 457 (6th Cir 2001) *Forest Park Pictures v Universal Television Network Inc*, 683 F 3d 424, 433 (2d Cir 2012) (in both cases, the Court concluding that a promise to pay lay beyond the scope of copyright and therefore the implied contracts remained enforceable.)
[673] For example, *Ritchie v Williams,* 395 F 3d 283, 287-88 (6th Cir 2005) (ruling that the contract related to the exclusive acts of distribution and performance and therefore that state law was pre-empted.)
[674] Rub, supra 670, n.297. For example, in *Davidson & Associates v Jung* while the courts found the EULA and Terms of Use negated a fair use defence Rub argues that "the main reason" for this decision was the plaintiff's violation of anti-circumvention measures of the DMCA.
[675] Rub, supra 670 n.279 and n.281 *Bowers v Baystate Techs Inc*, 320 F 3d 1317, 1326 (Fed Cir 2003) in rejecting pre-emption the Court held that the shrink wrap licence which forbade reverse engineering was enforceable. *Lowry's Reports v Legg Mason Inc,* 186 F Supp 2d 592, 593, 594-55 (D Md 2002) ruling that "the contract created a private law governing fair use of the copyrighted works *inter partes*." In *Fox Broad Co v Dish Network LLC*, 160 F Supp 3d 1139, 1178–79 (CD Cal 2015) stating that "[f]air use is not a defense to a breach of contract claim".

use”,[676] a close examination of the landmark rulings in *Sega*[677] and *Sony*[678] shows that the defendants were not subject to any form of contractual relationship or EULA.[679] Thus, even in comparatively settled areas of law such as this, whether rightsholders can expand their copyright privileges by contract to prevent fair use, including the utilisation of non-protected elements of a work, remains uncertain.

Away from shrink-wrap and browse-wrap licences, in-depth academic commentary by US scholars on the topic of contract override of exceptions is not as fulsome as one might imagine. One leading commentator’s view is that “[i]t seems unlikely that courts would accept that fair use should either always or never override contractual restrictions”.[680] Samuelson asserts that a mass market licence that forbade constitutionally guaranteed acts such as freedom of expression would be treated very differently to a fully negotiated licence where the licensor insists on a clause restricting acts to which the licensee actively assented.[681]

Assuming Samuelson’s reasoning to be correct, the nature and character of the user’s assent constitutes a crucial factor. US case law nevertheless reveals that the threshold to assent to a contract can be surprisingly low. Whilst courts have generally found that standard form agreements (so-called “browse-wrap” or “shrink-wrap” licences) that require no form of assent by the visitor to a website or purchaser of a DVD to be unenforceable against consumers, according to Lemley, even this as a matter of law is not settled.[682] Where there is some form or active assent by, for example, ticking a box to agree to terms and conditions (so-called “click-wrap” licences), the US courts have in line with a strong tradition of freedom of contract usually found such agreements to be enforceable.[683] This is in spite of the fact that there are many legitimate reasons for believing most computer users have not read or indeed agreed to the terms of use they are presented with.[684] A contract law phenomenon wryly referred to by Lemley as the “no need to agree” revolution.[685]

[676] Rub, supra 670, n.279.
[677] *Sega v Accolade Inc* (9th Cir 1992).
[678] *Sony Computer Entertainment v Connectix* (9th Cir 2000).
[679] See also Renée Zmurchyk, ‘Contractual validity of end user licence agreements’ (2006) 11 Appeal: Review of Current Law and Law Reform 55, 66 (“The Davidson case differs from Sega and Sony because of the existence of a EULA.”)
[680] Pamela Samuelson, ‘Possible futures of fair use’ (2015) 90 Wash L Rev 815, 859.
[681] Ibid. 859 ff.
[682] Mark Lemley, ‘Terms of Use’ (2006) 91 Minnesota Law Review 459, 460 n.4 (citing examples of cases involving businesses where courts have enforced browse-wrap licences.)
[683] Id. 459, n.2. One notable exception to this is *Specht v Netscape,* 306 F3d 17 (2d Cir 2002) (where the court rejected the enforceability of a click-wrap licence as it lacked mutual consent and reasonable notice of the licence.)
[684] See Nathan Davis ‘Presumed assent: The judicial acceptance of clickwrap.’ (2007) 22 Berkeley Tech LJ 577 (for an in-depth discussion of the situation surrounding the judicial treatment of click-wrap licences.)
[685] See Lemley & Casey, supra 481 n.75.

Given the overwhelming arc of pro-contract case law, it is therefore of no surprise that in the context of data mining Carroll joins other US commentators,[686] maintaining that where licensees agree to limit their rights to enjoy exceptions, contracts are most likely to be enforceable.[687] Given the particular importance of scientific literature to AI, as with click wrap licences, the knotty question of what constitutes agreement therefore deserves further scholarly attention. For example, whether not just an explicit prohibition against AI but contractual silence on actions beyond those that are expressly authorised might render machine learning a breach of contract. Certainly, in the case of contracts which an organisation has not actively negotiated – as is common in the academic sector owing to the fact most licences constitute contracts of adhesion – a licensee organisation may not even be aware of its contents.[688] Thus, a university may well "agree" to curb the fair use rights of researchers to undertake AI, unbeknownst to the library staff who manage the contracts.

Examining the landmark TDM cases for further guidance, neither *Google Books* nor *HathiTrust* address the issue of contractual annulment of fair use. Tantalisingly, in *iParadigms,* the court had the opportunity to examine in detail whether the click-wrap agreement signed by the plaintiffs was enforceable against Turnitin. Nonetheless, the court neglected to do so, putting its failure to rule on the issue down simply to it being "in the light of our fair use analysis".[689] Nevertheless, thereby indicating that fair use may in certain circumstances supersede contractual restrictions. Although not a fair use judgement, more recently the interplay of exceptions with contract law has been explored more fully in *Universal Instruments Corp v Micro Systems Engineering Inc et al.*[690] Concerning the enhancement of medical device handling software belonging to the plaintiff, in this case, the defendants' adaptation of Universal Instrument's code was ruled by the Second Circuit Court of Appeal not to be in breach of an agreement willingly entered into by all parties because the licence was pre-empted by federal copyright law. The exception in question being §117, which renders lawful archiving[691] as well as the adaptation of software by its owner on the proviso it is "an essential step in the utilization of the computer program".[692]

Providing more direct locus to the question of whether state law is pre-empted by §301(a) of the Copyright Act and thus, by extension, whether the exercise of exceptions is permissible,[693] is a recent

[686] See Samuelson, supra 680, 859 *in fine.* Derclaye & Favale, supra 667, 143. Jonathan Band and Masanobu Katoh, *Interfaces on Trial 2.0* (MIT Press 2011) 121. David Nimmer, Elliot Brown & Gary Frischling 'The metamorphosis of contract into expand' (1999) 87 Cal L Rev 17, 60.
[687] Carroll, supra 487, 897, 935, 940.
[688] For example, due to the volume of content licensed and the imbalanced nature of the relationship between a publisher and a university, it is not uncommon for university libraries to be unaware of the contents of a licence.
[689] *Vanderhye v iParadigms* (4th Cir 2009) 645 n.8 (the plaintiffs arguing that Turnitin's archiving of their work was in contravention of contractual terms submitted when signing up to the service.)
[690] *Universal Instruments Corp v Micro Systems Engineering Inc,* No.17-2748 (2d Cir 2019).
[691] §117 (a) 2.
[692] §117 (a) 1.
[693] Ongoing AI cases in the US which implicate contractual override include *J Doe 1* v *Github, X Corp v Bright Data Ltd,* and *New York Times v Microsoft et al.*

Californian case in which X asserted that the scraping of data from its platform was in contravention of its terms of use under state law.[694] Roundly finding in favour of user rights superseding contractual exclusions, the Judge opined that the exercise of fair use in Federal Copyright Act would in all instances "be obliterated by X Corp", thus "flout[ing] Congress's intent" if X's claims situated in unilaterally imposed set of terms and conditions were to succeed.[695]

Despite the potential negative externalities that private ordering may give effect to, the lack of legal clarity on this issue creates a largely unfavourable environment for technological innovation and the democratic paradigm more broadly.[696] Whilst a number of academic commentators have highlighted the undesirable effects of user privileges being overridden,[697] nevertheless the majority of circuit courts still take the mere existence of a contract as a justification for the pre-emption of federal law, and following on from this, the supremacy of most types of contract over copyright law. As shown above, although there are a number of notable exceptions to this, this may even be the case where they are un-negotiated or constitute a contract of adhesion with no meaningful assent – traditionally the cornerstone principle of contract law. Although clear from the case law that many types of non-generative AI are fair use, there nevertheless remains the very real prospect that contracts, standard licences, and in certain circumstances potentially even website terms of use will render fair use and other exceptions in the US ineffective. Elevating contract law above and superior to copyright law – and thereby undermining the balance between public and private interests at its core – poses a threat not only to future innovation and scientific progress, but also as Goldstein cautions, to the "acceptable conditions of cultural and political life [that have sustained us] for more than two centuries."[698]

3.8 Technological Protection Measures (TPM)

Intertwined in the digital age with licensing, technological protection measures if unchecked, create further barriers which subvert the copyright framework and undermine the public policy goals which inhere in limitations and exceptions. As noted by Reichman and Okediji, when "technological fences and electronic contracts are further supported by anti-circumvention measures that forbid decryption or other means of cutting through such fences, the publisher's control becomes virtually absolute."[699] Thus, as has been well documented in the literature, the copyright bargain that balances the rights of authors with the public interest is being further eroded by technological tools.

---

694 *X Corp v Bright Data Ltd,* No. C 23-03698 WHA (ND Cal May 9, 2024).
695 Ibid.
696 See Netanel, supra 659.
697 Samuelson, Band, Katoh, Nimmer, Carroll, supra 686-687.
698 Paul Goldstein, 'Copyright and Its Substitutes' (1997) 45 J Copyright Soc'y USA 151, 152.
699 Jerome Reichman and Ruth Okediji, 'When copyright law and science collide: empowering digitally integrated research methods on a global scale' (2012) 96 Minnesota Law Review 4 1362, 1418.

Numerous types of technological protection measures exist that can frustrate data scientists wishing to undertake machine learning.[700] These range from copy control mechanisms, such as download-blocking technologies, to access-control measures such as passwords and CATCHA technologies, through to firewalls and overly sensitive server intrusion detection systems which mistake legitimate use for piracy and prevent machine learning that is otherwise lawful.[701]

As signatories of the WIPO Copyright Treaty (WCT) and WIPO Performances and Phonograms Treaty (WPPT), both Japan and the US are obliged to provide authors with "effective legal remedies against the circumvention of effective technological measures".[702] In Japan, regulation of copy control technologies takes the form of a mixed framework where penalties are targeted solely against vendors of circumvention software and persons who circumvent TPMs for the purpose of private copying.[703] Thus, the bypassing of copy control technologies for any other user privilege is non-infringing. As a consequence of its ratification of the Comprehensive and Progressive Agreement for Trans-Pacific Partnership (CPTPP) on 6th July 2018, a supplementary regime for access-control measures has now been realised in Japan, thus further extending the legal and technological reach of authorial entitlements. Despite being a suspended clause, and there being no contractual requirement to implement 18.68.1(a) of the CPTPP, all the same in 2018 stipulations against the circumvention of access control measures were enacted as a new addition to the JCA.[704]

Where does this leave a data scientist in Japan wishing to undertake machine learning of works protected by TPMs? As noted above, bypassing copy-protection measures for AI is non-infringing as the only act to which anti-circumvention provisions apply is personal copying. As to circumventing access-control mechanisms, Article 113 (3) of the JCA provides for lawful circumvention for the "purpose of research or technological development in connection with technological exploitation restriction measures, *or otherwise does not harm the interests of the owner of the copyright."*[705] As by virtue of Article 30-4, the JCA has positioned data analysis as an act of non-enjoyment lying exogenous to the goal of copyright, it therefore seems probable that the circumvention of access-control measures for most ML purposes would be judged by the courts to be non-infringing.

Whereas, under Japanese law, circumvention by a data scientist under their own steam is in many instances lawful, nevertheless if it involves software or hardware sourced from a third party, it constitutes an infringing act. Article 120-2 of the JCA establishes both civil and criminal penalties for

---

[700] For example, see Séverine Dusollier's categorisation of TPMs in 'Electrifying the fence: the legal protection of technological measures for protecting copyright' (1999) 21 European Intellectual Property Review 6 285.
[701] See LACA and LIBER survey results relating to TPMs frustrating data analytics in universities: https://libereurope.eu/article/tdm-technical-protection-measures/
[702] Article 11. Obligations concerning Technological Measures. WIPO Copyright Treaty.
[703] Article 30 (1) (ii). Reproduction for Private Use. Article 120-2 JCA.
[704] Article 113 (3) JCA.
[705] Emphasis added.

persons who supply third parties with circumvention technologies.[706] As a consequence, if an AI company or researcher is to bypass TPMs free from interference by the law, circumvention technologies must be developed in-house. As Kohashi notes, in restricting the breaking of digital locks in this way, the JCA creates a contradictory half-way house where circumvention of TPMs for AI purposes is lawful, but the means to enjoy these statutory freedoms are circumscribed.[707] While no doubt in a privileged position compared to AI developers in many other countries, all the same, the requirement to rely on internal resources to circumvent TPMs introduces cognisable time and economic hurdles and thus create an AI innovation penalty that a more ambitious approach to TPMs could have otherwise avoided.

Much like its approach to contracts, the legalities of anti-circumvention in the US are singularly complex. While the Digital Millennium Copyright Act (DMCA)[708] only prevents circumvention of access-control measures, leaving copy-control measures unregulated,[709] like Japan, trading in circumventing technologies is unlawful.[710] However, it is its two-pronged dynamic approach to permitted circumvention that singles the US out as unique in the global copyright system. Unlike Europe and Japan, §1201 of the US Code contains an enumerated list of codified activities which automatically render the bypassing of TPMs as lawful. These include circumvention of digital locks by libraries in the course of deciding whether to acquire a work,[711] law enforcement and intelligence activities,[712] reverse engineering,[713] and encryption research.[714] Supplementing this, in an extremely time-consuming and onerous procedure, whereby a review of further lawful acts of circumvention is undertaken *ab initio* every three years. By dint of §1201(a)(1)(c), the Librarian of Congress is obligated to hold a triennial review of TPMs to establish non-codified cases where protection measures are unduly hampering access to copyrighted works.[715] As a consequence, this obliges applicants whose acts have already been deemed non-infringing to devote additional time and resources to renewing their applications in each subsequent review.

Despite the undeniable stature that fair use holds within US copyright jurisprudence, whether under federal law it is a legitimate defence for circumvention is surprisingly ambiguous. On first blush, §1201 seemingly permits fair use circumventions, as the Copyright Act establishes that “nothing shall

[706] Article 120-2. For providing third parties with the means to circumvent TPMs a court can impose a fine of up to ¥ 3,000,000 (€ 20,500) or a term of up to three years imprisonment.
[707] Kaoru Kohashi「技術的保護手段と著作物の自由利用」641 in Naoki Koizumi & Yoshiyuki Tamura『羽ばたきー２１世紀の知的財産法』(Kōbundō 2015).
[708] Public Law 105-304 (1998).
[709] §1201(a) (1) (A) (2012). See Sag, supra 488 (“The law does not prohibit circumventing a TPM that controls specific uses of a work without denying access altogether.”) 352. Samuelson, supra 680, 860.
[710] §1201 (a) (2).
[711] §1201 (d).
[712] §1201 (e).
[713] §1201 (f).
[714] §1201 (g).
[715] §1201 (a) (1) (B)-(C).

affect … defences to copyright infringement, including fair use."[716] Nevertheless, differing and indeed less favourable interpretations of this section's intent can be pointed to. Some commentators argue that it creates a fair use savings clause,[717] while others credibly suggest that the intent is to seek to clarify that §1201 regulates the trafficking of anti-TPM software and the act of circumvention, and not the *use* of a copyright work once the technical protection measures have been circumvented.[718] Furthermore, this second interpretation has been given weight by the US Copyright Office which in guidance accompanying the enactment of the DCMA expressly stated that fair use is not a defence against circumvention.[719]

Turning to case law concerning ML specifically, whether fair use is a legitimate defence for circumventing access-controls for the purpose of text and data mining was not addressed in *Google Books*, *Hathitrust* or *iParadigms*. Looking more broadly to case law that has abutted on fair use, whether, and in what circumstances, §107 may override anti-circumvention prohibitions, reveals an absence of determinative precedent, as a fair use defence was either not claimed[720] or ruled out of scope.[721] Compared to their Japanese competitors, US businesses wishing without authorisation to circumvent access-control measures for machine learning purposes therefore face a precarious and difficult decision.[722]

The only clear-cut case where circumvention of access-control mechanisms for TDM has been rendered lawful relates to the use of motion pictures and literary works pursuant to the triennial review of permitted exemptions to anti-circumvention measures. Repeating the previously adopted exemption,[723] as of the 28th of October 2024, the bypassing of TPMs for the purpose of data analysis of motion pictures and literary works contained in on-line and off-line media is afforded a non-infringing act if performed by a researcher, student, or IT staff affiliated with a not-for-profit institute

[716] §1201 (c) (1).

[717] See testimony by the Electronic Frontier Foundation on the Digital Millennium Copyright (19/5/2000) (asserting that Congress intended to protect fair use and other exceptions from TPM override.) See also Pamela Samuelson, 'Intellectual property and the digital economy: Why the anti-circumvention regulations need to be revised' (1999) 14 Berkeley Tech LJ 519, 539 ("This seemed to recognize that circumventing a technical protection system for purposes of engaging in fair use or other non-infringing acts would be lawful, although it did not directly say so".) Sag, supra 488 ("The anti-circumvention provisions of the DMCA were not intended to limit or restrict fair use…") 352.

[718] James Boyle and Jennifer Jenkins, *Intellectual Property: Law & The Information Society. Cases and Materials. Centre for the Study of the Public Domain.* (5th Edition, Duke Law School 2021) 566, referencing *Universal Studios Inc v Corley* 273 F 3d 429 (2d Cir 2001) which endorses such a view.

[719] *Digital Millenium Copyright Act of 1988. US Copyright Office Summary* (1998) ("the fair use doctrine is not a defense to the act of gaining unauthorized access to a work, the act of circumventing a technological measure in order to gain access is prohibited.") 4.

[720] *Chamberlain Grp Inc v Skylink Techs Inc,* 381 F 3d 1178, 1203 (Fed Cir 2004) (in a case about automated garage door openers, although Skylink did not claim fair use the court did express the general opinion that §1201 in principle does not invalidate a fair use defence, although it simultaneously declined to clarify when a fair use defence would be ruled as valid.) n.14. See also *MDY Indus LLC v Blizzard Entm't Inc,* 629 F 3d 928 (9th Cir 2010) (in referring to *Chamberlain* stating "§1201 (a) does not limit the traditional framework of exclusive rights created by §106, or defenses to those rights such as fair use.") n.12.

[721] *Universal City Studios v Corley* (formerly Reimerdes) 273 F 3d 429 (2d Cir 2001) (a case relating to the posting of links to film decryption software on mailing lists that facilitated circumvention of TPMs. It is noteworthy that the defendants did not raise a fair use defence regarding their own actions, only referring to possible fair uses by users of the software, which the court failed to evaluate due to a lack of evidence.) 459.

[722] Sag, supra 488, 352 ff.

[723] Granted first on the 28th October 2021.

of higher education.[724] The 2024 enactment expands the 2021 provision by allowing the secure sharing of works to researchers affiliated to other higher education institutions. The authorisation nevertheless contains a number of caveats.

Firstly, "computer programs and compilations that were compiled specifically for text and data mining purposes" are excluded from the authorisation granted by the US Register of Copyrights.[725] Secondly, the use of the term TDM potentially restricts circumvention to instances of data analytics only, thus leaving unanswered whether generative AI – whose ultimate purpose is the production of a creative output – is a non-infringing act.[726] Thirdly, invalidating much potential TDM is the requirement that the content is "*owned* by the institution, or licensed to the institution *without a time limitation* on access."[727] Much online digital content acquired by educational establishments and their libraries is not owned but licensed on a temporary basis, thus rendering the exemption considerably less useful than it otherwise might have proven.

Storage obligations create further burdensome conditions for beneficiaries. Reflective of the court's position in *Google Books* and *HathiTrust,* the Register of Copyrights' Recommendations emphasise the need for secure storage of data. However, rather than leaving the modalities of security to universities, the Recommendations prioritise those which "have been agreed to by all interested copyright owners … and institutions of higher education".[728] It is only in the absence of such measures that universities are permitted to store the works in a manner that "the institution uses to keep its own *highly confidential* information secure".[729] Thus as noted by Carroll the proportionality principles inherent in the first and fourth fair use factors that take into account "the probability of third-party harm and the foreseeable magnitude of its impact" are disregarded in favour of uniformly burdensome obligations.[730]

In conclusion, the commonplace nature of anti-circumvention measures presents data scientists with both practical and legal hindrances. While e-Books, journals, websites, broadcasts, and multiple other forms of digital product provide a wellspring of information for data-driven innovation, the ubiquitous presence of TPMs and contractual terms and conditions can gridlock computational knowledge discovery tools, thus creating poor incentives for both businesses and scientific discovery. In the US particularly, when seeking to extract knowledge from digital materials, investors have to factor these

[724] Section 1201 Rulemaking: Ninth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention: Recommendation of the Register of Copyrights (October 2024) 99-102.
[725] Id. 101.
[726] Id. The Recommendation simply states that the Register of Copyrights "has declined to incorporate language specific to generative AI." 97.
[727] Id. 100 ff.
[728] Id. 101 ff.
[729] "[t]he option for institutions to use the security measures they use to protect their own highly confidential information provides a fallback in the absence of consensus security measures." Id. 79.
[730] Carroll, supra 487, 950.

unavoidable features of the digital landscape into their business decisions.[731] Indeed, to the detriment of the public interest so high and complex are these interwoven legal and technological hurdles it seems unlikely that data analysis and information retrieval projects as ambitious and valuable as *Google Books* and *HathiTrust* would have ever been realised if they had had to contend with born-digital copyrighted works.

Conclusion

Faced with the challenge of updating a copyright framework ostensibly devised for a bygone analogue era, the response of the different jurisdictions subject to this study display distinct asymmetries. Despite the plethora of national investment strategies designed to support AI, paradoxically the UK has outlawed commercial TDM entirely, whilst the CDSM Directive's support of for-profit actors is weak at best. By contrast, relying on open and flexible limitations and exceptions, the legal approach adopted by Japan and the US is one more friendly to both business and science. The United States, as it has so many times before, relies on the ever-developing judicial interpretations of fair use to facilitate AI. While this has cemented automated data analysis, producing an array of output types as lawful,[732] a significant number of cases concerning the legal contours of generative AI works are ongoing and will inevitably shape the future of this branch of machine learning. Meanwhile, in a clear lesson for Europe, the introduction of an open norm in Japan has transformed its once rigid approach to exceptions into one that supports technological advancement and innovation balanced with appropriate safeguards for those who undertake cultural production.

As discussed, extra-copyright barriers to machine learning in the form of contract and technical protection measures can provide a formidable hindrance to AI innovation. Both Article 4 of the CDSM Directive and the US Copyright Act fail to adequately address the near ubiquitous presence in the digital environment of extra-copyright mechanisms which serve to hinder innovation and scientific discovery. Likewise, although since 2014 the UK has outlawed the contractual override of many of its exceptions, and despite no longer being bound by the European copyright acquis, it has failed to capitalise on this by amending the CDPA to allow circumvention of technical protection measures for otherwise lawful acts.

By contrast, Article 7 of the 2019 CDSM Directive offers a clearly codified response to the challenges posed by both private ordering and digital locks, by forbidding contractual and technological override

[731] For example, as of November 2024, X has been granted permission by the Northern District of California to amend their filing against Bright Data to assert that their servers have been improperly accessed as part of a web scraping process.
[732] For example, in *Google Books,* snippets of verbatim text were provided end users. In *Hathi Trust* only word frequencies and positions are provided, while in *iParadigm,* the software identifies the likelihood of similarity with pre-existing academic papers etc.

of TDM undertaken by research organisations and cultural heritage institutions for the purposes of scientific research. Despite the doctrinally significant and ambitious nature of this new addition to the copyright acquis, its real-world impact is nevertheless limited as the Directive provides no practical guidance, timelines, or procedures for permitted circumvention when legitimate data analysis is being blocked.

Although not as clearly articulated as the Directive, in Japan a mix of codified law and authentic interpretational guidance leads to the conclusion that with the potential exception of generative AI works, many forms of machine learning are protected from both contractual and technological pre-emption. Moreover, distinct from Europe, the encumbrances caused to a business from rightsholders' application of TPMs are comparatively few. No lengthy circumvention process involving legal authorities and rightsholders is required because beneficiaries of Article 30-4 are at liberty to develop their own circumvention tools as they say fit.

From the perspective of strategic innovation policy, in contrast to Japan which has gone the furthest to support AI, it is evident that the other countries in this study have failed to undertake a systemic analysis of the broad impact that copyright has on AI markets and scientific advancement. Perhaps most surprising of all, despite America's e-commerce multinationals dominating many aspects of the AI economic landscape, the authorities there to date have left TDM to the vagaries of ongoing judicial rulings, onerous triennial reviews of encryption rules, and conflicting state and federal laws.

While traditionally more interventionist than the US, the response in Europe has been both normatively flawed and technically uneven in its execution. From the inadequate support for commercial entities, to the failure to provide a derogation to the exclusive right of communication to the public, there has been a symptomatic lack of inter-institutional and intra-governmental join up around AI. Whilst the copyright response of western countries to this new transformative technology has been distinctly mixed, by contrast Japan has adeptly navigated the legislative pitfalls to deliver comparatively high levels of certainty for AI businesses based there. Through a combination of statute and authentic interpretation, in amending its once rigid copyright system, it has – at the very least from a legal perspective – been able to deliver what half in jest Ueno refers to as a "machine learning paradise."[733]

[733] For example see Tatsuhiro Ueno『人工知能による"発明"と"創作"』("Invention" and "Creation" by Artificial Intelligence) Japio Year Book 2017, 20.

Chapter 4: Protection of Computer Programs – A Cuckoo in the Copyright Nest

Introduction

Analysis in the previous two chapters shows the varying extent to which copyright exceptions have supported the application of machine learning technologies in the four jurisdictions that make up this study. This chapter now turns to another facet of the role copyright can play in incentivising AI; the jurisprudence associated with the protection of computer programs. This will act as an essential undergirding for the analysis undertaken in Chapter 5, which applies these newly formed tenets of copyright law to the norms of AI software production.

Arising from tensions in the book trade over 300 years ago,[734] and still very much rooted in the norms of an analogue world, software copyright law has from the very start vexed the authorities.[735] In the words of the Court of Appeals of the First Circuit of the United States, the application of copyright law to computer programs resembles "assembling a jigsaw puzzle whose pieces do not quite fit."[736] Furthermore, it seems inevitable that the new norms that are characteristic of AI software production will likely exacerbate the poor fit that commentators and the courts have repeatedly highlighted prior and since 1980 – the year in which the United States became the first country in the world to explicitly afford protection in its copyright act to computer programs.[737]

While it is settled as a matter of law that originality expressed in computer code is subject to copyright, those elements that should remain open and accessible for others to use and reproduce has proven a continual challenge. The aim of this chapter is to outline the prevailing doctrines which have come to govern the scope of copyright as applied to software. First is an evaluation of whether data-dependent AI software, where data literally "writes" elements of the final program, falls under the existing legal definitions of computer programs. Following this, after a brief discussion of the legislative history, the copyright statutes of the US, Japan, and the EU are examined, describing how originality and functionality have been addressed respectively across these different bodies of law. This chapter focuses mainly on these three jurisdictions, as the fundamental tenets of UK software copyright law are derived from that of Union law. The final subpart of the chapter will then focus on

[734] The Statute of Anne (1710) was the result of a consortium of influential stationers who submitted a petition to Parliament stating "divers Persons have of late invaded the Properties of others, by reprinting several Books, without the Consent, and to the great Injury, of the Proprietors, even to their utter Ruin, and the Discouragement of all Writers in any useful Part of Learning." Journal of the House of Commons. 16.240.

[735] For example, most recently a series of rulings (2013-2021) relating to Google's unlicensed use of 37 Java APIs belonging to Oracle has upended US landmark case law from between 1992 and 2004 that had found consensus that interoperability information is not copyrightable. See Mark Lemley, Pamela Samuelson, 'Interfaces and Interoperability After Google v Oracle' (2021) 100 Tex L Rev, 1.

[736] *Lotus Development Corporation v Borland International Inc,* 49 F 3d (1st Cir 1995) 67.

[737] The Computer Software Copyright Act of 1980 Pub. L. No. 96-517, 94 Stat. 3015, 3028 (1980) amended the US Copyright Act of 1976 by introducing a definition of computer programs and defining them as literary works.

norms developed by the courts in the light of the unique set of copyright challenges that computer programs present.

1. Is an AI a Computer Program in Copyright Law?

The prominent role of data, and the revolution in computing that the internet has engendered, has created a very different programming paradigm to that envisaged by lawmakers in the 1970s and `80s. In spite of the central role that data plays in shaping the encoding and architecture of some AI models, nonetheless as a matter of law, machine learning models are still governed by the same statute and case law-based jurisprudence that regulates more traditional computer programs.

To analyse this assertion in more detail, an evaluation of the definitions in different enactments is first required. Significantly, in the interests of interpretational flexibility, both the Copyright, Designs, and Patents Act (1988) as well as Computer Programs Directive 2009/24/EC[738] entirely eschew any definition of a computer program.[739] Thus, subject to the presence of a human author in the EU,[740] adopting a natural language interpretation as the courts in both these jurisdictions may well do, where the work in question constitutes "operating information used by a computer",[741] there is likely to be little barrier to AI software being afforded copyright protection.

By contrast, US law highlights both the functional and executable role of programs in its statutory definition:

Copyright Act of the United States (USC:Title 17) §101:

A "computer program" is a set of statements or instructions to be used directly or indirectly in a computer in order to bring about a certain result."

Thus, unlike other copyright works, a computer program both instructs and effectuates a response. It is the set of instructions in the software code that cause the computer to run and produce a pre-defined "certain result." In the light of the substantial role that data plays in AI, the lack of a subject in the form of *who* or *what* provides the instructions that comprise the program is arguably highly pertinent. While the focus of the debate at the time, and the very reason for making software expressly subject to

[738] The Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs, was amended by 2009/24/EC of 23 April 2009.
[739] Despite the absence of a definition of computer programs in the Directive, Part 1, 1.1.1 of the Commission Proposal for a Council Directive on the Legal Protection of Computer Programs COM (88) 816 does explain that a computer program is a "set of instructions the purpose of which is to cause an information processing device, a computer, to perform its function". It goes on however to advise against the inclusion of a definition in Directive "to avoid it becoming outdated".
[740] See Chapter 2 for a discussion of the requirement for authorship in the EU being dependent on a human author.
[741] Oxford English Dictionary (OUP 2003). "OED". Definition of software.

copyright law, was the centrality of the human programmer viz. author, the formulation of §101 evidently neither negates the important role data plays in establishing the final form of some AI programs, nor places it beyond the scope of US copyright law.

The words *statement* and *instruction* evidently play a pivotal role in the Act's definition of a computer program. At the creational level of the AI software, therefore in common with the UK and EU, the 1976 Act's definition could be argued to also cover AI training data even though at the time this was evidently not the original intention.[742] Although the data provides no more than a mechanism for the AI model to learn for itself how to solve the problem in question, and does not instruct in the sense that it does not give deliberate or pre-determined 'direction [s] or order[s]"[743] in the way that executable code does, data unquestionably as defined by the OED "give[s] information" to which the operations of the computer respond.[744] That is to say, it informs the structure, hyperparameters, and parameters of some AI models[745,746] which in turn direct the model how to operate and bring about a broadly pre-defined "certain result." Thus, just as is the case with human created instructions to a computer that gives rise to prescribed results and actions, the same causation can be found with the data.

More germane to the question of whether in principle the 1976 definition covers AI software code is whether the presence of "data-written" parameters[747] negate the application of the definition as a whole and therefore protection by copyright law to machine learning programs will fall away.

This question arises because, as discussed in Chapter 1, copyright in authorship rests solely in the creative works and contributions of human authors. It is not afforded to the works of any other agent, whether animals,[748] non-terrestrial beings,[749] computer programs[750] or as is the case here, inanimate

---

[742] Interestingly, in one Japanese case the court was faced with the issue of whether stand-alone data (in this case numerical values) would fall under the definition of a computer program as set out in the Japanese Copyright Act. Unclear if values by themselves that cause a computer to perform a certain function would constitute a computer program, the question was avoided by concluding that in combination with human-written code they would reach the appropriate threshold to be covered by the definition. *Trainline Design Program Case*「電車設計用プログラム事件」Tokyo District Court Heisei 15/1/31 Hanji 1820 / 127 (2003).
[743] Definition of instruction. OED.
[744] Id. "give information to".
[745] In an SVM model for instance, the data affects the position of the hyperplane, and as a result determines the results of classification in the generalised model. In an Artificial Neural Network, the data shapes how the neurons and layers interact with each other, and therefore how critically important functions in the program perform.
[746] Strictly speaking not all AI models abstract from and generalise the data they process. A strict definition of machine learning would exclude instance-based learning models like K-nearest neighbour and Lazy Bayesian Rules. In these models nothing is learnt, and generalisation does not occur as the model simply stores the data verbatim. See Brett Lantz, *Machine Learning with R* (Packt Publishing 2019) 75.
[747] For example, weights in a neural network, the support vectors in a support vector machine, the coefficients in linear and logistic regression.
[748] *Naruto et al v David John Slater et al,* Case No 3:15-cv-04324-WHO (a much referenced series of cases in the US regarding a selfie taken by a macaque, which in line with copyright jurisprudence resulted in a ruling denying the monkey copyright.)
[749] *Cummins v Bond*, 1 Ch 167 (1927) (a case involving the writings of a medium who purported to be channelling the spirit of a follower of St Paul. The court rejected the defendant's claim that the psychographic work could vest in a someone "domiciled on the other side", instead granting copyright to the plaintiff.)
[750] *Stephen Thaler v Shira Perlmutter,* No. 22-CV-384-1564-BAH (18/8/2023) (a case upholding the decision by the US Copyright Office to reject an application for copyright in artwork which was produced by the software belonging to the plaintiff which was described by him as "autonomously created by computer algorithm running on a machine".)

data. Where elements of the machine learning code are created by the data, or indeed other computer programs,[751] the question of whether copyright arises will in most jurisdictions[752] centre on the role of the human author *qua* programmer. At the present time, however, in spite of exaggerated discussions of artificial consciousness,[753] even the most cutting-edge AI programs that result from generalisation will constitute a synthesis of both human-written and data-written code.[754] Data is not the sole actor in the production of a machine learning model, but crucially from a copyright perspective, it works in tandem with human-programmed code, which if original and creative, is undeniably the subject matter of exclusive rights. In combination, they provide a set of executable instructions and orders for the computer grounded largely in human creativity. Thus, as will be examined in this chapter and the next, although which elements of a particular program will enjoy protection will depend on copyright jurisprudence and the facts at hand, *ab initio* the role of data in shaping the make-up and function of the program intertwined as it is with human programmed code will in no sense invalidate the applicability of the §101 definition to a machine learning computer program in its aggregate form.

2. A Brief Legislative History – Fitting the Proverbial Square Peg in a Round Hole[755]

From the 1960s onwards,[756] whilst law makers increasingly recognised the protectability of software, historically speaking, the decision to bring computer programs into the copyright fold nevertheless represented a radical departure from intellectual property traditions.[757] Although first grounded in the norms of late 17th and early 18th century book trade, over the centuries, copyright has expanded to cover virtually any creative work of an aesthetic or informational nature including photographs, film, and music. Thus, copyright protects works which create, in the minds of the audience, a sense of enjoyment, whether of an intellectual or more aesthetic and sensual nature. The contrast with computer programs could not be starker. As technical devices, they have wide industrial application, and unlike artistic and literary works, computer code's role is not to be read or to enjoy and entertain but to perform and function.

Unlike artistic works – which speak to their audience at an emotional level – computers are functional and utilitarian machines. Other copyrights are aimed at a human audience and are intended to "be

[751] *J Doe 1 et al v GitHub Inc et al*, ND Cal No. 3:22-cv-06823 (3/11/2022) (a case against GitHub, Microsoft and OpenAI for failure to attribute the authors of software used in the training of the assistive programming products Codex and Copilot.)
[752] In the case *of works generated* by a computer, as outlined in Chapter 2, a number of common law jurisdictions including the UK, Ireland, India etc have afforded copyright to the agent who has made arrangements leading to the creation of the work.
[753] Richard Luscome, "Google engineer put on leave after saying AI chatbot has become sentient" (*The Guardian*, 12 June 2022) <https://www.theguardian.com/technology/2022/jun/12/google-engineer-ai-bot-sentient-blake-lemoine>
[754] See Karpathy. op. cit. 53.
[755] In the landmark US case *Computer Associates International Inc v Altai Inc,* 982 F2d 693 (2d Cir 1992) the Second Circuit expressed the challenges for courts of applying copyright norms to "a highly functional, utilitarian component in the larger process of computing" using the well-known metaphor of fitting a square peg into a round hole. (712).
[756] The US Copyright Office first registered computer programs under the "rule of doubt" procedure as a copyright work in 1964: *Copyright Registration for Computer Programs – Announcement from the Copyright Office,* reprinted in 11 Bull. Copyright Soc'y USA 361 (1964).
[757] Dennis S Karjala, 'The Relative Roles of Patent and Copyright in the Protection of Computer Programs' (1998) 17 J Marshall J Computer & Info L 41, 47.

read, heard, or seen, for either pleasurable or practical ends",[758] something that self-evidently cannot be said for computer code given its very limited audience. Rather, information is stored and conveyed by magnetic polarity, electrically charged particles,[759] and electric signals concealed from view. These signals are represented at the most granular level by object code: the zeros and ones of binary code, which are derived from, as well as translatable into, human-readable source code.[760] As "machines that have been constructed in the medium of text",[761] the prime role of the software is to control the electromagnetic signals of a computer.

Historically speaking, this is not the first time that the lack of a human audience has provoked doubts about the suitability of technically oriented works for copyright protection. Around 120 years ago, in the both the UK[762] and the US,[763] perforated pianola sheets were rejected by the courts as being appropriate copyright subject matter on the grounds that they were "not addressed to the eye" and constituted "only parts of the machine."[764]

Due to its abundantly different and primarily functional nature, the decision to codify a type of machine into copyright law ineluctably led to an array of doctrinal challenges. From the outset, the customary divide between patents, which are granted to apparatuses and devices, and copyright works that arise from creative and oftentimes artistic endeavours, has predictably acted as a faultline across the norm setting activities of both the legislature and the courts. One major consequence of this has been that the authorities have been forced to elaborate repeatedly not just on the contours of the exclusive rights, but also on instances as to whether copyright should regulate elements of the code at all. Programming languages, methods of operation and algorithms form several of the areas that the legislature has decided to explicitly shield from the scope of protection, and which find few equivalents in the precepts that govern more orthodox copyright works.[765] Likewise, particularly in the US, this uneasy fit led the US courts to further refine pre-existing doctrines in order that the functional and ideational elements of computer programs could be copied by competitors. In this way, acting as an incentive to a strategically important and growing industry.

The 1978 Report of the National Commission on New Technological Uses of Copyrighted Works (CONTU) paved the way for the protection of computer programs in the US Copyright Act. Symbolic

[758] *Final Report of the National Commission on New Technological Uses of Copyrighted Works* (National Commission on New Technological Uses of Copyrighted Works 1978) (hereafter CONTU). Dissent of Commissioner Hershey.
[759] NAND technology does not require power as it relies on electrically charged particles to record the binary data values computers use.
[760] Source code and object code are no more than approximations of each other. For example, the sequence and structure of the two sets of code may differ, and object code will not contain the annotations of computer programmers.
[761] Pamela Samuelson et al, 'Manifesto concerning the Legal Protection of Computer Programs' (1994) 94 Colum L Rev 2308, 2316.
[762] *Boosey v Whight*, Court of Appeal, 13 December 1899 [1900] 1 Ch 122.
[763] *White-Smith Music Publishing Company v Apollo Company*, 209 US 1 (1908).
[764] See Maurizio Borghi and Stavroula Karapapa, *Copyright and Mass Digitization* (OUP 2013) 53-6.
[765] While human languages are like programming languages not subject to copyright, as far as this author is aware, lawmakers have not found it necessary to explicitly clarify this in statute.

of the doctrinal challenge that instantiating machines in the form of text into copyright law presented, CONTU was not unanimous in its recommendation that copyright should protect software. Two of the thirteen commissioners alive at the time the report was submitted to Congress issued opposing opinions.[766] Commissioner Hersey, in his 24 page dissent, gave particular attention to, in his view, the fallacious analogy underpinning the report: that because computer code was a form of text, like a book or a poem, it should be protected by copyright. In rejecting copyright law as an appropriate mechanism for legal protection, he gave particular prominence to the functional and executable nature of computer code. Hersey drew a distinction between other copyright works that can be sensed and enjoyed, and programs which have "no purpose beyond being engaged in a computer to perform mechanical work."[767] Memorably, in highlighting his deep-seated jurisprudential concerns, Hersey drew on the constitutional division between copyrights and patents when he reminded the Committee that "printed instructions tell how to do; programs are able to do."[768]

At the same time as CONTU, similar questions around the desirability of including computer software as a copyright work under the Berne Convention were being examined by the World Intellectual Property Organisation (WIPO). Although never adopted, the result was WIPO's 1978 Model Provisions on the Protection of Computer Software. Modelled in no small part on copyright, the model law nevertheless proposed a *sui generis* right lasting no more than 25 years from first creation of the program, and in the case of it being made available to the public, a protection period of no more than 20 years.

Despite the many reservations expressed by legal experts globally, starting in 1980 with the US, governments around the world began to explicitly bring programs into the ranks of works protectable by copyright law.[769] This codification arguably reached its ultimate conclusion during the Uruguay Round of the General Agreement on Tariffs and Trade in 1994. As a result of intense lobbying by the United States government and its allies, and strongly assisted by vocal US business interests,[770] the Agreement on Trade-Related Aspects of Intellectual Property Protection (TRIPS) was incorporated into the 1994 Marrakesh Agreement, the founding document of the World Trade Organisation

---

[766] Commissioner William S Dix having passed away before the final report was submitted, it is interesting to note that in Commissioner Karpatkin's dissent supporting that of Commissioner John Hersey, she wrote "the late Commissioner Dix … indicated that he shared them [i.e. Commissioner Hersey's views] as well." 37.
[767] CONTU, supra 758, 70.
[768] Id. 71.
[769] For example, the UK a private member's bill (Copyright (Computer Software) Amendment Act 1985) expressly recognised a computer program as a work of copyright. The same year West Germany followed suit via the Copyright Amendment Act of 1985. Similarly in Japan, coming into force on the 1st of January 1987, the 1985 amendment to the Japanese Copyright Act (Act No 48 of 1970) brought computer programs into scope. By contrast, perhaps influenced by the WIPO Model Provisions, the French Copyright Amendment Act of 1985 granted related rights protection to computer programs.
[770] See Peter Drahos and John Braithwaite, *Information Feudalism* (Earthscan 2002) (an exhaustive account of how American corporations were able to make strong IPR a top objective of US foreign policy). Susan Sell, 'Industry Strategies for Intellectual Property and Trade: The Quest for TRIPS, and Post-TRIPS Strategies' (2002) 10 Cardozo J Int'l & Comp L 79, 86 (outlining the lobbying activities of American industry groups such as the International IP Alliance in the run up to the adoption of TRIPS.)

(WTO). With TRIPS defining computer programs as a literary work under the Berne Convention,[771] in one fell swoop 128 countries[772] were legally bound to treat computer programs as copyright works. At a time when WIPO member states only comprised 104 countries,[773] this represented an unambiguous watershed moment for the protectability of software globally.[774] No longer were the technology exports of the US and other developed countries subject to unclear legal protection in local courts; copyright, with its long duration and broad monopolistic scope of protection, had become the foremost tool for regulating computer programs internationally.

Strategically this was the result of well-laid plans by US trade negotiators who, in the run up to the Marrakesh Agreement, *inter alia* had prioritised the technology industries and copyright in their negotiations.[775] TRIPS signified a material ratcheting up of penalties for countries whose governments and courts were weak on IPR enforcement, a prime US focus given the high rates of software piracy reported at the time.[776] According to the Business Software Alliance, who were extremely active in their support of TRIPS,[777] worldwide 50%[778] of all software being used in the early 1990s was pirated, with an estimated lost sales value of $12.8 billion.[779] This, combined with concerns of its own industrial decline in the face of a burgeoning technology trade deficit with the countries of East Asia in particular,[780] led the US to protect its IP intensive industries as a matter of international industrial policy, with copyright being of particular strategic importance for a number of different reasons.

Firstly, copyright law is subject to international agreements, and subsequent to the adoption of TRIPS into the global trading framework, one now subject to a dispute settlement procedure at the WTO, bringing with it the potential for meaningful economic penalties. By creating an international legal norm backed up with sanctions for states that turned a blind eye or who lacked capacity to deal with software infringement, the US government was able to complement the domestic action of companies pursuing infringements through local courts with a potentially more effective state-to-state dispute resolution process overseen by the WTO's Dispute Settlement Body.

---

[771] Article 10.1 "Computer programs, whether in source or object code, shall be protected as literary works under the Berne Convention (1971)."
[772] See <https://www.wto.org/english/thewto_e/gattmem_e.htm>
[773] See <https://www.wipo.int/edocs/pubdocs/en/intproperty/120/wipo_pub_120_1994_01.pdf>
[774] Unlike Article 33 of the Berne Convention which only provides for de facto voluntary referral of disputes to the International Court of Justice, Article 3.3 of Marrakesh Agreement Establishing the World Trade Organization and Annex 2 (Understanding on Rules and Procedures Governing the Settlement of Disputes) allows a signatory to report another country for non-compliance with TRIPS obligations. If the Dispute Settlement Body establishes non-compliance, trade sanctions against the non-complying state are then held to be permissible.
[775] Drahos and Braithwaite, supra 770, 73, 85-107. Chapter 11 on the trade agreement related advocacy of the US computer and entertainment industries.
[776] Sell, supra 770.
[777] See *Potential Impact on the U.S. Economy and Industries of the GATT Uruguay Round Agreements: Investigation No. 332-353. Volume I* (Chapter 48 VII-19, United States International Trade Commission 1994).
[778] *Eighth Annual BSA Global Software Piracy Study – Trends in Software Piracy 1994-2002* (The Software Alliance 2002) 1.
[779] Moshe Givon et al, 'Software piracy: Estimation of lost sales and the impact on software diffusion' (1995) 59 Journal of Marketing 1 29, 29.
[780] In the 1970s and 1980s Japan was the main hardware and software competitor to the US. In the face of severe and politicised trade friction resulting from the large trade imbalance between the two countries, Japan ultimately caved into US demands and abandoned its own *sui generis* proposals for the protection of software. (Ministry of International Trade and Industry proposals were made in 1972 and 1983.) See Nobuhiro Nakayama『ソフトウエアの法的保護』(Legal Protection of Software) (Yūhikaku 1990) 11-16.

Secondly, in establishing a global norm, it prevented other countries from either not protecting US computer imports at all by IP law, or introducing legal regimes that in the eyes of the US unduly benefitted domestic competitors. The US, having decided to protect programs in 1980 by copyright law, had created a comparatively high benchmark within its own borders for American businesses, leaving its international competitors free to operate within a more flexible IP environment of their own choosing. This lack of legal harmonisation resulted in a situation where, in the 1980s domestically, American firms received protection for their software of 75 years,[781] but once exported, the duration of protection in a vast number of countries remained opaque. Consequently, the American government was keen to avoid situations arising like the *sui generis* proposals made by the Japanese government[782] and WIPO, both with relatively short durations of protection, which in theory would have resulted in increased competition against US software exports. From the perspective of public welfare, irrespective of what an appropriate duration for computer programs may be, as a consequence of CONTU the US had already decided on a long copyright duration and as a matter of industrial competitiveness it was highly desirous that other countries were made to follow suit.

3. The Protectability and Non-Protectability of Software in Legislation

Having briefly outlined the historical background to software's incorporation into copyright law, this section now turns to the norm-setting activities of the legislature and the courts, much of which took place between 1978 and 1992.[783] First, the response from the legislature will be examined, starting initially with the treatment of ideas and processes in statute. The subsequent section will then proceed to outline the treatment afforded by the courts to copyrightable expression, including the development of interrelated doctrines that act as tools to demarcate protectable from non-protectable subject matter.

While challenges such as ensuring that competition in the technology industry is not unduly hampered can find parallels in the past, ensuring that societally beneficial activities, such as device interoperability, were entirely new copyright challenges. Faced with these questions, two sets of responses from the legislature can broadly be identified. First, was the perceived need to explicitly delineate protectable matter from that which is non-protected, arguably reflective of the fact that the functioning of technology markets has an economic and social impact far beyond that of more traditional copyright works. Not an explicit feature of the law's treatment of other copyright works, this work-specific delineation was deemed essential to ensure that the very building blocks of

[781] Assuming that that the vast majority of software produced will be subject to the work for hire principle under §302 of the Copyright Law of the United States.
[782] Nakayama, supra 780, ibid.
[783] This period represents the span between CONTU, the 1991 Computer Programs Directive and a significant number of the landmark software cases in the US including: *Plains Cotton Co-op v Goodpasture Computer,* 807 F 2d 1256 (5th Cir 1987), *Computer Associates International v Altai,* 982 F 2d 693 (2d Cir 1992) and *Atari Games Corp v Nintendo of America Inc,* 975 F 2d 832 (Fed Cir 1992), *Sega Enterprises v Accolade Inc,* 977 F 2d 1510 (9th Cir 1992).

technological innovation were safeguarded from the negative welfare effects arising from an excessive monopoly control. The second more common response, paralleled in all other copyright works, concerns the boundaries of exclusive rights and the introduction and exercise of accompanying limitations and exceptions.

3.1 The Idea-Expression Dichotomy

Although the extent to which software would go on to control people's lives could arguably not have been imagined at the time, nonetheless, from the very beginning, policy makers were aware of the imbalanced incentives that the over-protection of computer programs could give rise to. A situation where competitors were unable to use a particular programming language, undertake analysis of code, reproduce functionality, or access information in the public domain, would not only have acted as a significant barrier to technological innovation but impacted negatively on societal advancement. Taking just one example of this, as can be seen in the development of the Android platform,[784] awarding overly broad copyright protection to code could result in the stymying or prevention of entirely new marketplaces from emerging worth many billions of dollars.[785]

From the standpoint of copyright, the legal challenge was to define an appropriate distinction between protectable and non-protectable elements of the computer code for a type of work entirely different in its nature to other subject matter. Unlike patents, which protect the real-world application of ideas by preventing others from replicating the same original methods and processes, copyright leaves the ideas and concepts encapsulated within original expression free for others to use and build upon. The reasoning for this is that by leaving the ideas of an author free for others to utilise, it engenders further creativity. This doctrinal division between protectable expression on the one hand, and non-protectable idea on the other, commonly referred to in the literature as the "idea-expression dichotomy", is a cardinal principle of copyright law rooted in both civil[786] and common law[787]

[784] *Google LLC v Oracle America Inc,* 141 S Ct 1183 (2021). The second appellate ruling by the Federal Circuit determined that Google's use of Oracle's code in its software (referred to as "declarations" in the proceedings) was an infringement. The Supreme Court (2021) subsequently reversed the decision based on a fair use analysis. The court concluded that the use by Google of 11,500 lines of declarative code specifically chosen to facilitate Android developers familiar with Java to call up functions written by Google, was transformative as it facilitated the ease with which the new Android mobile platform could grow as a new market. As part of its summing up the Court justified its conclusion by relying on public welfare arguments including a quote from the Amicus Curiae brief from Microsoft: "[A]llowing reasonable fair use of functional code enables innovation that creates new opportunities for the whole market to grow." 27.
[785] Id. The 2021 ruling states that "As of 2015, Android sales produced more than $42 billion in revenue" for Google, 9.
[786] See Maurizio Borghi, 'Owning Form, Sharing Content: Natural-Right Copyright and Digital Environment' in Fiona MacMillan (ed), *New Direction in Copyright Law* (Edward Elgar 2007) (discussing the different terminologies for idea-expression used in 19th and early 20th century civil law courts such as "concept", "shaped thought" on the one hand and "form and content" "expressed thought" and "form of expression" on the other.)
[787] For example, in the UK: *Jefferys v Boosey* 1 0 ER 681 (1854) IV HLC 869 "the claim [to copyright] is not to ideas, but to the order of words, and that this order has a marked identity and a permanent endurance." *Pike v Nicholas* (1869) L.R.5 Ch.App.251. Vice-Chancellor Sir W M James stated there is "no monopoly in the main theory of the Plaintiff, or in the theories and speculations by which he has supported it." In the landmark US case *Baker v Selden* 101 US 99 (1880) relating to whether book-keeping systems and processes were subject to copyright, the court ruled "The very object of publishing a book … is to communicate to the world the useful knowledge which it contains. But this object would be frustrated if the knowledge could not be used without incurring the guilt of piracy of the book." In 2012 the idea-expression doctrine was ruled a First Amendment safeguard by the Supreme Court. *Golan v Holder* 565 US 302 (2012).

traditions. Acting to underline its role as a cornerstone of much copyright jurisprudence, it is explicitly enshrined at the international level in differing forms within the Berne Convention,[788] TRIPS,[789] as well as the WCT.[790]

Turning first to the United States, doctrinal concerns surrounding the application of copyright law to computer programs were already apparent even prior to CONTU. Evincing of this being the 1964 decision by the US Copyright Office to limit the registration of software to the "rule of doubt procedure".[791] Moving forward another ten years, given the centrality of the idea-expression dichotomy to copyright jurisprudence, it is not surprising that it formed a key theme running through the Congressional hearings preceding the enactment of the 1976 Act. Several leading academics and representatives of the education sector[792] expressed concern that extension of copyright to software would ineluctably result in the unwanted protection of both ideas and process on account of the fact that the coded expression contained in computer programs is inextricably entwined with the very function of the computer. This, they argued, would go against not only centuries of jurisprudence, but in the United States, the constitutional division between patents and copyright.[793]

In direct response to these concerns,[794] Congress enacted §102(b) as part of the 1976 Copyright Act, which sets forth the division between copyrightable and non-copyrightable matter as follows:

§102(b) In no case does copyright protection for an original work of authorship extend to any idea, procedure, process, system, method of operation, concept, principle, or discovery, regardless of the form in which it is described, explained, illustrated, or embodied in such work.

Situating the background to this newly enacted provision firmly in copyright jurisprudence, the US Copyright Office stated its purpose was to underline the fact that the "basic dichotomy between idea and expression remains unchanged".[795] Viewed by leading commentators as rooted in the seminal nineteenth century Supreme Court case *Baker v Selden,*[796] §102 (b) can be said to have codified for

---

[788] Although the Berne Convention does not explicitly rule out protection of ideas, implicitly its focus on expression in Article 2.1 suggests that ideas are not subject to copyright. See *Guide to the Berne Convention for the Protection of Literary and Artistic Works-Paris Act 1971* (WIPO 1978).
[789] Article 9.2, TRIPS. (Relation to the Berne Convention) "Copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such."
[790] Article 2, WIPO Copyright Treaty.
[791] Supra 756.
[792] See Pamela Samuelson, 'Why Copyright Law Excludes Systems and Processes from the Scope of Its Protection' (2007) 85 Tex L Rev 1921 (outlining the US legislative history leading up to the introduction of §102(b) and the pivotal roles that Prof Arthur Miller, Professor Benjamin Kaplan and W Morton Brown Jr representing EDUCOM had in its inception.)
[793] Article I, Section 8, Clause 8, of the United States Constitution grants Congress the power "To promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their respective writings and discoveries."
[794] See *1976 House of Representatives Report* 94-1476 (1976) 57.
[795] Chapter 1, 2. Quoting from the 1974 Senate Judiciary Committee Report (no full citation provided). *Second Supplementary Report of the Register of Copyrights on the General Revision of the U.S. Copyright Law: 1975 Revision Bill October – December 1975* (United States Government 1975).
[796] See §2A.06. David Nimmer, *Nimmer on Copyright* (Matthew Bender & Company Inc 2021).

the computer age not only the constitutional division between copyright and patents, but the doctrinal divide between protectable expression and non-protectable idea.

Turning to Europe, it is perhaps no surprise that the sole explicit referral to the non-protection of ideas in the copyright acquis appears in the Computer Programs Directive. Article 1(2) unequivocally states that "*ideas and principles which underlie any element of a computer program … are not protected by copyright*".[797] Examination of the Directive's Explanatory Memorandum reveals that, much like the United States, the Commission was keen to ensure that the independent development of computer programs using the ideas of a third party should not to be troubled by claims of copyright infringement. While jurisprudentially this is self-evidently correct, in terms of application of the idea-expression principle it is interesting to observe that the Memorandum's focus is comparatively narrow. Rather than pointing European developers to the broad innovational benefits of competition in the marketplace that can arise from the free flow of ideas and other non-protected copyright matter, the Memorandum concerns itself solely with interoperability.

3.2 The Process - Expression Dichotomy

Intertwined with, and frequently undifferentiated from the idea-expression dichotomy is what Englund has referred to as the "process-expression dichotomy".[798] In the US, coming from a line of cases that broadly define the contours between patent and copyright law, the process-expression dichotomy denies protection to processes, procedures, and methods.[799]

Thus, while the Copyright Office stated that through the promulgation of §102 (b) it wished to underline that the idea-expression dichotomy remained unchanged, as Nimmer implies *stricto sensu* this does not bear up to scrutiny.[800] This is because four of §102 (b)'s eight enumerated exclusions constitute not ideas, but functional processes and procedures.

Underlining that the process-expression dichotomy is conceptually autonomous from the more well-known idea-expression dichotomy, Nimmer points to what he calls the "non-surplusage canon" of law.[801] That is to say, Congress's conscious decision to treat each of the eight categories of 102(b)

---

797 In the UK, the only explicit reference to ideas in the CDPA comes as a result of the transposition of the Computer Programs Directive.
798 See Steven Englund, 'Idea, Process, or Protected Expression?: Determining the Scope of Copyright Protection of the Structure of Computer Programs' (1990) 88 Michigan Law Review 866, 877.
799 For example, *Baker v Selden,* 101 US 99 (1879) (opining that "the methods of an art are the common property of the whole world" thus denying protection to systems, methods of operation and processes.). *Tilghman v Proctor*, 102 US 707, 728 (1881) ("A machine is a thing. A process is an act, or a mode of acting. The one is visible to the eye … [t]he other is a conception of the mind.") *Burk v Johnson*, 146 F 209, 213 (8th Cir 1906) (concluding that mutual burial association's business methods are not protected by copyright.) *Seltzer v Sunbrock*, 22 F Supp 621, 630 (SD Cal 1938) (ruling a system for a roller-skating contact sport known as "roller derby" was not protectable).
800 Nimmer, supra 796, §2A.06. 2[a].
801 Ibid.

separately is confirmed by an accepted canon of statutory interpretation that every word of statute should "be given meaning and not be treated as superfluous."[802]

Adopting a teleological approach, relegating ideas to the public domain while continuing to protect methods of operation would, in practice, foreclose the use of the very ideas upon which such processes are founded. In the context of computer programs, this would mean that once a process has been adopted by a software developer, no competitor could use that same process in their own products. It goes without saying that such an approach would have been highly detrimental to the emergence of a functional and competitive software market.

While unlike the 1976 Act, which explicitly removed process from the scope of copyright law, there is on first blush a less evident exclusion in the statutes of the EU and the UK. Prior to 2007,[803] in the United Kingdom, this appears to have caused some significant confusion as to whether methods and procedures were protectable matter or not. A 1994 report[804] from the UK Patent Office (now UK Intellectual Property Office), reflecting the predominant views of 150 practising lawyers and IP experts, observed that while unprotectable function was a feature of US law, this was not necessarily the case in the UK. Referencing *Ibcos Computers*[805] and *Anacon,*[806] the report states "there was a reluctance to accept that we should be limiting the ambit of copyright law to take functional material out of its scope, and … it was considered that … TRIPS did not require us to do this."[807] Despite the rather unexpected view from attendees concerning the UK's explicit legal obligations under Article 9.2 of TRIPS,[808] a close reading of the 1991 Computer Programs Directive reveals as a matter of law processes are not protected. Article 5(3), though narrowly circumscribed in that it permits a person, solely in the course of running a computer program, to "observe, study or test the *functioning* of the program in order to determine the idea and principles which underlie any element of the program",[809] nonetheless, signals that processes as such are not intended to be protected by the Directive. Additionally, Recital 10 and Article 6 permit reverse engineering for the purposes of accessing and evaluating the functional elements of the code which allow interoperability. The justification being that the parts of a computer program that permit interoperability with other programs comprise a "*functional* interconnection and interaction".[810] Thus, while nowhere near as explicit an exclusion as

[802] Ibid.
[803] Infra 812.
[804] *Legal Protection for Software-Related Innovation. A Report of the Public Forum held at the United Kingdom Patent Office on 19 October 1994* (Patent Office 1994) 5.
[805] *Ibcos Computers v Barclays Mercantile Highland Finance* (1994) FSR 275. Jacob J holding that even if functional elements of a program could be expressed in one or a limited number of ways it still could be protected.
[806] *Anacon Corp Ltd v Environmental Research Technology Ltd* (1994) FSR 659. Jacob J ruling that an image of the connectivity of electronic components in a circuit board was subject to copyright and therefore the actual reproduction of the connectivity was an infringement.
[807] Supra 804. Ibid.
[808] Article 9.2 TRIPS ("Copyright protection shall extend to expressions and not to ideas, procedures, methods of operation or mathematical concepts as such.")
[809] Article 5 (3) Computer Programs Directive (Emphasis added).
[810] Id. Recital 10.

that provided by §102(b), and putting aside the thorny issue of how effective the Directive is at facilitating access to functional elements of the code,[811] the Directive nevertheless reveals that in principle functional aspects of a program are as a matter of law to be unprotected.

Irrespective of the need for a detailed analysis of the Directive's text, the question of whether function is protectable has been subsequently addressed decisively a number of times by both the UK courts[812] and the CJEU. In *BSA,* a decision concerning components of a graphic user interface, the CJEU held that "where the expression …is dictated by … technical function, the criterion of originality is not met."[813] Highly reminiscent of the US doctrine of merger,[814] where the idea is said to have merged with the expression, the Court denied the application of exclusive rights not just to the processes and functions of the software, but also to the expression directly determined and prescribed by the function. In other words, where expression is not original (as it is dictated by the processes inherent in the software or the hardware), the source code resides in the public domain. Again, two years later in *SAS v WPL*, in a case concerned with emulated software, the Court opined that functionality does not "constitute a form of expression of th[e plaintiff's] program and, as such, are not protected by copyright in computer programs".[815] In doing so, albeit nearly two decades after the Directive first entered into law, any lingering questions in the European Union around the protectability of function were decisively addressed.

The reason why both the EU and the US, and thereafter ultimately TRIPS, all concluded it was desirous to explicitly and broadly codify in legislation the boundaries of protectable and non-protectable rather than leave it entirely to the courts, in large part reflects the poor fit with copyright law software has had since the start. Unlike other copyrightable material where the ideas, concepts, and unprotectable elements contained in a work are freely accessible, commercially distributed software buries its contents deep within the object code. This is because programs are essentially industrial and functional devices displaying many similarities with patentable machines and not works intended to entertain or enlighten. From a doctrine perspective, as pithily remarked on by CONTU Commissioner Hersey, like other machines, "programs are able to do."[816]

---

[811] Other than for the purposes of decompilation (Article 6), reverse engineering of computer programs unless by implication required to make a back-up copy (Article 5(2)) is an infringement under the Computer Programs Directive. Article 5(3) only permits the lawful user of the software to "observe, study or test" if done while "loading, displaying, running, transmitting, or storing the program."

[812] *Navitaire Inc v Easyjet Airline Company and Bulletproof Technologies Inc* (2004) EWHC 1725 (Ch), ruling that replicating the functional aspects of an airline booking system was lawful. (In this case the defendant had no access to the plaintiff's code.) See also *Nova Productions Limited v Mazooma Games Limited & Others* (2007) EWCA Civ 219 determining that the use of ideas that appear as functional elements of a computer game are not subject to copyright.

[813] *BSA,* 49.

[814] Infra subsection 4.2.

[815] Case C-406/10 *SAS Institute Inc v World Programming Ltd Case* [2012] ECLI:EU:C:2012:259 (hereinafter *SAS v WPL*).

[816] CONTU, supra 758, 71.

Expressly clarifying in statute that the ideas and functions which inhere in software are not protectable by copyright served to answer some of the numerous questions that hung over the introduction of software programs into a body of law hitherto associated with expressive works overwhelmingly connected with the arts. In conclusion, therefore, perhaps what is most significant is not that there appear to be clearly distinct stratifications of jurisprudential thinking surrounding the idea-expression dichotomy,[817] but that the incorporation of computerised text-based machines into copyright law demanded a marked recalibration of approach and design by the legislature. As the chapter will later outline, the decision not to leave software to protection by patent law or, as proposed by WIPO, a highly customised *sui generis* right, has had long-lasting and as yet unresolved ramifications for the courts.

3.3 Algorithms – Idea, Expression and Process

In the light of how germane algorithms are to any discussion of AI, this subpart shall examine how they traverse the boundary between idea, process, and expression. The OED defines an algorithm broadly as "a *process* or *set of rules* to be followed in calculations or other problem-solving operations".[818] Knuth's definition, however, is arguably more instructive as the imperative nature of an algorithm is more clearly apparent: "a finite set of rules that gives a sequence of operations for solving a specific type of problem."[819] Whilst both these definitions are illuminating and undoubtedly accurate, the reality is that the word *algorithm* has a wide array of meanings[820] in both public discourse, computer science, as well as in intellectual property law. As aptly expressed by one computer scientist "the fact we have an intuitive notion of what an algorithm is, does not mean that we have a formal notion."[821] Commenting on the confusingly large number of ways the term is used, Hill observes "we see evidence that *any* procedure or decision process … can be called an 'algorithm' in the press and in public discourse."[822] Even computer scientists themselves blur the lines between it, as a step-by-step formula – and therefore in the abstract no more than an idea how to solve a particular problem – and its concrete expression in the computer code. Computer engineers will even use the term to refer to the entire software program and machine learning model. To substantiate this, Mittelstadt et al have pointed out that, irrespective of the discipline or context, often "the term is used not in reference to the algorithm as a mathematical construct, but rather the implementation and interaction of one or more algorithms in a particular program, software or information system."[823]

[817] See Samuelson, supra 792, 1923 (stating that given the existence of the idea-expression, process-expression and the fact-expression dichotomies "[t]o be more consistent with §102(b), courts would be well advised to speak of the "protectable/unprotectable distinction" in copyright law.")
[818] Emphasis added.
[819] Donald Knuth, *The Art of Computer Programming: fundamental algorithms* (Addison Wesley Longman 1997).
[820] Brent Daniel Mittelstadt et al, 'The ethics of algorithms: Mapping the debate' (2016) Big Data & Society 3, 2 (commenting that the term "Algorithm has an array of meanings across computer science, mathematics and public discourse.")
[821] Moshe Vardi, 'What is an Algorithm?' (14th International Congress of Logic, Methodology and Philosophy of Science 2011).
[822] Robin Hill, 'What an algorithm is.' (2015) 29 Philosophy & Technology 1 35, 47 (Emphasis added).
[823] Mittelstadt et al, supra 820, 2.

On the essential question of what exactly constitutes an idea when "described, explained, illustrated or embodied in … a work",[824] this has been the object of much interrogation by the courts but surprisingly little elaboration by the legislator. In the context of AI, of note therefore is the Copyright Office's justification of §102(b), when it explained that "in response to the great debate over computers and copyright, …[it] is intended to disclaim any intention to protect a programmer's algorithm under the bill."[825] Perhaps due to its wide range of uses and the definitional opacity that arises from this, of note is that the term algorithm has not been widely adopted as a legal term of art in statute. The US and UK Copyright Acts forego the term entirely, and although the word is used in the Software Directive, it is not a defined term. The only elaboration comes in Recital 11, which make it clear that an algorithm falls as an idea, rather than as protectable expression.[826] As far as this author is aware, Japan is one of the few countries to use and define the term in statute.[827] Article 10.3 of the Japan Copyright Act sets forth that copyright protection in a computer program does "*not extend to the programming language, coding conventions, or algorithms used to create the work".* In doing so, it defines an algorithm as "*a process in a computer program, which consists of a combination of instructions for the computer."*[828]

As already touched upon, when the Ministry of International Trade and Industry issued its 1972 recommendation for a *sui generis* form of protection for software, the Japanese authorities were clearly concerned that the application of copyright to programs would inappropriately hamper its domestic technology industry. Given this background, combined with the additional pressure that the Japan came under in the late 1970s and early 1980s from the US to subsume programs into copyright law, Japanese policy makers were particularly aware of the need to create appropriately well-defined limits to the copyrightability of software. Alongside the Japan Copyright Act's exclusion of programming languages and coding conventions, using the term algorithm to explicitly exclude process from the scope of protection was undoubtedly viewed as necessary, not only in terms of legal doctrine, but arguably more importantly, in order to protect domestic technological innovation and competition.

In defining the term algorithm and therefore expressly excluding ideas which inhere in process from the scope of the Act, the Japanese legislature – as with §102(b) of the US Copyright Act –

[824] §102(b), US Copyright Act.
[825] Scope of Section 102. Register Of Copyrights, Second Supplemental Report on the General Revision of the US Copyright Law: 1975 Revision Bill 2 (1975) 1.
[826] Article 11 of the Computer Programs Directive states "For the avoidance of doubt, it has to be made clear that only the expression of a computer program is protected and that ideas and principles which underlie any element of a program, including those which underlie its interfaces, are not protected by copyright under this Directive. In accordance with this principle of copyright, to the extent that logic, algorithms and programming languages comprise ideas and principles, those ideas and principles are not protected under this Directive."
[827] The Korean Copyright Act (Act No 14634 21/3/2017) as with the JCA in terms of its treatment of computer programs also uses the term algorithm defined as "a combination of instructions and commands in a computer program". Article 101-2.
[828] The 1985 amendment to the Japanese Copyright Act (Act No 62) bringing computer programs into scope came into force on the 1st of January 1987.

acknowledged the economic and commercial policy imperative of incorporating both the "idea-expression", and in particular the less well-known "process-expression" dichotomy, into its statutory framework.[829]

Despite the common use of the term algorithm in everyday discourse, its omission from the US Copyright Act and its lack of definition in the Software Directive, in part arguably explains why the term is infrequently used by Western courts in overseeing claims of copyright infringement. Other than *Whelan v Jaslow,* where the court nebulously defined an algorithm as a "mechanical computational procedure",[830] none of the foremost US copyright cases dealing with software,[831] or indeed those from the European Court of Justice,[832] have engaged with what an algorithm is, or how to define it.

One reason for this rather surprising fact may be found in jurisprudential theory itself. The term algorithm works across the idea and process-expression divide and, in the absence of a statutory definition, its usage may well have confused matters further at a time when courts were already grappling with the issue of copyrightability in this new and conceptually challenging addition to the ranks of copyright subject matter. A computer program consists of not one single algorithm, but a vast number of them, and as discussed above at a higher level of abstraction the entire software program can even be thought of as a single algorithm. From the perspective of copyright law, in applying the idea-expression dichotomy to Marr's classic three level characterisation of computer programs, it is evident that an algorithm exists at the level of expression (the hardware implementational level) as well as the level of ideas (computational level and algorithmic level): [833]

1) *the computational level*: including the "the goal of the computation" and the "*logic of the strategy* by which it can be carried out" (Emphasis added);
2) *the algorithmic level:* including "How can this computational theory be implemented" and "what is the algorithm?"; and
3) *the hardware implementation level:* "How can the representation and algorithm be realised physically?"

Using this widely known taxonomy, an algorithm therefore encompasses the entire strata of a computer program from detailed expression in code to the ideational level where it serves to define

[829] Englund, supra 798.
[830] *Whelan Associates Inc v Jaslow Dental Laboratory Inc*, 797 F 2d 1222 (3d Cir 1986) 21.
[831] *Computer Associates International Inc v Altai Inc* (2d Cir 1992) 1992. *Sega Enterprises v Accolade* (9th Cir 1992). *Lotus Development Corporation v Borland International Inc*, 49 F3d (1st Cir 1995). *Sony Computer Entertainment Inc v Connectix Corp,* 203 F 3d 596 (9th Cir 2000). *Lexmark Int'l Inc v Static Control Components Inc,* 387 F 3d 522 (6th Cir 2004). *Google v Oracle America Inc* (2021).
[832] *BSA*, *SAS v WPL*.
[833] David Marr, *Vision: A computational investigation into the human representation and processing of visual information* (Henry Holt & Co 1982) 25.

the function and purpose of the entire program. In the absence of a clear definition in law, this fluidity and imprecision, encompassing three doctrinal pillars of copyright law – namely ideas, functions and expression – may go some way to explain why western courts have avoided referring to algorithms in their rulings. A further plausible reason why courts have eschewed discussion of algorithms is that the authorities overseeing copyright cases have themselves defined algorithm, in the context of patent rulings, to be no more than a mathematical formula.[834] Plausibly therefore the reason why there has been no move to evaluate the application of copyright law to algorithms stems from the courts' own conceptualisation of algorithms as mathematical formulae which lie outside the scope of copyright protection.

3.4 Programming Languages

In dealing with the interplay of ideas and functions, the Japan Copyright Act codifies the most explicit guidance for the courts of the three jurisdictions addressed in this thesis. In doing so, it excludes not only programming languages but coding conventions and a wide variety of protocols from protection. Turning first to programming languages, unlike the Directive, that expressly excludes them but leaves the term undefined, and the US Copyright Act, which ignores them entirely, the JCA not only precludes programming languages from protection but provides a concise definition also: "*letters, other symbols and structures used as a means to express a computer program.*"[835]

As with a natural language, the structure of programming languages in the form of words, grammar, and syntax are the building blocks of expression. Programming languages can be divided into at least two different types. On the one hand, there exist higher level languages such as C++, Java, and Python, which are commonly referred to as source code and are "human readable and writable". On the other hand, there are the low-level languages like machine code, which are sometimes referred to as object code. It is this "translation" of the source code which instructs the computer how to work. Once the code is translated into a machine language, they become executable, meaning that they form the direct means by which a computer performs its tasks. It is this machine code which directly engenders the electromagnetic signals that cause the computer to function.

[834] It is interesting to note the contrast with patent cases where the term is used in its original sense as a mathematical formula. In *Gottschalk v Benson,* 409 US 63 (1972) the US Supreme Court ruled against patentability of software which converted decimal numbers to binary code equating the instantiation of the algorithm in the program to a "mathematical formula." In *Parker v Flook 43*, US 584 (1978) the Supreme Court similarly rejected patentability equating the algorithm used in a computerised catalytic convertor in a fire alarm to the mathematical formula $2\pi r$. More recently in *State Street Bank & Trust v Signature Financial Group* in assessing software that resulted in no physical result, the court stated "We hold that the transformation of data … by a machine through a series of mathematical calculations into a final share price, constitutes a practical application of a mathematical algorithm, formula, or calculation, because it produces a useful, concrete and tangible result".

[835] Article 10-3(i), JCA.

Higher level languages are nearly always developed to simplify programming within the particular environment for which they have been designed. For example, FORTRAN was developed for high-performance numeric computation, and Java for interoperability within consumer electronic markets. It is interesting to observe, therefore, that although intellectual property law exists to protect the investment made by a person or organisation in their own creations, in this instance lawmakers have not afforded programming languages copyright protection. Despite the great cost associated with their development, from a public welfare perspective foreclosing the building blocks of expression in the computer industry would have had serious ramifications on market competition, not to mention the whole development of the digital age. Having to license programming languages would limit their uptake and directly impact on the pool of programmers conversant in a particular computer language. This in turn would result in a slowdown in the development of new technologies. As was noted by the Supreme Court in *Google LLC v Oracle Inc*,[836] the use of Java is not only an investment on behalf of its estimated six million programmers,[837] but its usage is the starting point for the creation of a whole raft of new and transformative technologies, digital ecosystems, and markets.

Thus, conferring monopoly control over programming languages – and therefore the very means of expression – would have entirely undermined copyright's policy objective of promoting the further creation of literary and artistic works. Had policy makers sought to extend protection to programming languages, it would have served to disincentivise the development of computer programs to the detriment of the economy, innovation, and technological progress.[838]

3.5 Programming Conventions and Interoperability

Another distinguishing facet of software expression which legislators sought to grapple with is that to function, computer programs must communicate with other programs around them. It is precisely this, their communicative and functional nature, that means they operate in an environment with many standards and coding conventions. Working within such a structured environment inevitably diminishes the scope for originality of expression, and required an answer from the authorities that did not overly restrain competition in this nascent sector of the economy.

The Japanese and European response was grounded in statute. Due to a politically charged trade deficit with the US, the policy imperative of safeguarding innovation and competition in the domestic

[836] *Google v Oracle America Inc* (2021) 34. Stating "About six million programmers had spent considerable time learning, and then using, the Java language." In this particular case, programmers familiar with Java had learnt standard expressions for calling on a function which had facilitated and contributed to the success of the Android platform which now has an estimated 2.5 billion users globally.
[837] David Curry, 'Android Statistics' (2022) https://www.businessofapps.com/data/android-statistics/ (accessed 4th April 2022).
[838] Despite apparently settled jurisprudence on programming language, it should be noted however that in *SAS v WPL* (2012) the court did envisage, with no elaboration or context: "the possibility that the SAS language … might be protected, as [a] work…, by copyright under Directive 2001/29 if they are their author's own intellectual creation." 45.

technology sector was felt particularly keenly in Japan. The diplomatic tension, alongside a civil law tradition where typically as many issues as possible are addressed upfront in the legislation, probably goes far to explain why the Japanese government sought to create flexibilities in its laws by explicitly excluding as well as defining a number of facets of software copyright law.

Focusing on coding conventions, Article 10-3 (iii) defines a convention (*kiyaku* 規約) as "*particular rules relating to the use of a programming language … deployed in a specific program*".[839] On first blush, the definition reveals little that distinguishes this provision from the preceding one that creates a derogation for programming language. Japanese jurists, however, contend that "convention" refers to the protocols and interfaces that allow the transmission of data within a network and between computers.[840] In other words, not the common conventions provided by the syntax covered by the definition of a programming language, but the standards and protocols which a computer programmer may actively select to employ in a specific computer program.[841] While this has not been the subject of elaboration in any of the prominent Japanese software rulings, this interpretation is lent considerable credence by textual interpretation. "[*R*]*ules … deployed in a specific program*" can be inferred to exclude standard conventions peculiar to a particular programming language, as these would be so common as to make the use of the phrase "*specific program*" redundant. Less ubiquitous standards expressed in the code, would infer *inter alia* standards relating to interoperability as well as more broadly other types of interfaces. Moreover, the number of authors of the leading treatises on Japanese copyright law positing this interpretation lends it significant authority.[842]

In contrast to Japan, where protocols and standards were expressly excluded from copyrightability, the European Union's approach was to only mildly truncate the exclusive rights of the author. The Computer Programs Directive only allows decompilation where it is "*indispensable*" for interoperability purposes,[843] and has "*not previously been [made] readily available*".[844] Thus, unlike their rivals in Japan, who are free to use an unlimited number and typology of coding conventions – for example, those that relate to the demands of a specific industry, security, or which safeguard the privacy of citizens – EU-based industries can only adopt protocols that are indispensable to achieve software interoperability.

[839] Article 10-3 (iii), JCA.
[840] Nakayama, 780, 43-6. Hideo Ogura & Shigehiko Kanai『著作権法コメンタールI「改訂版」』(A Commentary on Copyright Law I) (Daiichihōki 2020) Vol 1 363. Shigehiko Kanai et al『DX時代におけるデジタルコンテンツ著作権法』(Digital Contents and Copyright Law in an Era of Digital Transformation) (Gyōsei 2021) 39.
[841] Article 10-3(i), JCA. "Programming Language: the letters, other symbols and structures used as a means to express a computer program".
[842] Nakayama, Ogura & Kanai, supra 840, ibid. Moriyuki Kato『著作権法逐条講義（六訂新版）』 (An Article by Article Commentary on Japanese Copyright Law) (Copyright Research Information Centre 2013) 129.
[843] Computer Programs Directive. Article 6.1.
[844] Id. Article 6(1)(b).

Furthermore, whereas the expression contained in software protocols and standards are in the public domain in Japan, the question of the extent to which the reproduction of code required to achieve interoperability is permissible as a defence against infringement remains unclear under the Directive. The Japanese formulation lends itself to a broad interpretation whereby the courts would likely view the reproduction of all elements of the code that constitute a standard to be lawful. By contrast, the Directive leaves the door open to a *de minimis* interpretation, where only the elements that are "indispensable"[845] to interoperability may be copied. As a consequence, other routines that nevertheless comprise a standard but could be reproduced and expressed differently would have to be reprogrammed using a different form of expression. From the point of view of international competition, even relatively subtle differences in law such as this can have significant ramifications. For example, while the Japanese courts would in all probability concur with the Supreme Court in *Google LLC v Oracle Inc* holding that Google's use of all 11,500 lines of Oracle's API code was lawful, a European court may draw a far narrower conclusion.

True to form, in America, the normative questions posed by protocols and standards were primarily addressed in common law. Whereas the Japanese approach affords no protection to protocols and standards, over time American courts have taken differing approaches to the question of copyrightability. In *Atari Games Corp v Nintendo of America* (1992), the Federal Circuit was at pains to stress that to access the "digital key code" that would unlock Nintendo's console – so that it would accept Atari's games – was lawful as its aim was to learn and extract "*unprotected* ideas and processes".[846] Similarly, in *Sega Enterprises v Accolade* (1992*)*, Accolade's decompilation of the Genesis Platform's interface in order that Accolade's own games would play on the Sega platform was determined in the light of §102(b), on the grounds that the interoperability information was not subject to copyright. The court opined that disassembly of a program was permissible if its aim was to discover the "ideas, facts, and functional concepts in the public domain".[847] Thus, the code that allowed interoperability between different products was designated uncopyrightable and could be used freely by businesses to develop commercial products as they so wished.

More recently, however, the question of whether standards, such as interoperability information, are copyrightable has come into question. In *Google LLC v Oracle Inc,* the Supreme Court declined to engage in whether application protocol interfaces (APIs), which were being used by Google for purposes mainly beyond simple interoperability, were copyrightable subject matter. Rather than rely on the precedent set by the *Atari* and *Sega* judgements, it decided to assume "for argument's sake, that

[845] Ibid. Article 6.1 "The authorization of the rightholder shall not be required where reproduction of the code and translation of its form … are *indispensable* to obtain the information necessary to achieve the interoperability of an independently created computer program with other programs… ." (Emphasis added).
[846] *Atari Games Corp v Nintendo of America Inc* (Fed Cir 1992) 884 (Emphasis added).
[847] *Sega Enterprises Ltd v Accolade Inc* (9th Cir 1992) at 69.

the copied lines can be copyrighted."[848] This resulted in a determination grounded not in §102(b), as precedent might have suggested, but in fair use. Thus, leaving unresolved the core question of whether information that has become a de facto market standard – including that required for interoperability whether a markets standard or not – qualifies for copyright protection. Samuelson and Lemley have characterised *Google v Oracle* as a "lost opportunity",[849] and assert that in failing to address the central issue of copyrightability with respect to standards and protocols, what appeared to be relatively settled as a matter of law is now once again open to question.

Be that as it may, the approach taken by Japan and the US, from an economic viewpoint, has much to recommend it. Shielding protocols and interfaces from the monopoly control of rightsholders is a fundamental enabler of e-commerce. Standards are essential for the functioning of many different types of software and technical infrastructures, not least the world wide web. As shown by the indispensable nature of the open internet and the Android platform, whole new market ecologies can grow and prosper where protocols and standards remain free from the over-reaching exercise of copyright. Granting an excessive scope to copyright presents a very real danger of foreclosing access to essential technological elements of a networked society, and serves to undermine both economic and societal progress.

4. Case Law and the Protectability and Non-Protectability of Software

Having outlined the main statutory doctrines that specifically regulate computer programs, this section now turns to the norm-setting activities of the judicial authorities often undertaken as it is in the face of meagre direction from the legislation itself. Arguably, the prime issue facing the courts was that while statute provided some guiding principles relating to the application of the idea-expression and process-expression dichotomies, when it came to direction on originality of expression, the legislative response was relatively limited. Other than the EU and Japanese carve-out for programming languages, standards and interoperability this lack of direction led the courts to fine-tune a number of pre-existing doctrines.

The functional nature of software, and the fact its production takes place mainly within industrial confines, creates a higher bar for originality, as expression can be influenced by extrinsic factors. This inevitably creates a different pathway for evaluating and judging originality compared to more traditional copyright works. While a book or a painting may be the result of no more than the independent artistic impulse and inspiration of its creator, writing software is an industrial goal-

[848] *Google v Oracle America Inc* (2021) 2.
[849] Lemley & Samuelson, supra 735, 4.

oriented process. Pre-defined technical parameters and external factors such as market norms, operational efficiency, or cost will strongly influence the form of its expression. Furthermore, its development is often highly incremental, with programmers learning from and building on the efforts of others. As a result, whether a particular computer program is original and should be afforded protection under copyright law can be highly equivocal.

While the preceding section addressed the codification of legal norms peculiar to software on a thematic basis, the remainder of the chapter will divide the norm-setting activities of the courts by doctrine and country. This is because the tests that have arisen, while similar and often overlapping, at the same time display a distinctive approach that does not readily lend itself to a theme-based analysis.

4.1 Scènes à Faire – US

The first doctrine to be examined is *scènes à faire*. Adapting a French theatrical term meaning "scenes for doing" or "scenes (that must) be done", it originally referred to standard or stereotypical elements of a given plotline. For example, common to many fantasy stories are medieval backgrounds, magic, fantastical creatures, and a person discovering they have powers they were previously unaware of. Although the term was initially used by the courts in regard to common plot elements of a film,[850] when applied to software it quickly began to evolve. As a consequence, the doctrine shifted from one being entirely concerned with features common to a particular scenario, to one that includes expression framed by market and technology standards as well as hardware and software requirements.

Historically speaking, one of, if not the first case to apply a scènes à faire analysis to computer software, was *Atari Inc v North American Philips Consumer Electronics* (1982).[851] In adjudicating on a rival videogame to Atari's Pac-Man, the judge held that stock elements common to both games (such as: the maze where the game is set, scoring tables, and dots depicting performance), should not be afforded copyright protection. While this is essentially a straight application of scènes à faire to software, as it revolves around commonplace features inherent in a particular scenario, subsequent to this, the doctrine was materially developed in *Computer Associates Intern Inc v Altai Inc.* The main focus of this landmark ruling related to substantial similarity in the sequencing and architecture of job-scheduling software. As part of its espousal of the abstraction, filtration, comparison test,[852] the

[850] *Cain v Universal Pictures Co*, 47 F Supp 1013 (9th Cir 1942) (where the writer James Cain unsuccessfully sued Universal for elements of a plot included in the film "When Tomorrow Comes" from his 1937 book Serenade. The court ruled that plot devices such as playing the piano, prayer and hunger were "inevitable" in scenes set in a church where people were sheltering from a large and ferocious storm.)
[851] *Atari Inc v North American Philips Consumer Electronics Corp*, 672 F 2d 607 (7th Cir 1982) 40.
[852] *Computer Associates v Altai Inc* (2d Cir 1992) first employed the abstraction filtration comparison test, which has been widely adopted in US court rulings. The test relied on previous case law (*Nichols v Universal Pictures Co* 45 F 2d 119, 121 (2d Cir 1930) as well as the work

Second Circuit's Court of Appeals found utility in scènes à faire at the filtration level where public domain or copyright materials not in question are filtered out. In stating that scènes à faire has "analogous application to computer programs",[853] the court proceeded to identify five scenarios where the expression was unduly limited by extrinsic factors, and therefore lacked the originality required for it to be afforded copyright protection. These were, 1) limitations dictated by hardware, 2) interoperability requirements, 3) manufacturing standards, 4) standards relating to the market being serviced, and 5) widely accepted coding practices.[854] Although this characterisation is not without its detractors,[855] the court determined that any similarities in expression in the computer program arising from one of these five causalities would not constitute grounds for infringement.

Turning to case law, irrespective of whether scènes à faire was explicitly mentioned in the proceedings, analysis of the five factors identified by *Altai* shows that they provide a coherent framework for evaluating or predicting when a court will find against claims for infringement. In 1987, in *Plains Cotton Co-Op v Goodpasture,*[856] holding against the plaintiff, the Appellate Court found similarities in the structure, sequence, and organisation of the program to be largely dictated by the goal of the software and standard practice in the market being served – in this case the cotton industry. Reflecting the fourth and third *Altai* factor, the Court of Appeals for the 9th Circuit in *Apple Computer Inc v Microsoft Corp*[857] found that the use of overlapping windows in Windows software was a stock feature of many graphic user interfaces, and therefore not protectable unless the precise design of its expression constituted copyrightable matter.

More recently, Google's reuse of 37 Java APIs, in *Google v Oracle* constituted the deployment of "widely accepted programming practices within the computer industry", an unequivocal enunciation of the fifth factor.[858] The Supreme Court observed that the Java declaratory code belonging to Oracle had become a market standard used by over 6 million programmers worldwide, and as a consequence its use by Google in its Android platform was non-infringing. Given the substantial investment made by programmers to learn Java and its APIs, rendering the code subject to permission in the opinion of the Court "would make … the Sun Java API's declaring code a lock limiting the future creativity of new programs. Oracle alone would hold the key."[859] Justifying its opinion in clear policy terms, the Court observed that while this would result in high profits for Oracle, from a public welfare

of Professor M.B Nimmer, *Nimmer on Copyright* (Matthew Bender Publishing 1978). It involves dividing the software into differing levels of abstraction from the features and purpose of the individual elements of the code, up to the overall goal of the entire program. After this, non-protectable elements of the code can be filtered out from the disputed elements of the software. Finally, the remaining parts can be evaluated as to whether an infringement has occurred or not.
[853] *Altai*, 709.
[854] *Altai,* 710.
[855] Walker, infra 864.
[856] *Plains Cotton Co-op v Goodpasture Computer* (5th Cir 1987).
[857] *Apple Computer Inc v Microsoft Corp,* 35 F 3d 1435 (9th Cir 1994).
[858] *Google LLC v Oracle Inc*, 750 F 3d 1339 (Fed Cir 2014) 1363 quoting *Softel Inc v Dragon Medical and Scientific Communications Inc*, 118 F 3d 955 (2d Cir 1997) at 963.
[859] *Google LLC v Oracle America Inc* (2021) 1208.

perspective "those profits could well flow from creative improvements, new applications, and new uses developed by users who have learned to work with that interface. To that extent, the lock would interfere with, not further, copyright's basic creativity objectives."[860]

Shifting from market and common programming practice to factor two and technical standards, interoperability has also been the focus of a significant number of cases. In two analogous rulings, *Atari Games Corp v Nintendo of America Inc*[861] and *Sega Enterprises v Accolade,*[862] copyright protection was denied to the interoperability information of the plaintiffs as it was deemed a prerequisite for their competitors to launch games playable on the consoles of Nintendo and Sega. As outlined in the previous section, without the "keys" to access the consoles, direct competitors as well as downstream and upstream markets could not emerge, and the utilitarian aims of copyright would be subverted. These two landmark cases are particularly noteworthy as the decision to permit reverse engineering reflected the view of the courts that interoperability information should be in the public domain in order to encourage widespread innovation in technology markets. In determining that such information was non-protectable and should be free for others to use, it can be argued that the courts have blurred the line between source code that specifically facilitates interoperability per se and market standards more generally. As a result of proprietary interoperability information becoming usable by others, it can then either act as if it were a standard by those who deploy it, or if widely adopted by businesses as is the case with Oracle's Java APIs, it then truly does become a market standard. Furthermore, as observable in *Oracle,* in certain instances interoperability code can be used for purposes other than two sets of software communicating with each other. As such, distinguishing the difference between interoperability information on the one hand, and a general technical standard on the other, can be extremely challenging.

Finally, it should be said that neither is *Altai*'s characterisation universally accepted,[863] nor is scènes à faire as a doctrine without its detractors.[864,865] Indeed, given the expanded scope given it by *Altai,* straying beyond its roots into the fields of interoperability and technological restraints, combined with courts sometimes confusing it with merger,[866] its coherence as a doctrine can legitimately be contested. Nevertheless, due to the many external factors that affect original expression in software, as a doctrine it undeniably has proven a useful jurisprudential tool for the courts. On that account alone,

[860] Ibid.
[861] *Atari Games v Nintendo* (Fed Cir 1992).
[862] *Sega Enterprises v Accolade* (9th Cir 1992).
[863] Pamela Samuelson, 'Reconceptualizing Copyright's Merger Doctrine' (2016) UC Berkeley Public Law Research Paper No. 2763903, 21, 31.
[864] Robert Walker, 'Breaking with Convention: The Conceptual Failings of Scènes à Faire' (2020) 38 Cardozo Arts & Ent LJ 435 (arguing for a change to the copyright registration rules in the US and a more rigorous definition and applications of scènes à faire.)
[865] Krista S Schwartz et al, 'Software and Scènes à Faire' (2009) 190 Copyright World, 11 (proposing that scènes à faire should be evaluated from the perspective of the author, not the alleged infringer's perspective and be based upon what was common practice at the time the software was authored.)
[866] Lemley and Samuelson, supra 735, 14.

even if not for its jurisprudential coherence, it remains a fundamental and centrally important principle for evaluating software-related claims of infringement.

4.2 Merger Doctrine – US and Europe

A second doctrine governing originality of expression developed by the US courts, and closely intertwined with scènes à faire, is that of *merger*. Whereas scènes à faire as applied to software relates to standard expression commonly associated with a given context, or that which arises due to external technical and market factors, at the other end of the spectrum, merger concerns situations where choices are so circumscribed that expression is limited to one or a handful of possible choices. As a result, the expression is said to have merged with the idea owing to the fact that so few options are open to the author that the expression lacks originality.[867] Despite this classical interpretation, it is perhaps worthwhile noting that in a study of 150 cases involving digital as well as analogue forms of merger, Samuelson outlines four further factors in addition to, or arguably subsumed under, the concept of idea. These include instances where facts, processes, systems, and legal requirements have merged with the expression.[868]

The courts and some leading commentators[869] attribute the origins of merger to the seminal case of *Baker v Selden,*[870] in which copyright protection in bookkeeping forms was denied as the "methods and diagrams are to be considered as necessary incidents to the art." [871] As the expression in the forms were "incident to", and resulting from, the bookkeeping system, and could not have been expressed in any way materially different from how they appeared in the plaintiff and defendant's books, the expression it is held merged with the idea. Other commentators, however, point to more recent origins.[872]

Irrespective of whether merger finds its doctrinal roots in *Baker v Selden* or *Apple Computer Inc v Franklin Computer Corp*[873] as asserted by Samuelson,[874] as a doctrine it arguably suffers from both a lack of definitional clarity,[875] and an overlap with scènes à faire. Wiley argues that merger can almost always be found if evaluated at a very detailed level, and therefore as a tool of law making, it is indeterminate.[876] Other scholars have argued that because both merger and scènes à faire rely on a

---

[867] On a point of law, it is interesting to note that different courts have held different opinions as to whether merger is merely a defence against infringement, or whether it results in a determination against copyrightability. Samuelson, supra 863, passim.
[868] Id. 21.
[869] Nimmer, supra 796, §2.18[D][1].
[870] See also *Altai* para 86, stating "Baker v Selden … appears to be the cornerstone of what has developed into the merger doctrine."
[871] *Baker v Selden*, 101 US 99 (1879).
[872] Samuelson, supra 863, 3.
[873] *Apple Computer Inc v Franklin Computer Corp* 714 F 2d 1240, 1253 (3d Cir 1983).
[874] Samuelson, supra 863, 3.
[875] For example, see John Shepherd Wiley, 'Copyright at the School of Patent' (1991) 58 1 The University of Chicago Law Review, 119.
[876] Ibid.

shared understanding that in certain circumstances the options available for expression are pre-determined, it has sometimes resulted in a confusion between the two doctrines.[877] Arguably, the root cause of this mix-up is that both doctrines are in some senses two sides of the same coin. Disentangling whether the lack of originality in an expression is due to it being a common or inevitable feature of the environment in which it was created, or because the creator had limited ways to express themselves due to that environment, understandably does not result in a bright line dividing these two important software doctrines.

In analysing the difference between these precepts, Samuelson has suggested that in software cases the merger doctrine is most applicable to those that deal with "constraints imposed by mechanical specifications of the computer on which a program is run, computer manufacturer design standards, and requirements for achieving compatibility".[878] As has been observed by the author herself, and as discussed in the previous section, this assertion is at odds with *Altai*'s five factor characterisation of scènes à faire. Without doubt, the utility of doctrines – where leading scholars question their application by the courts – raises a number of legitimate questions regarding the coherence of the jurisprudence itself. Ultimately however, it perhaps matters less whether there is a meaningful normative distinction between scènes à faire and merger, than the fact that the US courts have found value in applying these two doctrines in computer program cases. The merger doctrine in particular, Samuelson argues, filled a gap for a period of time where courts were reluctant to embrace the broad functionality exclusions provided by §102 (b).[879] Furthermore, it served as a refinement mechanism because, where raised, it resulted in judges engaging in a thorough investigation of the appropriate boundaries applicable to copyright, and as a consequence arguably sounder law-making in a technically complex area of law.

Turning to Europe, consistent with the Design Directive's exclusion from protection of features "solely dictated by its technical function",[880] an analogous concept to the merger doctrine has become an established feature of Union copyright law, stemming from CJEU rulings involving works as disparate as GUIs,[881] software architecture,[882] sporting rules,[883] industrial[884] and clothing design.[885] It is too early to determine whether its European sibling will find the wide applicability Samuelson has attributed to it,[886] nevertheless, in common with the US, it is evident from careful analysis of the

[877] Samuelson, supra 863, 30 (giving examples of cases where the two doctrines have been confused. For example, *CMM Cable Rep Inc v Ocean Coast Properties, Landsberg v Scrabble Crossword Gameplayers* etc.)
[878] Pamela Samuelson, 'Functionality and expression in computer programs: Refining the tests for copyright infringement' (2016) Berkeley Tech LJ, 15.
[879] Samuelson, supra 863, 49.
[880] Article 7.1, Directive 98/71/EC of the European Parliament and of the Council of 13 October 1998 on the legal protection of designs OJ L289/28.
[881] *BSA*, 49.
[882] *SAS v WPL.*
[883] *FAPL*, *Football Dataco*, 39.
[884] Case C-833/18 *SI and Brompton Bicycle Ltd v Chedech / Get2Get* [2020] ECLI:EU:C:2020:461 24 (hereinafter *Brompton Ltd*).
[885] *Cofemel,* 31.
[886] Samuelson, supra 863, 21.

rulings that where “technical considerations, rules or other constraints”[887] dictate expression, a work is unprotected in cases where not just one, but also a number of choices, are possible. Further, alive to the policy need to promote competition and innovation,[888] the Court casts the exclusionary net of the merger doctrine wide. For example, in *FAPL*, the CJEU seems to exclude all sporting events from copyright protection on the basis that the “rules of the game, leav[e] no room for creative freedom for the purposes of copyright.”[889] Thus indicating that in certain circumstances the very presence of overarching constraints may entirely preclude free choice, and as a result, originality.

Moving to the issue of its jurisprudential status, an examination of the case law reveals two distinct aspects of the European merger variant. First, subsequent to *Cofemel,* is that as a doctrine it now operates as a core facet of the originality requirement operative as an autonomous concept in Union law.[890] Second, unlike the US, where Samuelson concludes the matter remains open,[891] merger creates an unequivocal determination against copyrightability as exogenous constraints preclude the free and creative choices of the author necessary for protection. As such, works whose expression is dictated to by external factors will fail to constitute original subject matter that can be protected under EU copyright law.

4.3 Room for Choice Doctrine – Japan

Analogous according to one leading commentator to merger,[892] similarly the “*room for choice*” [893] doctrine in Japan, renders a work unprotectable. As discussed above, determining originality in a computer program presents unique complexities owing to the many external factors that impose limitations on original expression in the code and the architecture of a program. For example, pre-written functions provided by programming languages and software libraries may, in certain circumstances, be tantamount to a required means of expression.

One notable software ruling that applies the “room for choice” doctrine – while simultaneously evincing of the Japanese courts’ standard for originality – is the *National Space Development Agency of Japan Program Case*.[894] In its ruling, the Intellectual Property High Court opined that “in order to be able to say that a program has copyrightable properties, it is necessary that there is a sufficient room for choice in the expression of the instructions, the combination of those instructions, or the

[887] *Brompton Ltd*, 24.
[888] *Brompton Ltd*, 16-18.
[889] *FAPL*, 98.
[890] *Cofemel,* 29-31.
[891] Samuelson, supra 863.
[892] Nobuhiro Nakayama,『著作権法第三版』(Copyright Law h Edition) (Yūhikaku 2020) 80.
[893] Japanese courts have frequently used the terms *sentaku no haba* or *sentaku no yochi.*
[894] *National Space Development Agency of Japan Program Case.*「宇宙開発事業団プログラム事件」Intellectual Property High Court. Heisei 18 NE 1003 (2006).

order in which they are expressed. Moreover, this should not be a commonplace expression, but one where the personality of the author is in evidence. If there is no room for choice in the expression of the program, or if the range of selection is very narrow, there is no scope for expressing the individuality of the author, and the program therefore lacks any copyrightable properties."

According to the Japanese Copyright Act, an author's own "thoughts or sentiments"[895] must be evidenced for a work to be afforded protection. If no imprint of the creator's individual personality is evidenced in the work, entirely in step with the horizontally harmonised originality test of the EU, copyright is denied.

As expressed by the IP High Court in the *Space Development Agency* ruling, restrictions placed on an author in how the code is written, or in how the constituent parts of the program are combined, is highly evocative of the merger doctrine. As identified by the Japanese authorities, exogenous factors limiting the room for choice available to computer programmers include: *inter alia* restrictions determined by the hardware and software, financial considerations, as well as limitations that spring from the need for the program to run efficiently.[896] For example, in the *Trainline Program Design Case,* the judge observed that in authoring source code "language is restricted, moreover syntax is strictly defined. In wanting the computer to run economically and efficiently the choice of functions can be limited, which often results in the concrete expression contained within the computer program being similar."[897] On the question of copyrightability, the Tokyo District Court went on to conclude that the form of expression in one sub-program "is *prescribed by* AutoCAD …" and because "the choices available … are extremely limited and *determined by* the syntax of the shape definition files, the program cannot be said to be creative".

Irrespective of the jurisdiction, terms such as "determined by", "dictated by", or "prescribed by" are indicative of the room for choice / merger doctrine. When expression is determined by externalities, such as software, hardware, or situational requirements, little room for choice may be afforded to the author to express themselves and thus, this will act to deny the conferral of copyright on a work.

In applying the room for choice doctrine, the Japanese courts have sometimes relied on an analogous test. Where, owing to external factors, a situation arises that "a work would virtually be the same whoever were to create it",[898] the personal expression of the author's own thoughts and sentiments

[895] Article 2 (1) "work" is defined as: "meaning a creatively produced expression of thoughts or sentiments… ." Japanese Copyright Act 1972 as amended.
[896] *Trainline Program Design Case*「電車設計用プログラム事件」Tokyo District Court. Heisei 15/1/31 Hanji 1820 / 127 (2003).
[897] Ibid.
[898] For example, this formulation was used in *New Masuda Ashi Case*「NEW 増田足事件」Tokyo District Court. Heisei 23/1/28 Hanji 2133/ 114 (2011) and the *Contact Angle Program Case*「接触角計算プログラム事件」Intellectual Property High Court. 28/4/27 Hanji 2321/85 (2016).

will be un-evinced. As a consequence, if a work's expression is determined by exogenous factors to the extent that it would be similar irrespective of who created it, the work will fall outside the scope of copyright protection.

4.4 Banal Expression and Brevity of Expression

Further interrelated doctrines that exclude a work from copyright protection relate to banal, facile, commonplace, and brief expression. Lacking the personal stamp of an author, when an expression lacks originality, it will fail to meet the required threshold for copyright eligibility. A facet of the author's own intellectual creation test, a number of national courts in Europe expressly deny protection to "trivial" or "banal" expression. In the Netherlands, copyright has been denied short sentences involving insufficient creativity.[899] Similarly in the UK, commonplace short expressions fail to receive protection.[900] Analogous to scènes à faire, more broadly in France, the term *banalités* is used by the courts to refer to instances where the expression is unoriginal as it naturally flows from the environment from which it stems.

A feature of a number of rulings in Japan, the "banal expression"[901] doctrine has often been brought to bear in software cases where sub-programs typical of the programming language and the task being performed are used, or where the combination and selection of functions are deemed to be generic, predictable, and lacking in character.[902] For example, overlapping with the scènes à faire and the merger doctrine, in the *Trainline Design Program Case,*[903] similarities in the claimant and defendant's software for the menu titles and their sequencing were ruled unoriginal as "determined by the requirements of designing in AutoCAD, and the demands of Japan Rail East". Reminiscent of *Plains Cotton Co-Op v Goodpasture,*[904] in the *Sekisan Kun Case,*[905] the court found that the detailed expression and sequencing within the GUI was commonplace, resulting from standard practice within industry – in this case, the construction industry. Similarities in the terms used in the GUI, and the order of inputting values into the software, were found by the court to lack the personal touch of the

---

[899] See Supreme Court (Netherlands) *Zonen Endstra v Nieuw Amsterdam* (2008) ECLI:NL:HR:2008:BC2153, HR 30.05.2008, NJ 2008, 556.

[900] *Noah v Shuba* [1991] FSR 14. 32-33. *Kirk v J & R Fleming Ltd* [1928–1935] Mac CC 44. *Costa v Dissociadid Ltd* [2022] EWHC 1934

[901] In Japanese the term used by the courts is *arifureta.*

[902] See *Trainline Design Program Case* where the deployment of functions common to AutoLISP were found to lack creativity「電車設計用プログラム事件」Tokyo District Court Heisei 15/1/31 Hanji 1820 / 127 (2003); *System Science* ① *Case* where printer functions in both plaintiff and the defendant's software were both ruled to be too generic to be afforded copyright protection.「システムサイエンス①事件」Tokyo High Court Heisei 1/6/20 Hanji 1322 138 (1989); *JR-CAD Program Case* ruling that combining certain functions common to AutoLISP into a program lacked creativity.「JR-CAD プログラム事件」Tokyo District Court Heisei 13 Wa 17306 (2001); *Disk Publisher Control* Case ruling that use of common Microsoft functions in both the defendant and plaintiff's software was not a creative act and therefore not copyright subject matter.「ディスクパブリッシャーソフト事件」Tokyo District Court. Heisei 24 WA 5771(2012).

[903] *Trainline Design Program Case* Tokyo District Court Heisei 15/1/31 Hanji 1820 / 127.

[904] Supra 856.

[905] *Sekisan Kun Case*「積算くん事件」Osaka District Court Heisei 10 WA 13577 (1998).

author, and were ascribed to the fact that both sets of software were used for designing and estimating building costs.
It is noteworthy that, as with the abstraction, filtration, and comparison test operative in the US, Japanese courts have also recognised that external factors influencing the functionality and form exist at different levels within the software making the analysis and balancing of these factors across different levels of abstraction a crucial step. In the *Cybers Case*,[906] commonalities in the expression of the program were deemed to result from the fact both plaintiff and defendant produced groupware for the same market. Both parties produced scheduling and facilities booking software, which reflecting the business task, contained features relating to the day-to-day running of a company, such as scheduling employee activities, managing remote members of staff, address books, to-do lists, facility booking, etc. In dismissing the claims of the plaintiff, whilst the Tokyo District Court agreed that the details of the expression were different, at a higher level of abstraction the similarities in the interface were found to be commonplace and "inevitable configurations that result from the function of the software, or stock features found in conventional bulletin boards, schedulers etc as well as being a feature of any related software products." Of note was that not only was analysis of the software undertaken at differing levels of abstraction, but the fact that the court also undertook a balancing act whereby similarities in the software design stemming from industry norms *outweighed* the significance of the granular details of its execution and expression.

Intertwined with denial of copyrightability for commonplace expression, the authorities in Japan have also pointed to length of code as an indicator as to whether a work should be afforded protection.[907] In the *Mobile Phone Contents Case,*[908] in which the defendant continued to supply astrological software for use on mobile phones and PCs beyond an agreed cut-off date, the short length of the code was a decisive factor for the court. In analysing one subprogram,[909] which calculated astrological information pertinent to the user, it was found to "be no more than an instantiation of the algorithm, and so short in length that the expression cannot be said to display the personality of the author". Similarly, a quantitative approach to analysing code was a key feature of another prominent software judgement, the *Trainline Design Program Case.*[910] In this ruling concerning similarities between two AutoCAD software programs supplied by separate companies to the East Japan Railway Company, the length of code was a determining factor for denying copyrightability in sub-programs present in both sets of software. The court held that two programs, one relating to printing and the other relating

[906] *Cybers Case*「サイボーズ事件」Tokyo District Court Heisei 14 /9/5 Hanji 1811/127 (2002).
[907] Sueyoshi Wataru,「プログラム著作物の創作性について」in Koizumi N and Tamura Y (eds),『はばたき－２１世紀の知的財産法－中山信弘先生古稀記念論文集』('Creativity in Software Works' in *Taking Flight: Intellectual Property Law in the 21st Century*) (Kōbundō 2015) 554 (analysing the *Mobile Phone Contents Case* where six composite scenarios for denying originality based on them constituting banal expression are cited, four of which include the length of the work as a determining factor.)
[908] *Mobile Phone Contents Case*「携帯電話コンテンツ事件」Intellectual Property High Court Heisei 23 NE 10051 (2011).
[909] Known as the "Kyūsei (九星) function".
[910] *Trainline Design Program Case.*

to setting span lines between linear entities, were too short in their expression to constitute copyright works.

While the courts in Japan are no different to any other in that decisions on originality are generally the result of qualitative assessment, nevertheless, the focus on the length of code is a distinctive feature of a number of prominent Japanese software cases. Although the authorities generally focus on the original and creative qualities which may inhere in a literary or artistic work, it is proposed that a causal link between quantitative and qualitative assessments of copyright indisputably exists. The shorter the expression, the more likely it is to be formulaic and unoriginal. In the case of computer programs, the length of the code arguably has a particular significance given that it is written in an environment where external limitations on personal expression are already particularly common. It is proposed that, as the Japanese courts have determined, the corollary of these multiple limitations on original creative expression is that as a rule, the shorter the code, the less likely it is that the originality and personal imprint of the author necessary for copyright to inhere in a work will be apparent. As discussed in detail in Chapter 5, this is of particular import for this thesis because brevity of expression is a notable feature of many machine learning software programs.

Conclusion

Despite computer programs being first recognised as copyrightable in the 1960s, they nevertheless continue to pose significant challenges to lawmakers to this day. As textual expression synthesized into functioning machine code, whether software is a suitable subject matter for copyright has been questioned by many. As discussed, what distinguishes software from other copyright works is its functional nature and industrial application. From the perspective of jurisprudence, not only did its codification into copyright law lead to a need to safeguard the ideational and functional aspects of software from exclusive proprietary control, but it also has meant the bar for determining originality in software is higher than with other protectable works due to the proliferation of technical, market, and other factors which engender non-original expression. Another factor that sets computer programs apart from other copyright works is that the creative expression is buried deep within the work, thus hidden away from other authors and potential competitors who wish to access it. Concealed behind a firewall or simply obscured as machine code, the creative expression that defines all other copyright subject matter is distinctly hard to locate in a computer program.

At the political level, the concern that overly broad copyright protection would harm innovation and competition in the growing technology sector was particularly evident in Japan. As strong competitors to the US in technology markets, Japan initially rejected copyright outright as a legal means of protecting software. By contrast, the US response to the software policy challenge was endogenous to copyright law in the form of §102(b) of the US Copyright Act. Creating comparatively limited practical guidance to software investors and developers, it was left to the courts to do much of the heavy interpretational lifting. This common law approach led to the application and enhancement of pre-existing doctrines including scènes à faire, merger, and in particular fair use. In contrast to this rich body of rulings, the main doctrinal thrust in Europe has been statutory, in the form of the Computer Programs Directive which sought to shield market competition from over-protection by introducing a handful of software-specific limitations and exceptions.

Of these, Article 6 of the Directive – which explicitly permits reverse engineering solely for commercial and non-commercial interoperability purposes – is arguably the most significant of the derogations outlined therein. Nevertheless, innovators in Europe wishing to decompile software devices for other purposes such as general R&D, or for the purposes of forward engineering more broadly, have been left ignored by the Directive. In comparison to Japan, where no restrictions are placed on reverse engineering,[911] thus the European Directive restricts decompilation solely to the purposes of interoperability. Europe-based businesses wishing to simply research and learn from the works of their competitors are as a result limited to passive observation and monitoring as part of the software's normal function and operation.[912]

Unlike the US[913] and Europe, the Japanese legislative response to the multiple doctrinal challenges posed, although offering explicit derogations for programming languages, conventions, and algorithms, introduced no accompanying limitations and exceptions. As outlined above, it was without doubt this absence of software-specific permitted acts which led the Japanese courts to apply a suite of pre-existing doctrines rendering certain software expressions in the public domain.

While it is contended that – irrespective of the jurisdiction – many of the juridical doctrines examined above blur and demonstrate significant overlap, from a public policy perspective this matters little. Most salient was the need to support competition in technology markets and safeguard from monopoly control the ability for programmers to reuse ideas, functionalities, and technical standards. In short, in the face of the many normative challenges that arose from introducing industrial machinery into a body of law associated with the arts, without strong action by the courts, the limitations inherent in statute would have overly stymied the technology industry to the detriment of both the economy and societal progress.

[911] Article 30-4 of the JCA allows reverse engineering based upon explicit confirmation by Tsukasa Nakaoka, the Deputy Director of the Agency for Cultural Affairs, in hearings on the 2018 amendments held on 6/4/2018 at the National Diet. In line with the principle of authentic interpretation as exercised in Japan, the opinion on reverse engineering from the Agency of Cultural Affairs as part of government should be regarded as valid until a decision is made by the courts to the contrary.
[912] Computer Programs Directive. Article 5.3.
[913] Public Law 96-517 (HR 6933) *inter alia* amended the US Copyright Act introducing an exception allowing in the course of running a program a copy to be made, as well as a reproduction for archival purposes.

As will become self-evident in the following chapter, the body of software copyright law developed in the 1980s and 1990s ineluctably reflects the norms of pre-internet production methodologies and markets. Although machine learning can be traced back to the 1950s with pivotal developments like Arthur Samuel's self-learning draughts program and IBM's Mark 1 Perceptron Machine, the copyright developments of the `80s and `90s focused entirely on more conventional software. The impetus for legislation was not the cutting-edge research being undertaken into artificial intelligence in academia and industrial labs, but the need to regulate commercial software markets as they started to grow in size and economic importance. The now mainstream nature of the internet, open source software, and the open environment where both academic and business communities share the latest machine learning algorithms via toolkits and the literature, creates a very different setting to that for which computer program related copyright law was first developed.

The question that naturally arises from this is whether – or more accurately, the extent to which – existing copyright legislation and the normative developments that took place in the courts three to four decades ago, designed to protect traditional software, still affords adequate protection to AI and machine learning programs. The next chapter will examine how far the legal response still adequately protects today's AI innovators and investors. In addition, based on this conclusion, the remainder of the thesis will evaluate whether a new AI-specific response from policy makers is required, and if so, what form it should take in order to maximise public welfare.

Chapter 5: Programming Code, Model Architectures and Parameters: Are Machine Learning Models Protectable Subject Matter?

Introduction

Designated as a literary work under the Berne Convention,[914] it is perhaps puzzling to even question whether AI software is adequately protected by copyright *de lege lata*. However, the distinct production modalities totally unlike those for which software copyright law was devised, combined with the prodigious economic importance of machine learning, nonetheless warrants a careful evaluation of copyright's role in safeguarding incentives to invest in this transformative new technology. This chapter will focus primarily on applying existing copyright norms to a handful of the most commonly used machine learning models. These "eager learners', which include decision trees, random forests, SVM models, linear and logistic regression, and neural networks are of particular interest from a copyright perspective as they are written and designed not just by human programmers – as necessitated by copyright law – but also by the data upon which they train.

This chapter begins by outlining two AI taxonomies. After this, in the light of the copyright doctrines examined in the previous chapter, core features of the model software – including the parameters,[915] hyperparameters,[916] source code, and architecture – will be evaluated to ascertain the extent to which copyright protection extends to eager learning AI algorithms. Third, the applicability of copyright and database rights to compilations of sub-programs and parameters will be examined. Fourth, whether copyright law requires recalibrating in order to safeguard innovation will be evaluated in view of the numerous extra-copyright legal and technological regimes that can be relied on to protect economic incentives.

1. AI Model Taxonomies

When AI is spoken of beyond the domain of computer engineering, it is often artificial neural networks (ANN) that are being referred to. Neural networks can be time consuming to create, costly to run, and understanding why a certain output was arrived at may be complex (XAI).[917] Contrary to the impression created by the legal literature, there are a vast number of different AI models that data scientists deploy in their everyday work. At the other end of the spectrum from ANNs lie the simpler and, according to Kaggle, the more commonly used models: linear and logistic regression (used by

[914] Article 4, WCT. Article 10.1, TRIPS.
[915] An internal value of the model which is calculated from or derived at with reference to training data.
[916] A configuration within a model set by the model creator or hyperparameter tuning software which is not estimated from the training data.
[917] This often heard criticism of neural networks is less valid than it once was, as explainability tools are increasingly commonplace. e.g. Leading explainable AI (XAI) tools include SHAP and LIME.

80.3% of data scientists), decision trees / random forests (74.1%), and gradient-boosting machines (59.5%),[918] which have the advantage of being computationally inexpensive and comparatively easy to interpret.[919]

Taxonomically, machine learning models can either be characterised by how they interact with data, or by the tasks that they are able to perform. Taking the first approach, there are typically said to be four types of machine learning model:[920]

1. Supervised Learning

Named so because during the training phase the model is "supervised" by labelled data containing the correct target information, once trained, it is able to make predictions with high levels of accuracy on previously unseen and unlabelled data.[921] Examples of supervised learning models include linear and logistic regression, K-Nearest Neighbours, decision trees, random forests, support vector machines (SVMs), and some neural networks. Multiple instance learning, where similar data is given one label only, may also be characterised as a type of supervised learning.[922]

2. Unsupervised Learning

Unsupervised learning models are tasked "blind" to detect patterns. Their prime use is for clustering ingested data into similar groupings, thus creating recognisable associations between datapoints. Unsupervised learning algorithms may also reduce the number of datapoints by aggregating them into broader categories in a process known as dimensionality reduction.[923] Examples of unsupervised learning models include K-Means, principal component analysis (PCA), the Apriori algorithm, and a type of recurrent neural network called a Boltzmann machine.

3. Semi-supervised Learning

A combination of supervised and unsupervised learning models, these models utilise a small amount of labelled data alongside a majority of unlabelled data.[924] Examples include deep belief networks (DBNs).

---

918 *State of Data Science and Machine Learning 2021*(Kaggle 2021) 33.
919 Vili Podgorelec et al, 'Decision trees: an overview and their use in medicine' (2002) Journal of Medical Systems 26 5 445, 447 (explaining the easy interpretability of decision trees.) Andrei Konstantinov and Lev Utkin, 'Interpretable machine learning with an ensemble of gradient boosting machines' (2021) arXiv:2010.07388 (outlining models that have been developed to enhance interpretation.)
920 Aurélien Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow* (O'Reilly 2022) 8.
921 Ibid.
922 Ian Witten, Eibe Frank, Mark Hall, Christopher Pal, *Data Mining – Practical Machine Learning Tools and Techniques* (Morgan Kaufmann 2011) 472 ff.
923 Géron, op. cit. 920, 10.
924 Id. 13.

4. Reinforcement Learning

Used in many different scenarios ranging from robotics to game theory, reinforcement learning models learn and update through trial and error, being "rewarded" for correct results and "punished" for incorrect responses. Examples include Q-learning, Sarsa, and a type of neural network called a Deep Q-network.

A second, equally informative non-exhaustive taxonomy is one based on the function undertaken by the model. Any model type is able to undertake at least one of these three tasks, with many being able to perform more:

1. Regression

Regression is the activity of predicting a numeric value from a linearly or non-linearly ordered set of values.[925] Models which undertake regression are used to make numeric predictions and forecasts. e.g. nearest neighbours, random forests, neural networks, decision trees, support vector regression etc. Typical applications include weather forecasting, drug responses, stock price predictions, and demand forecasting in retail.

2. Classification

Classification models identify within which category a datapoint is likely to fall. They are utilised in a wide range of applications including image classification, disease prediction, image recognition, optical character recognition, and disaster prediction. Typical classification models include support vector machines (SVM), nearest neighbours, convolutional neural networks (CNN), and random forests.[926]

3. Clustering

Clustering algorithms are usually associated with unsupervised learning, which, as the name suggests, group similar objects together. Clustering algorithms facilitate dimensionality reduction and are used in a whole host of applications ranging from identifying fake news,[927] customer segmentation, image

[925] Simon Haykin, *Neural Networks and Learning Machines* (Pearson 2009) 68 ff.
[926] Géron, op. cit. 920, 79 ff.
[927] Claire Whittaker, 'Innovative Uses of Clustering Algorithms in the Real World' (*Datafloq,* 4 April 2019) https://datafloq.com/read/7-innovative-uses-of-clustering-algorithms/

analysis in the health sector, to defining relationships in the financial sector.[928] Examples include K-Means clustering, DBSCAN, and agglomerative clustering.[929]

4. Representation or Feature Learning

Representation learning is a process whereby the models learn the most important features in the input data (observations), and how to generate them.[930] These representations can be used for classification, predictions, regression, clustering. Examples include but are not limited to multilayer neural networks, K-Means clustering, principal component analysis, and autoencoders.

5. Generative AI

These are probabilistic neural network models that identify patterns and structures within training data. By incorporating a stochastic component in their calculations, the models are able to generate new and original content. e.g. Naïve Bayes models, Variational Autoencoders (VAE), Generative Adversarial Networks (GANs), LLMs, Generative pre-trained transformers (GPT), etc.[931]

2. Copyright Protection of Computer Programs

Before exploring the extent to which copyright protects AI models, it is first necessary to delineate between those elements of a software program which may be afforded copyright protection from those which are unprotected. As described in detail in the previous chapter, anchored in international treaties and jurisprudential tradition, the authorities have developed a rich canon of legal precepts which *a priori* exclude certain expressional features of a computer program from copyright protection.[932] These include ideas, algorithms *qua* mathematical concepts, functionality, programming languages, interoperability information, commonplace expressions, as well as forms of expression dictated by exogenous factors.[933]

In terms of form, computer programs stand apart from other copyright works as expression operates simultaneously at differing levels of abstraction. Unlike natural language, which triggers proprietary entitlements from original expression directly tangible on the surface of the work, by contrast

[928] Jiamin Li & Harold W Lewis, 'Fuzzy Clustering Algorithms – Review of the Applications' (2016) IEEE International Conference on Smart Cloud 282.
[929] Andreas C Müller and Sarah Guido, *Introduction to Machine Learning with Python* (O'Reilly 2016) 168 ff.
[930] See David Foster, *Generative Deep Learning* (O'Reilly 2019) 23-7.
[931] Ibid.
[932] Article 9 (2) TRIPS. Article 2 (1) Berne Convention. See M Ficsor, *Guide to Copyright and Related Rights Treaties Administered by WIPO and Glossary of Copyright and Related Rights Terms* (WIPO 2003) (suggesting that "whatever may be the mode or form of its expression" is a reference to the idea-expression dichotomy.) BC 2.2 and BC 2.3.
[933] See in terms of statute: Article 9 of TRIPS, Article 2 of WCT, Computer Programs Directive 2009/24/EC Recitals 11, 13-15, Articles 1.2, 5 and 6, Copyright Law of the United States (Title 17) §102 (b).

expression in source code has a unique duality in that it simultaneously acts as a command corresponding to or instructing machine code elsewhere in the programming environment. As a result of electromagnetic signals, the "words" communicate with the computer making the computer function.

Despite its multi-layered complexity, in common with natural language, individual words or a very small number of words in the source code will at the textual level not reach a requisite threshold for copyright protection.[934] However, the point at which short commands and instructions become subject to protection by copyright law lacks a distinctive bright line. While the CJEU has stated that an individual word in isolation is not the intellectual creation of an author, nevertheless it holds that even a limited number of words – even as few as eleven – if sufficiently original, may reflect the intellectual creation of their author.[935] Thus, it is axiomatic that in theory even relatively short commands may be afforded copyright protection where they reflect the original conception and expression of the programmer.

In common with other literary works, copyright protection for software does not cease to operate at the textual level. As Advocate General Bot set forth in *SAS v WPL,* "protection of a computer program is not … confined to the literal elements of that program, that is to say, the source code and the object code, but extends to *any other element* expressing the creativity of its author."[936] Thus, where originality in the program is discernible in the arrangement and structure,[937] copyright may inhere in what computer scientists routinely refer to as the architecture of the software.[938] It is for this, the structure and arrangement of the individual coded instructions, sub-programs, and other elements of the program, that courts and scholars alike have coined a number of variant terminologies. These include "non-literal" or "non-textual" expressional elements, and particularly in the United States, "structure, sequence and organisation" (SSO).[939]

3. Changed Modalities of AI Software Production

[934] For example, in Europe, SAS *v WPL* held that "keywords, syntax, commands and combinations of commands, options, defaults and iterations consist of words, figures or mathematical concepts which, considered in isolation, are not, as such, an intellectual creation." 66. See also infra 970 regarding the short words and phrases doctrine in the US.
[935] *Infopaq I*, 45, 48.
[936] *SAS v WPL,* Opinion of AG Bot, 50. See also the final judgement (stating that it is only through the "choice, sequence and combination" of words, commands etc that "the author may express his creativity in an original manner".) 67.
[937] Putting aside the formal statutory protection of object code it is submitted that in the case of software the structure that should properly be awarded copyright is at the theoretical design level expressed in the source code rather than at the level of the executable machine code given that the interaction and "arrangement" of the elements of the program in the object code is different to the source code and determined by the operating system. See Ralph Clifford, 'Intellectual Property in the Era of the Creative Computer Program: Will the True Creator Please Stand Up' (1997) 71 Tul L Rev 1675, 1689 ff.
[938] See Noam Shemtov, *Beyond the Code: Protection of Non-Textual Features of Software* (OUP 2017). Mark Lemley et al, *Software and Internet Law* (Aspen Publishing 2011) 36 ff. Ken Moon, 'Another nail in the coffin for non-literal software copyright infringement?' (2015) 10 Journal of Intellectual Property Law & Practice 12, 921.
[939] A term coined by the court in *Whelan Associates Inc v Jaslow Dental Laboratory Inc,* 797 F2d 1222, 1224 (3d Cir 1986).

Important to understand in any discussion relating to copyright's applicability to modern day software practices, and ML in particular, is that the common production norms fifty years ago around which regulators first congregated to justify copyright protection have changed dramatically in the intervening years. This is particularly evident in the field of artificial intelligence. As explained by Karpathy, former Director of AI at Tesla,[940] not only does data "write" the all-important parametric values in the code, but the open-sourced toolkits provide pre-coded the most popular algorithms *qua* software, thus significantly reducing the need for developers to write their own routines. As a consequence, these very different realities of AI code programming create a significantly less certain picture for copyright protectability than is the case for more traditional forms of software design.

3.1 Object Oriented Programming and Open Source

In the 1970s and 1980s, when legislatures around the world first began to embrace copyright as an automatic and unregistered means of protection for computer programs, the predominant programming paradigm was a style known as procedural programming. This form of coding is particularly noted for the length of its code, in part due to the need to repeat the same commands throughout the program whenever they are called. The centrally structured and protracted nature of this type of code means that managing and working on it in parallel across different teams and organisations is extremely challenging.[941]

By contrast, object-oriented programming (OOP) – which entered the mainstream in the 1990s and is predicated on languages such as Java, Python, C++ – was an answer to the systemic issues of complexity and concurrency inherent to procedural programming. OOP is a decentralised programming model where functions and data are grouped into independent objects. As a result, sub-elements of the code can easily be reused without the need to repeat it each time within the same program. The flexible nature of OOP facilitates not only extensibility of the software, but its reusability, thus facilitating the creation of new and potentially unrelated software applications.[942]

Significant for AI, in the mid 1990s, the conjunction of the wide-spread uptake of object-oriented programming – contemporaneous with the adoption into the mainstream of the world wide web – contributed directly to the popularity of software libraries.[943] Interlinked with these transformative technical developments was the dramatic rise of open source software. The use of openly licensed software, where the code is free for programmers to access and reuse under a variety of different open

[940] Karpathy, op. cit. 53.
[941] Martin Prause and Juergen Weigand, 'Industry 4.0 and object-oriented development: incremental and architectural change' (2016) 11 Journal of Technology Management & Innovation 2 104.
[942] Brian Henderson-Sellers and Julian Edwards, 'The object-oriented systems life cycle' (1990) 33 Communications of the ACM 9 142
[943] Bertrand Meyer, *Object-oriented Software Construction* (Prentice Hall 1997) 297.

licence types, is now commonplace across the whole technology industry, with the value of the open source service market estimated to reach $66.84 billion by 2026.[944]

Central to this study, given its implication for authorial originality, is the ubiquitous reliance on pre-programmed AI software libraries and toolkits to build ML models.[945] According to research by the AI platform Kaggle, machine learning frameworks are used by 96% of data scientists.[946] Of particular note is that open source machine learning libraries such as PyTorch, TensorFlow, PaddlePaddle (supported by Meta, Google and Baidu respectively), scikit-learn and Keras, take centre-stage in both commercial and non-commercial machine learning production environments, providing a wide array of pre-programmed openly licensed machine learning models, programs, and tools.[947, 948]

3.2 A New Software Production Paradigm: What significance for copyright law?

The fact that the creation of machine learning models is largely predicated on readily available open source software code,[949] combined with the formative role of data which part-programs eager learning models[950] raises fundamental questions about the continuing relevance and applicability of copyright in this important sector of the software industry.

In terms of first principles, it is important to stress, however, that the ubiquitous use of open source toolkits to design task-specific AI models does not negate the role of copyright *per se* given that it will almost certainly inhere in part or in whole in the library's pre-programmed code. Nevertheless, a corollary of data scientists' heavy reliance on AI toolkits is that the writing from scratch of qualitatively substantive amounts of new and original code in the core of the model is in all probability likely to be rare.[951, 952] This can be surmised from the fact that if there was the substantive

---

[944] $66.84 Billion Open Source Services Market by Industry, Service Type, and Geography (*businesswire*, 13 November 2020) https://www.businesswire.com/news/home/20201113005374/en/66.84-Billion-Open-Source-Services-Market-by-Industry-Service-Type-and-Geography---Global-Forecast-to-2026---ResearchAndMarkets.com
[945] PyTorch and SciKit Learn are licensed under the 3-clause BSD licence. TensorFlow, JAX, Keras and PaddlePaddle are licensed under the Apache 2.0 licence.
[946] Kaggle, supra 918.
[947] According to *State of Data Science and Machine Learning 2021* (Kaggle 2021) 82.3% of data scientists reported using scikit-learn, 52.6% TensorFlow, 47.3% Keras and 33.7% PyTorch. For a taxonomy of machine learning frameworks see Fig. 2 of Nguyen Giang et al, 'Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey' (2019) Artificial Intelligence Review 52 77.
[948] Sci-kit learn for example provides a wide range of common model types including linear models, support vector machines, nearest neighbours, decision trees, random forests, and multi-layered perceptrons. Data scientists wishing to use more complex neural network models or undertake reinforcement learning are more likely however to turn to Keras, PyTorch and TensorFlow which offer fuller support for different ANNs such as Long short-term memory (LSTM), Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN).
[949] No-code AI applications also allow the creation of models without a knowledge of programming. For example, Amazon's SageMaker, Google Cloud Auto ML tables, MonkeyLearn, Akkio etc. See also a description of Copilot which can create code in Giorgio Franceschelli and Mirco Musolesi, 'Copyright in generative deep learning' (2022) Data & Policy 4: e17, 12 ff.
[950] Karpathy op. cit. 53.
[951] Kaggle, supra 918, 35 (showing only 4% of data scientists reported using no machine learning frameworks.)
[952] See Fig.1-2, Chapter 1. See also *The State of Data Science 2020: Moving from Hype Toward* Maturity (Anaconda 2020) finding data scientists on average spent 19% of their time on loading data, 26% on data cleansing and 21% on data visualisation. Only 11% was spent on model training, and 11% on its deployment.

intent by a developer to write their own code there would be little reason to rely on the library in the first place.

While developers may rely on various extrinsic incentives,[953] from a theoretical standpoint, the crucial question endogenous to copyright is whether the final task-specific AI program, created with the assistance of a software library, manifests original authorial expression. Thus, in terms of economic incentive rationales, whether copyright provides sufficient inducements to investors in circumstances where trained models are comprised predominantly of pre-written open source software and data-written code, constitutes a question of considerable normative significance. If trained AI models lack copyright protection for the developer of the task-specific model, or are only thinly protected, according to utilitarian legal theory this will have a negative public welfare effect. Accordingly, in the face of free-riding, costs may not be recouped and as a result the incentive to further invest is diluted or rendered entirely absent.[954]

Owing to the ubiquitous reliance placed on ML frameworks by developers, the prime focus of this subsection will be on the toolkit-based programming environment that has become the industry norm. First, focusing on the literal level of copyright protection, the framework-based textual coding of the program instructions will be examined to identify the extent to which copyright is likely to arise. Second, as a precursor to evaluating exclusive rights in the non-textual elements of the AI program, the implications for protectability that the common reliance on toolkits, public domain knowledge, and heuristics in building ML models will be considered. The final subpart will assess whether copyright inheres in the non-literal elements of a data-specific trained AI software model.

4. Literal Copyright Protection

At the literal level, framework-based human programming can typically be divided into two broad categories. The first involves the data scientist writing short calls for the pre-written sub-programs provided by the software library. The second category of programming entails stipulating the hyperparameter values, often numerically, that define the data-specific structure and / or functioning of the model.

When writing code to implement a model with an AI framework, as a rule most of the expression is pre-defined by the vocabulary and syntax of the toolkit. Python in particular dominates.[955] Taking

[953] For the purposes of this analysis only copyright and *sui generis* database rights are taken into consideration. Other IP incentives include patents, trade secrets, contract law, as well as non-legal incentives such as first-mover advantage etc.
[954] William Fisher, 'Theories of intellectual property' in Stephen Munzer (ed), *New Essays in the Legal and Political Theory of Property* (CUP 2001).
[955] *State of Data Science and Machine Learning 2022* (Kaggle 2022) 14.

Keras [956] as one example of an open source framework, it interacts with TensorFlow to streamline the creation of neural networks. As a library, it provides a pre-programmed collection of functions and topologies for data scientists which, once the data has been prepared, only requires "metaprogramming"[957] in order to execute the training of the model.

To illustrate framework-based programming in more detail, and to aid the evaluation of the extent to which copyright is implicated at the textual level, the code of two different models created with the aid of an AI toolkit will now be evaluated.

(a) **Decision Tree**[958]

Taking an example of a decision tree that predicts whether medical images depict benign or malignant growths, as can be seen from Fig. 5-1 below, the expression required to train and evaluate the model is remarkably compact. Seven lines of Python achieve the selection of the decision tree algorithm (1); ingest of the training data (2); splitting of the data in to training and testing sets (3); "pruning"[959] of the tree to avoid overfitting to the training data (4); training the model on the test data (5); and then finally evaluation of the performance of the model on the training set (6). Finally, an entirely new and unseen set of images is passed through the model in order to validate how accurate the trained model's predictions are (7):[960]

```
1.      from sklearn.tree import DecisionTreeClassifier
2.      cancer = load_breast_cancer()
3.      X_train, X_test, y_train, y_test = train_test_split(cancer.data, cancer.target,
        stratify=cancer.target, random_state=42)
4.      tree = DecisionTreeClassifier(max_depth=4, random_state=0)
5.      tree.fit(X_train, y_train)
6.      print("Accuracy on training set: {:.3f}".format(tree.score(X_train, y_train)))
7.      print("Accuracy on test set: {:.3f}".format(tree.score(X_test, y_test)))
```

Fig. 5-1. Source Code for Training and Testing a Decision Tree in Sci-kit Learn

Whereas, as explained in Chapter 1, much labour and effort goes into collecting and labelling training images, data cleansing, ensuring the processing power of the infrastructure reflects the task at hand, etc,[961] by contrast, as part of the model implementation process, the writing of source code is noteworthy for its brevity.

[956] https://keras.io
[957] The term metaprogramming is used to signify that code is written at a high level using very brief and pre-determined commands which engage other pre-provided functions within the programming environment.
[958] Despite legal commentary on AI almost exclusively focussing on neural networks, the topmost common machine learning algorithms are linear regression / logistic regression (80.3%), decision trees / random forests (74.1%) and gradient boosting machines (59.5%). Kaggle, supra 918, 33.
[959] A data compression technique where nodes on the tree are removed in order to improve generalisation of the model to unseen data.
[960] An amended scenario to that in Andreas Müller and Sarah Guido, *Introduction to Machine Learning with Python* (O'Reilly 2016) 75
[961] See Fig.1-2, Chapter 1.

Taking the decision tree as a typical example of metaprogramming in AI toolkits,[962] the literal textual elements of the source code will unquestionably not create any new copyright. The vast majority of the code lacks originality as its expression is almost entirely dictated exogenously by the vocabulary and syntax of the toolkit. Thus, pursuant to the merger doctrine in the US[963] and the functionality of a highly analogous body of law in Europe,[964] providing no leeway for creativity and the personal stamp of the programmer, it will fail to fulfil the requisite criteria for copyright protection.

Turning from the metaprogramming of a series of short calls to the activity of hyperparameter setting, unlike text which when combined in an original manner attracts copyright, the use of numerals as commands (e.g. In this example random_state = 42 / 0) (max_depth=4))[965] will not constitute protectable subject matter for overlapping doctrinal reasons.

First, this is because similar to the short calls examined in the previous paragraph, it is determined endogenously entirely by the programming language. As a consequence, lacking original creative character, it is ineligible for copyright protection. Even if such a line of argument was found to be invalid,[966] from the viewpoint of copyright jurisprudence, short expressions are *de minimis* and as such fail to reach both the author's own intellectual creation threshold in Europe,[967] and the "minimal degree of creativity" required in the US pursuant to the landmark copyright ruling in *Feist.*[968] Further doctrinal reasonings are also implicated. In the US, such truncated[969] forms of expression which directly instruct a computer how to function would unquestionably be non-protectable, falling simultaneously under the short words and phrases doctrine[970] and §102(b) of the US Copyright Act as

---

[962] Kaggle, supra 918 (showing that decision trees and random forests (an aggregation of individual decision trees) are the second most commonly used machine learning algorithm after linear or logistic regression, with 74.1% of respondents reporting using them.)

[963] §5.2. See also Pamela Samuelson, 'Reconceptualizing Copyright's Merger Doctrine' (2016) 63 J Copyright Soc'y USA 417 (outlining 150 cases involving merger.)

[964] *BSA*, 49. *FAPL* para 98) 1079, *Football Dataco* (para 39), *Cofemel* (para 31), *Brompton Bicycle Ltd* (para 24) (determining that where external constraints such as rules or technology constrain expression to the extent that the idea and expression have become indissociable, that such works will not be subject to copyright protection.) See also in *SAS v WPL* the Advocate General Opinion, later confirmed in the final ruling para 40 (implying that the architecture of a software program has merged with the idea because the combination of sub-programs is "dictated by a specific and limited purpose". i.e. The purpose for which the software is created.) 54

[965] Random_state controls the shuffling of the data before the split and max_depth instructs the model how many times to split before making a prediction.

[966] It is of note that controversially a number of US rulings in the second and ninth Circuits have upheld copyright protection for numbers. See Justin Hughes, 'Size matters (or should) in copyright law' (2005) 74 Fordham L Rev 575. cf. According to the Courts in both cases the valuations deemed protectable were however the result of a process that involved some creativity and, unlike the textual hyperparameter code in machine learning software libraries, there was no direct and self-evident correlation between the number and the information it represented. In *CCC Information Inc v Maclean Hunter Market Reports Inc*, 44 F 3d 61 (2d Cir 1994) 67 ("CCC Info") the Second Circuit found that the numerical "valuations themselves are original creations" as they were predictions based on at least fifteen other considerations. In *CDN Inc v Kapes,* 197 F 3d 1256 (9th Cir 1999) 1262, on appeal, the Ninth Circuit in keeping with *Maclean*, held that individual coin valuations "contain sufficient originality to sustain copyright protection".) In the case of *CCC Info,* Judge Leval found car valuations subject to copyright as based on different data sources and "professional judgement and expertise."

[967] *Infopaq I* ("Words as such do not, therefore, constitute elements covered by the protection.") 46. See also *SAS v WPL* ("… words, figures or mathematical concepts which, considered in isolation, are not, as such, an intellectual creation of the author of the computer program.") 66.

[968] *Feist Publications Inc v Rural Telephone Services Co,* 499 US 340 (1991) 348, 359, 362, 363.

[969] *Reductio ad absurdum* it appears clear there must be a link between length of code and whether it can be rendered unprotected as a procedure. If this were not the case, pursuant to Article 9.2 of TRIPS all software would be rendered unprotected as a method of operation.

[970] Title 37 Code of Federal Regulations, Material not subject to copyright §202.1 (a).

an unprotected "system" or "method of operation".[971] Likewise, in Union Law, pursuant to Article 1.2 of the Computer Programs Directive, and the CJEU ruling in *SAS v WPL*,[972] short function-inducing commands will constitute unprotected ideas.[973]

(b) **Convolution Neural Networks**

Our inquiry into copyright at the textual level of AI software now shifts to a more complex ML topology. While the construction of neural network models involves more convoluted software, and consequently affords the programmer with more choices as to the combination of sub-programs to use or which hyperparameters to select,[974] nevertheless, for the same doctrinal reasoning as the decision tree example, original creative expression in the code will rarely be present at the textual level.

As shown below, textual expression in the inputted code will frequently be *de minimis* in nature. It will also be defined by the programming language and toolkit being utilised. Furthermore, for the reasons explained, even if numeric hyperparametric settings which shape the training and ultimate structure of the trained model are not viewed as inherently beyond the scope of copyright,[975] as non-protected functional methods of operation that initiate a process, they most certainly will.

Despite the heavy reliance on toolkits, developers may still be required to write their own supplemental code which, if original, will be eligible for protection. Examples of this might include code to facilitate program interoperability, model monitoring, visualisation, ingesting and transforming data. Such programs however fail to encompass the core economic value of the trained and optimised model, as they do not directly implicate the implementation of the model algorithm and hence the learning outcome. From the viewpoint of economic effort, this therefore leaves the core focus of the investment at the textual level unprotected by copyright law, and absent of other legal or technological measures, *a priori* free for second-comers to reproduce with none of the upstream investment.[976]

---

[971] See Chapter 4, subsection 3.2. See also discussion in Chapter 4 subsection 3.4 concerning *Google LLC v Oracle America Inc* (2021) which ruled that commands determined by broad definitions of what constitute a programming language are excluded from copyright protection.
[972] Supra 934 ibid. (The CJEU excluding individual commands from copyright protection.)
[973] See also Article 2, WCT.
[974] See definition at supra 916. For example hyperparameters of a neural network include the model choice itself, the number and connectivity of the layers (architectural hyperparameters) and in the training phase the number of examples of data processed before the model updates (batch size), the number of times to run the data through the model (epoch), the setting the learning rate for the model in the optimisation software etc (functional / non-architectural hyperparameters).
[975] Supra 966.
[976] *batch_size* sets the number of images processed before the model updates. *num_classes* sets the number of image classes the model divides the data into. *learning_rate* determines the speed at which the model updates the weights in the neurons thus affecting ultimately how well the model can generalise, and *num_epochs* established the number of times all the data runs through the models.

To further illustrate the weak and unsystematic protection afforded to toolkit-derived neural networks at the textual level, the source code of a type of neural network widely used for audio analysis and image classification known as a convolutional neural network (CNN) will be examined.[977] Immediately evident from Figs. 2 and 3 below is that the length of code is extremely short and, as with the decision tree example, this is grounded in the fact that the expression in the code is not original but defined by the programming language and AI toolkit.

The task of this CNN model is to extract and ultimately recognise common features from the images which it has learnt on in the training phase, so that it can make accurate classifications of unseen pictures once trained. It does this by learning how to recognise standard features associated with the training images. Much like a narrowly focused torch beam in a pitch-black room gradually sweeping across a large image on a wall, a relatively small number of pixels are analysed at any one time. These separate snapshots of the data are then aggregated together to enable the final prediction.

In more technical terms, the input data (i.e. the values representing the pixels) are multiplied by an array of weighted neurons,[978] called a kernel, and as the kernel works across the unseen data, where a feature the program has been trained to recognise is observed, its presence is combined into a number of different feature maps holding the locations where a relevant shape, shading, line etc is found to be present. This filtration process is repeated multiple times, creating varying feature maps, each representing a different attribute elicited from the data. Once this information has been collated together, using the max pooling algorithm in the final layer of the model, the image is then classified according to the classes of data that have been specified by the programmer.[979]

To implement the CNN, having imported the necessary functions into the workspace:[980]

```
import torch
import torch.nn as nn
import torch.nn.functional as F
```

Fig. 5-2. Importing the Model

the model is selected and initialised (_init_):

[977] Source: Defining a Neural Network in PyTorch <https://pytorch.org/tutorials/Recipes/recipes/defining_a_random_neural_network.html>
[978] A neuron or a node is a calculation common to all types of neural network that takes in a defined number of input values, and applies a weighted calculation to them before making a non-linear transformation (activation function) to the sum. This is then passed on to the next node, with the sole exception of the output layer where no further outputs are necessary. (see Fig. 7).
[979] Rikiya Yamashita et al, 'Convolutional neural networks: an overview and application in radiology' (2018) 9 Insights into Imaging 4 611.
[980] The code in Figs. 2 & 3 are derived from tutorials in www.pytorch.org.

```
class Net(nn.Module):
    def __init__(self):
      super(Net, self).__init__()
      self.conv1 = nn.Conv2d(1, 32, 3, 1)
      self.conv2 = nn.Conv2d(32, 64, 3, 1)
      self.dropout1 = nn.Dropout2d(0.25)
      self.dropout2 = nn.Dropout2d(0.5)
      self.fc1 = nn.Linear(9216, 128)
      self.fc2 = nn.Linear(128, 10)

    def forward(self, x):
      x = self.conv1(x)
      x = F.relu(x)

      x = self.conv2(x)
      x = F.relu(x)

      x = F.max_pool2d(x, 2)
      x = self.dropout1(x)
      x = torch.flatten(x, 1)
      x = self.fc1(x)
      x = F.relu(x)
      x = self.dropout2(x)
      x = self.fc2(x)

      output = F.log_softmax(x, dim=1)
      return output
```

Fig. 5-3. Defining the Architecture of a Convolutional Neural Network in PyTorch

To build the CNN, like all neural networks, the data scientist is required to specify the number, types, and sequence of layers in the model architecture. In this example, the model comprises two convolutional layers which apply a filter to create a feature map (conv), one max pooling layer (max_pool), two dropout layers (dropout), and two fully connected layers (fc). To avoid the model overfitting to the training data, various techniques are used by the developer. In this instance: deploying a pooling layer after the convolutional layers (which has a dual function of reducing the data and complexity of the model and helping extract the most salient features), as well as removing neurons (dropout)[981] before passing the data into the model for training. As with the decision tree example, no further programming of the training algorithm itself is required as these sub-programs are native to the PyTorch environment.

As in any AI model, the hyperparameters which determine how the model is trained, structured, and ultimately functions have to be selected and inputted by the programmer. In the case of neural networks, this includes choosing values routinely expressed numerically (see Fig. 5-4).[982] Examples of hyperparameters that require tuning include number of layers and nodes in a neural network, the learning rate in an ANN, the depth of a decision tree, c or gamma values in an SVM model, etc. More

[981] Dropout is used to prevent overfitting of a neural network as surrounding neurons to those removed have to provide a calculation for the "missing" neuron. This results in the model becomes less sensitive to a particular calculation at each node, meaning the model is better able to adapt to unseen data.
[982] See supra 976.

specific to a CNN is the size of the kernel (3 in the above example), the number of features outputted by the convolutional layers (1 = 32, 2 = 64), and the number of steps across the data the kernel takes when adjusting its position before making a new calculation (stride).[983]

```
batch_size = 64
num_classes = 10
learning_rate = 0.001
num_epochs = 20
```

Fig. 5-4. Examples of Hyperparameter Tuning in a Neural Network

After passing the data through the convolutional layers, as is common in image classifiers, the non-linear activation function ReLU is applied which removes negative values from the image, thus making it computationally simpler for the machine to process. Finally, the softmax program applies a probability to each class of image, thus providing the means upon which the model can classify an image type in the output layer.

As is evident again from this example, despite the somewhat more complex nature of the source code, unless a skilled programmer qualitatively or quantitatively adapts or augments the underlying code with their own original commands, at the literal level the short and highly regimented nature of the toolkit metaprogramming environment militates against copyright protection. As with the decision tree, the expression in the commands is dictated exogenously by the syntax and vocabulary of the programming language and the ML framework and therefore lacks originality. Simultaneously, the truncated and highly standardised mode of expression required by the framework specification, acts to reinforce that the code should be rendered an unprotected method of operation pursuant to Article 9.2 of TRIPS.

Arguably the more skilled and therefore economically more valuable activity undertaken by the data scientist is the "tuning" of the model's numerous hyperparameters. Depending on their particular role, the hyperparameter settings inform the functioning and / or structure of the model. As discussed in the previous subpart, at the literal level the *de minimis* and functional nature of the numerical expression used for setting the hyperparameters renders them beyond the reach of copyright protection according to TRIPS, and in the US pursuant to §102(b) of the US Copyright Act.[984]

[983] A kernel or filter is a matrix of weights which convert inputs into an output allowing it to extract features from an image.
[984] See Chapter 4, §4.1-2. See also *Exxon Corp v Exxon Insurance Consultants International Ltd* [1982] RPC 69 (holding there is no copyright in a single word.)

In sum, while model programmers may need to augment the toolkit and write new software for loading, data transformation, etc, owing to the near universal use of machine learning libraries, the core of the model's code at the textual level will frequently fall beyond the scope of copyright. In line with jurisprudence, the highly regimented and sparse programming vocabulary, acting as it does as a method of operation, renders the expression ineligible for protection. Thus, contrary to traditional software, at the textual level, copyright protection for much of the investment made in building ML models is either absent, or frequently will only protect peripheral or fragmentary and uneven elements of the code.

5. Non-Literal Copying

De lege lata given that the majority of models created using AI frameworks will rarely attract substantive copyright protection at the textual level, the extent to which the structure and arrangement of the model topology is the result of authorial creative choice constitutes a vitally important legal and economic policy question.[985] As implied by Drexl, in an environment of policy activism where technology driven work-specific copyright protection is not uncommon, irrespective of the ensuing economic deadweight loss, weak copyright protection may yet spur regulatory interventions in the field of AI.[986]

As repeatedly observed by the courts[987] copyright may in theory "extend[] beyond a literary work's strictly textual form to its non-literal components." Thus, where the "choice, sequence and combination"[988] of sub-programs and other elements in software expresses authorial originality, exclusive rights may be invoked. While within the framework of copyright jurisprudence, this is conceptually coherent, as Karjiker notes, what exactly constitutes "non-literal" or "non-textual" copying lacks definitional precision.[989] Compounding this further is the fact that – where it has been considered – many of the most prominent rulings have found against it.[990] While these judgements can

[985] The point at which a model fulfils the fixation requirements of copyright will vary according to national law. See Drexl, Hilty et al, *Position Statement on IP Law and Artificial Intelligence* (Max Planck Institute 9/4/2021) (stating that for eager learning models "it might be very difficult or even impossible to establish the moment when copyright protection may arise.") §17. See also *Cofemel* in which, citing *Hengelo* the CJEU ruled expression must be "identifiable with sufficient precision and objectivity" 32. cf. Pertinent to a discussion on copyright protection for a model in the training phase is *Hengelo* contrasting taste as subjective and imprecise with pictures, musical works etc which can be evaluated with both precision and objectiveness, where the court states that fixation does not have to be in permanent form (para 40).
[986] Reto Hilty, Jörg Hoffmann and Stefan Scheuerer, 'Intellectual property justification for artificial intelligence' 50 ff. in Jyh-An Lee, Reto Hilty, and Kung-Chung Liu (eds), *Artificial Intelligence and Intellectual Property* (OUP 2021)
[987] Rulings acknowledging non-literal copyright in Europe include: *SAS v WPL* (Opinion) 50, (final judgement) 25, 67. *Navitaire Inc v Easyjet Airline Company and Bulletproof Technologies* [2004] EWHC 1725 (Ch), *John Richardson Computers Ltd v Flanders* [1993] FSR 497*, Ibcos Computers v Mercantile Highland Finance* [1994] FSR 275. In the US: inter alia *Computer Associates International Inc v Altai Inc, Whelan Associates Inc v Jaslow Dental Laboratory Inc, Lotus v Borland.*
[988] *SAS v WPL*, 66.
[989] Sadulla Karjiker, 'Copyright Protection of Computer Programs' (2016) 133 South African Law Journal 1 51, 60 ("non-literal copying is not a term of art, and there have been various definitions of what constitutes the non-literal components of a computer-program.")
[990] Moon, supra 938, 921(citing inter alia *Navitaire v Easyjet* (architecture determined by the purpose and function of the software or their function as part of the programming language), *Altai* (structure determined by the operating system and APIs), *Cantor Fitzgerald International v Tradition (UK)Ltd* [2000] RPC 95 (commonplace structure). See also Pamela Samuelson, supra 878 (citing rulings where copyright in the structure of computer software has been rejected by the courts including *Lotus v Borland* (as a method of operation) and

by and large be attributed to the defendants' lack of access to the plaintiff's code,[991] this was not the case in the recent ruling in *Google v Oracle*.[992] Calling into question the seeming judicial consensus, of particular note in this case was the Supreme Court's clear challenge to the very notion of copyright protection for a program's structure, sequence, and organisation (SSO). Using the analogy of organising paperwork into "cabinets, drawers, and files",[993] it designated the "task division and organization"[994] of Oracle's APIs as "uncopyrightable ideas".[995] Arguably seeking to challenge longstanding jurisprudence, the Supreme Court emphatically concluded that the arrangement and organisation of a computer program is something that "no one claims is a proper subject of copyright."[996]

Whereas it remains to be seen whether this dramatic reversal by the Supreme Court of decades of copyright jurisprudence will hold sway in the US, nevertheless, in order to reach an informed understanding of copyright's application to AI model architectures, a full evaluation of its theoretical operativity is still nevertheless indispensable.

This section will proceed in three parts. First, the types of structural choices that a data scientist may make that potentially qualify for copyright protection will be delineated from those that are unprotected. Second, the implications for copyright protection that arise from the AI knowledge commons, where ML software architectures are freely created and shared in the literature as well as in toolkits, will be evaluated. Third, the role of heuristics in model production is examined. Having situated model architecture design in its production context, in the final subpart the extent which the selection of hyperparameters in the code can be considered appropriate subject matter for IP protection at the data and task-specific model-level is discussed.

6. AI Model Architectures and Hyperparameters: A Dichotomy of Functional and Structural Choices

Copyright will afford protection to software architectures if choices, creative or derived from other criteria,[997] are made by the author in the selection and organisation of sub-elements that make up the

---

Google LLC v Oracle Inc 872 F Supp 2d 974 (ND CAL 2012) (as a system), *Plains Cotton Co-op v Goodpasture Computer* 807 F 2d 1256 (5th Cir 1987) (owing to the purpose of the software.)

[991] *SAS v WPL*, *Navitaire*, *Lotus v Borland, Plains Cotton* (with the exception of one subroutine*), John Richardson Computers Ltd v Flanders, Cantor Fitzgerald.*

[992] *Google LLC v Oracle America Inc* (2021).

[993] Id.1201.

[994] Id.1202.

[995] Ibid.

[996] Id.1201.

[997] Estelle Derclaye, 'Assessing the impact and reception of the Court of Justice of the European Union case law on UK copyright law: what does the future hold?' (2014) Revue internationale du droit d'auteur 240 1 (stating that after a slew of cases including *Infopaq I, BSA, Painer, FAPL, Football Dataco* it has become clear that in the UK "neither labour nor skill is required, but only free and creative choices.") 12. Moreover, pursuant to the European Union (Withdrawal) Act 2018 s.2&3, in the absence of a change to legislation or case law to the contrary, copyright rulings of the CJEU constitute "retained law". The continuing application of the CJEU author's own intellectual creation criterion in the UK has been reaffirmed most recently in *THJ Systems Limited & Anor v Daniel Sheridan & Anor* [2023] EWCA Civ 1354.

work.[998] Thus, if no countervailing legal doctrines apply, in Europe, architectural choices that involve the original and independent "choice, sequence and combination"[999] of sub-elements of software will induce proprietorial copyright entitlements.[1000] In the US, ostensibly the same copyright standard is also operative, applying to the "selection and arrangement" of the subparts of a program.[1001]

To evaluate the degree to which copyright law may protect the non-literal structural elements of a trained AI model, first an analysis of the type of programming choices the author of a machine learning model may make is briefly required. While fully accepting the somewhat artificial and overlapping nature of the division, hyperparametric decisions made by the designer of a task-specific model can be divided into two broad categories. The first of these are purely functional in their objective and involve the author directing a subset of functions how to operate. For example, in a K-Nearest Neighbour algorithm, instructing the model as to the number of clusters to divide the data into, determining how far the filter in a CNN moves in each step when analysing pixels (stride), setting the learning rate and number of epochs or what size to divide the data into for ingest into the model (e.g. in a CNN the kernel size) and so on. As commands which instruct the model *how* to process data, bringing about no immediate direct effect on the resultant structure of the AI software beyond their literal – and sometimes temporary – presence in the code,[1002] what may be called "non-architectural hyperparameters" are immaterial for the purposes of evaluating non-literal copyright in AI models.

A second category of hyperparameters, however, are more relevant to the discussion as they have a one-to-one direct impact on the architecture of the model. Typical architectural hyperparametric choices a model author will make that fall into this category include the depth or number of layers of a decision tree, random forest, or neural network,[1003] the combinations of layer types in an ANN (e.g. Fig 4 above), the numbers of neurons[1004] in each layer, the connectivity between the neurons and the layers, and suchlike. Moving beyond the structure of one individual model, a programmer may also decide to combine differing types of models in a process known as "ensemble learning".[1005]

---

cf. Elizabeth Judge and Daniel Gervais, 'Of silos and constellations: comparing notions of originality in copyright law' (2010) 27 Cardozo Arts & Entertainment Law Journal 375 (reviewing common law jurisdictions, the authors note that Canada requires "skill and judgement" while Australia, South Africa and New Zealand retain the former UK test of skill, labour and judgement.)

[998] In the case of an AI model the sub-elements are the "human-written" code and the parameters of the model which are derived primarily from the training data.

[999] *Infopaq I*, 65. *SAS v WPL,* 67.

[1000] cf. *Google LLC v Oracle* (see discussion in subpart 5).

[1001] *Feist* 349.

[1002] Functions including those such as learning rate, epoch, batch size etc are only used in the training phase as a rule.

[1003] For example, a neural network can have as few as one hidden layer. It requires two layers to be considered deep learning (John D Kellerer *Deep Learning* (MIT Press, 2019)) and in theory can have as many as needed, however the larger the model the longer and more expensive it is to train. In terms of disclosed information, the largest natural language processing neural network to date, GPT-4, has 120 layers in total.

[1004] A computational unit which activates an output based on inputs. Géron, op. cit. 920, 256 f.

[1005] For example, see Zhice Fang et al, 'A comparative study of heterogeneous ensemble-learning techniques for landslide susceptibility mapping.' (2021) 35 International Journal of Geographical Information Science 2 321 (incorporating four different ensemble techniques combining convolutional neural network, recurrent neural network, support vector machine, and logistic regression to predict the susceptibility of landslides in a given area.)

6.1 The Open Norms of AI

This subpart now moves to assess the background, industrial context, and influence that the open AI programming environment has on the originality of program implementation and design. This is indispensable before proceeding because, as discussed in Chapter 4, according to jurisprudence, commonplace stock expression falls beyond the reach of *in rem* protection by copyright law. Thus, choices in the architecture which are well-known and run-of-the mill will, depending on the jurisdiction, either lie in the public domain, or be subject to a defence against claims of infringement.[1006]

The free availability of open-source models in machine learning toolkits[1007] creates a programming environment where not only the software but the underlying theory of individual model architectures are often part of the knowledge commons. Thus, which models, combination of subprograms, and what structural choices to make in defining the model frequently constitute an industry standard. This acts as a strong inhibitor to copyright. From the standpoint of jurisprudence, where trained or pre-trained architectures are well-known or provided "as is" in the toolkit, they almost certainly constitute a commonplace industry standard and thus devoid of originality pursuant to scènes à faire and the similar doctrines which exist in Europe.[1008]

The reason why this open programming environment has become the norm stems in no small part from the AI strategy of the "technology moligopolies".[1009] As a consequence of the indomitable market advantage that has accrued to them from network effects, proprietorial ownership of big data and, as a result, their unparalleled technological infrastructures, means an open approach to knowledge sharing in AI has a number of strategic advantages. As Thompson contends, given the seemingly unassailable market power of the tech giants, defensive IP protection of each and every element of their business portfolio is strategically less important than being able to directly and indirectly accrue benefits from an open AI environment.[1010] Through their support of open toolkits,[1011] Meta, Google, Microsoft etc, are able to leverage the free flow of information and collective knowledge of machine learning experts as they work in environments reflective of the superplatforms' own ML infrastructures. Having set the standards for AI programming globally, not only have they ensured that the software engineers of today and tomorrow arrive on their doorsteps already trained in

[1006] See Chapter 4, subsection 5.1 - 4 regarding doctrines that render commonplace expression beyond the scope of protection albeit, reliant on different doctrinal underpinnings according to the jurisdiction.
[1007] Examples of platforms supporting AI include Hugging Face, Kaggle, wandb.ai, GitHub etc.
[1008] See Chapter 4, subsection 4.1-4.
[1009] Nicholas Petit, *Big Tech and the Digital Economy: The Moligopoly Scenario* (OUP, 2020) (a portmanteau term derived from monopoly and oligopoly.)
[1010] Ben Thompson, 'TensorFlow and Monetizing Intellectual Property' (12 November 2015) https://stratechery.com/2015/tensorflow-and-monetizing-intellectual-property/ (discussing the commercial strategy underpinning Google's decision to make TensorFlow open source.)
[1011] Google (TensorFlow, JAX), Meta (PyTorch), Baidu (PaddlePaddle) and Microsoft (Cognitive Toolkit, Synapse ML).

the systems deployed internally, but more broadly that their corporations remain strategically indissociable from the operation of the whole AI domain.[1012]

Over and above ML frameworks, another key strand of the AI knowledge commons is the open access availability of scholarly literature. Unsurprisingly, much of the high-impact, publicly accessible AI research originates from academic researchers either independently or in partnership with Big Tech.[1013] Evidence suggests that there is a relatively direct relationship between academic model designs and their industrial application. Illustratively, some of the most widely adopted AI architectures – Generative Adversarial Nets (GAN)[1014] and AlexNet (an early type of high performing CNN)[1015] – originated from the universities of Montreal and Toronto. Moreover, not only do some of the world's top AI scientists divide their time between academia and industry,[1016] but the widespread availability of scholarly articles from institutional repositories supports an open learning environment where direct knowledge transfer of scientific knowledge, including ML designs and architectures from academia is common.[1017]

The nexus between scholarly literature and industrial investment is further borne out by a 2019 study by WIPO.[1018] With the exception of "rule learning", as can be seen from Fig. 5-5, whereas the upswing in patent registrations for AI technologies around the mid 2010s is symptomatic of a

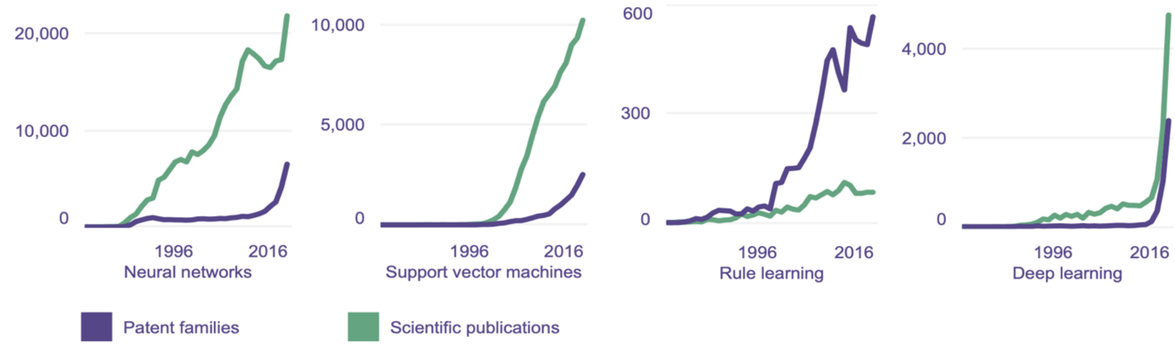

Fig. 5-5. AI Patent Family and Scientific Publications by Publication Year [1019]

[1012] Thompson, supra 1010.
[1013] Sergi Castella i Sapé, 'Must read: the 100 most cited papers in 2022 (*Zeta Alpha*, 8 Mar 2022) https://www.zeta-alpha.com/post/must-read-the-100-most-cited-ai-papers-in-2022 . Based on a study of the top 100 AI articles cited in 2022 (sourced from Semantic Scholar, Google Scholar and Zeta Alpha Platform) 21 articles were from purely academic institutions, 41 the result of academic-private partnerships and 38 from private corporations.
[1014] Ian Goodfellow et al, 'Generative adversarial nets' (2014) 27 Advances in Neural Information Processing Systems 1.
[1015] Geoffrey Hinton et al, 'Imagenet classification with deep convolutional neural networks' (2012) 25 Advances in Neural Information Processing Systems 1.
[1016] Well known data scientists Geoffrey Hinton worked at Google while based at the University of Toronto, and Fei-Fei Li worked for Google while at Stanford. Ruslan Salakhutdinov worked as Director of AI at Apple while based at Carnegie Mellon University.
[1017] José Guadix Martín, Milica Lilic, Marina Rosales Martínez, *AI Knowledge Transfer from the University to Society* (CRC Press 2022) (providing practical case studies of AI knowledge transfer from universities to industry in the following sectors: health and social welfare, renewable energies, energy efficiency, sustainable construction, transport, mobility and logistics, tourism, culture and leisure; agribusiness and digital economy.)
[1018] *WIPO Technology Trends 2019: Artificial Intelligence* (WIPO 2019) 45 ff.
[1019] Ibid. 45.

number of factors coalescing together, including the growth of big data,[1020] new transformative algorithms,[1021] and the uptake in use of GPUs,[1022] the WIPO Report evidences the availability of a large pool of freely accessible research also made a significant contributory factor.

From a copyright perspective, even the most cursory evaluation of technical AI literature reveals that as a matter of routine copyright-protectable architectures are freely shared by data scientists. For example, while the topology of AlexNet was novel and would unquestionably have reached the originality criteria for protection at the non-textual level,[1023] having won four AI competitions between May 2011 and September 2012, its architecture quickly became an unprotected industry standard.[1024]

Arguably underlining the commonplace and stock nature of even trained AI model designs is the comparative straightforwardness with which model extraction attacks may take place.[1025] Thus, owing to the commonality of AI model architectures, the simple querying of model responses through a firewall with an API,[1026] or even monitoring the electromagnetic waves emitted from the hardware, reveals enough to facilitate the unauthorised reproduction of the core of the model.[1027]

In sum, original architectures will frequently enter the knowledge commons through the scholarly literature or as code deposited in an open repository or library. Even where a new model's topology is *ab initio* protected by copyright, as discussed in Chapter 4, once the model architectures is adopted widely by the community, lacking originality, from a copyright perspective it will become an unprotected industry standard.[1028]

6.2 Heuristics and Commonplace Expression

[1020] Saša Batistič and Paul van der Laken, 'History, evolution and future of big data and analytics: A bibliometric analysis of its relationship to performance in organizations' (2019) 30 British Journal of Management 2, 229 (depicting a dramatic increase in primary papers on big data after 2010.)
[1021] For example, Goodfellow, supra 1014. Geoffrey Hinton et al, 'A fast learning algorithm for deep belief nets' (2006) 18 Neural Computation 7 1527. Geoffrey Hinton et al, 'Deep belief networks for phone recognition' (2009) NIPS Workshop on Deep Learning for Speech Recognition and Related Applications 1, 9.
[1022] Rajat Raina et al, 'Large-scale deep unsupervised learning using graphics processors' (2009) Proceedings of the 26th Annual International Conference on Machine Learning, 873 (a seminal paper prompting the uptake of GPUs.)
[1023] Inter alia*,* AlexNet had more layers than its predecessors, the first CNN to stack convolutional layers directly on top of each other and deployed the activation function ReLU in place of tanh which until then had been standard. Géron, op. cit. 920, 367.
[1024] Hinton, supra 1015.
[1025] Xing Hu et al, 'Neural network model extraction attacks in edge devices by hearing architectural hints' (2019) CoRR, arXiv: abs (1903) 11 ("The standardization through the whole stack of neural network system [sic] facilitates the model extraction.")
[1026] Thomas Schneider et al, 'Efficiently stealing your machine learning models' (2019) Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society ("MLaaS schemes are inherently vulnerable to extraction attacks.")
[1027] Lejla Batina et al, 'Reverse engineering of neural network architectures through electromagnetic side channel' (2019) 28th USENIX Security Symposium.
[1028] See Chapter 4, subsection 4.1 - 4 (discussing scènes à faire and similar doctrines.)

Intertwined with the factors outlined above is the heavy reliance by data scientists on heuristics.[1029] Although different definitions exist, including simple trial and error, a heuristic in machine learning can be best defined as a rule of thumb – something which is derived from the previous experience of the individual data scientist or others who work in the same domain.[1030]

The authorities have routinely recognised that commonly adopted programming practice such as those informed by well-known heuristics act as a limitation on copyright. As observed by the court in *Altai:* "in many instances it is virtually impossible to write a program to perform particular functions in a specific computing environment without employing standard techniques. This is the result of the fact that a programmer's freedom of design choice is often circumscribed by extrinsic considerations such as … widely accepted programming practices within the computer industry."[1031] Likewise, while not included in the final Directive, the Proposal for a Computer Program Directive underscores that software has the propensity to lack expressive originality owing to the industrial nature of its production.[1032]

To sum up therefore, choices in the organisation and structure of the software frequently will fail to qualify for protection owing to the conjunction of differing factors. With reference to copyright doctrine, owing to the heavy reliance on toolkits, industry standards, and the widespread sharing of architectures in the literature, many base models are therefore likely to fall beyond the scope of copyright protection pursuant to scènes à faire in the US[1033] and kindred doctrines operative in Europe.[1034] Whereas the industrial programming backdrop to software production is nothing novel in and of itself, the overwhelming predominance of open norms in ML markets evidently constitutes a step-change in the routine applicability of copyright protection to software. As a consequence, it will behove the courts to look more keenly for original copyrightable expression in the architecture of AI models than has been the case up until this point with its software predecessors.[1035]

6.3 Trained Model Architectures

This subpart now turns from the background environmental factors which will often induce weak copyright protection for the most common AI base architectures, to investigate the granular design details of individual model building. This is to answer the question of whether copyright may inhere

---

[1029] See for example, Josef Drexl and Reto Hilty et al, 'Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective' (2019) Max Planck Institute for Innovation & Competition Research Paper No. 19-13 (discussing the reliance on heuristics in AI model building.) passim.
[1030] Hugh Schwartz, 'Heuristics or rules of thumb' (2010) Behavioral Finance: Investors, Corporations, and Markets 57 ff.
[1031] *Altai*,709-10.
[1032] Proposal for a Council Directive on the legal protection of computer programs COM(88) 816 final – SYN 183. Part One.§ 2.7. Part Two Article 1.4 (a).
[1033] See Chapter 4, subsection 4.1.
[1034] See Chapter 4, subsection 4.4.
[1035] See Annex I for questions a court considering claim of unauthorised reproduction of ML software should consider.

in the unique architectures that emerge in response to training from the oftentimes generic, and as a consequence unprotected, starting topology.

Once the higher-level choices for the untrained model, including the algorithm type, which combination of models and functions to use, and the sequence of sub-programs have been defined, the data scientist is required to set starting values for many of the hyperparameters in a process known as initialisation. This is before training on the data commences.

As outlined in subpart 6, hyperparameters may be divided into two distinct typologies. First, are the hyperparameters which are functional in purpose and therefore do not contribute to the wider architecture of the model. Second are those which directly shape the topology of the trained model. For example: the number of layers in a neural network, decision tree, or random forest; the number of neurones in a hidden layer of a neural network and their connectivity; number and types of filters per layer, etc.[1036] The unmistakeable structural purpose of these sub-elements of the code allows us to refer to these configurations as structural or architectural hyperparameters.[1037]

Important from a copyright perspective, some architectural hyperparameters provide no leeway for creative or original decision making by the model's author. For example, in a neural network, the number of neurons in the input layer will be determined by the data structure and the nature of the model selected. Thus, a model ingesting images of 28 x 28 pixels will require 788 neurons on the input layer. Similarly, the number of neurons in the output layer will be influenced by the target output. Therefore, if the business task is to classify data into ten differing categories, the output layer would also routinely and most obviously consist of 10 neurons.

Connectivity and arrangement between elements of the model can likewise, in certain instances, be indissociable from the model chosen. For example, it may be determined by the choice of a sub-program, such as a pooling function, where multiple input connections are outputted to a single connection. Likewise in the case of a recurrent neural network, connections between the neurons are determined in part by the fact that the model is designed for values to travel both forward and backward through the model. Thus, in all these examples controlled exogenously by the data, task, or algorithm, the expression and the idea merge, rendering them ineligible for copyright protection.[1038]

Not all structural hyperparameters have such a bijective relationship with the data or the model in which they are deployed, however. In a process referred to as optimisation, hyperparameters are

[1036] Ahmed Akl et al, 'Optimizing deep neural networks hyperparameter positions and values' (2019) Journal of Intelligent & Fuzzy Systems 37 5 6665-6681.
[1037] Ibid.
[1038] See Chapter 4 subsection 4.1.

"tuned" so the base model achieves optimal performance without overfitting to the training data. Architectural hyperparameter tuning may involve adding layers of neurons or nodes in a neural network, determining the number of trees or splitting rules in decision trees / random forests, or experimenting with the number of filters or kernel size in a CNN until the performance of the model in the training phase degrades.[1039] This process can be done entirely manually using domain knowledge, heuristics, trial and error,[1040] or with the aid of hyperparameter tuning software.[1041]

Taking the example of a neural network below: as a result of tuning a neural network, and as informed by the GUI, the data scientist can see that having 67 neurons in each of the hidden layers, and dropping 15% of the neurons in the dropout layer, results in the highest accuracy for the model.

iter	target	activa...	batch_...	dropout	dropou...	epochs	layers1	layers2	learni...	neurons	normal...	optimiz
14	0.5764	2.531	285.0	0.4263	0.2522	28.83	2.973	1.467	0.7242	69.48	0.07776	4.881
15	0.768	2.388	921.5	0.8183	0.1198	85.62	1.396	2.045	0.4184	93.33	0.8254	3.507
16	0.7684	1.051	209.3	0.9132	0.1537	87.45	1.19	2.607	0.07161	67.19	0.9688	2.782
17	0.5144	5.936	371.9	0.8899	0.296	79.09	2.283	1.504	0.4811	34.13	0.8683	1.868
18	0.5719	8.757	370.8	0.2978	0.221	21.03	1.06	2.468	0.5033	29.63	0.00893	5.955

Fig. 5-6. Example of Hyperparameter Tuning[1042]

Having established the best combination of architectural settings in the testing environment, the data scientist can then freeze and deploy the model subject to any business constraints that may exist.[1043]

Taking this example, the final structure of the core of the model is therefore not an autonomous creative act but largely determined by its optimal configuration and performance, resulting from the model's interactions with the training and testing data. Absent is the authorial personal stamp in the software that derives from autonomous selection and choice, indispensable for copyright protection.[1044] Instead, the architecture is ultimately defined and dictated by the requirement to

---

[1039] J van den Hoogen et al, 'Hyperparameter analysis of wide-kernel CNN architectures in industrial fault detection: an exploratory study' (2023) 18 Int J Data Sci 423.

[1040] While heavy reliance is placed on software to perform optimisation, hyperparameters can also be adjusted manually. See Zoltan Czako et al, 'AutomaticAI–A Hybrid Approach for Automatic Artificial Intelligence Algorithm Selection and Hyperparameter Tuning' (2021) Expert Systems with Applications 182 115225.

[1041] Petro Liashchynskyi et al, 'Grid search, random search, genetic algorithm: a big comparison for NAS' (2019) arXiv 1912.06059 (comparing popular hyperparameter optimisation algorithms.)

[1042] Rendyk Kernia, 'Tuning the Hyperparameters and Layers of Neural Network Deep Learning' https://www.analyticsvidhya.com/

[1043] Note that deployed models will in many instances still be dynamic as they may require frequent updating and therefore how they are treated from a copyright perspective will depend on fixation requirements in national law. For example, see, Oleg Pianykh et al, 'Continuous learning AI in radiology: implementation principles and early applications' (2020) 297 Radiology 1 6 (outlining AI models which continually learn and update.)

[1044] For example, *Feist Publications Inc v Rural Telephone Services Co,* 499 US 340 (1991) 48 "[C]hoices as to selection and arrangement, so long as they are made independently … and entail a minimal degree of creativity, are sufficiently original." The European Court of Justice in *Painer* stating that copyright applies to a work which "is an intellectual creation of the author reflecting his personality and expressing his free and creative choices." 99.

minimise the generalisation error and create the most accurate outputs. Consequently, the arrangement and structure of the topology does not arise from independent creative efforts but is the result of the model's interaction with the training and validation data as the software analyses repeated data feeds according to iterative combinations of hyperparameter settings. As Karpathy succinctly states, crucial elements of the core software, and by extension its architecture, are not written by an author but "written by the [software] optimization" as the model learns from the data.[1045]

Where the interplay between the architectural and non-architectural hyperparameters learned from the data alone does not inform the specific implementation of the trained model's structure, external business constraints almost certainly will.[1046] As with conventional design, irrespective of the maximal performance of the software in a testing environment, the trained model will need to reflect the technological and business requirements of the environment within which it is deployed. Processing power, latency limitations, the software and hardware environment, customer requirements, and so on will all shape the final architecture of the model. Thus, the expression embodied in the model's topology is largely determined by the training data in combination with infrastructural and efficiency considerations, with the result that such elements will rarely meet the threshold required for copyright protection.

In the United States, pursuant to *Altai*'s characterisation of scènes à faire[1047] and Samuelson's broad portrayal of the merger doctrine,[1048] expression dictated by exogenous influences is either rendered unprotected by copyright outright or enjoys a defence against infringement.[1049] Likewise, in accordance with a long line of rulings from the CJEU,[1050] in the European Union, where expression is constrained by extrinsic factors, it will fail to reach the author's own intellectual creation standard as set forth inter alia in Article 1(3) of the Software Directive.

To conclude both sections on literal and non-literal copyright, while copyright will frequently inhere at the textual level in the underlying, openly licensed software available from an AI toolkit, at the local implementational level – where data-specific models are designed – the picture is far less certain given the almost universal reliance on software libraries. The brief metaprogramming expression in the code designed to call pre-written functions lacks authorial presence as it is almost entirely defined by the programming environment. Where some expressional leeway exists with the framework, the

[1045] Karpathy, op. cit. 53.
[1046] See Chapter 4 subsection 4.1 - 2. See Shemtov, supra 938 (discussing business externalities limiting copyrightability in the context of abstraction, filtration, comparison test.) 129. Samuelson, supra 878(in view of the decision in *Plains Cotton* stating that business considerations and markets affect whether a software program attracts copyright protection.) 1229.
[1047] See Chapter 4, subsection 4.1.
[1048] See Samuelson, supra 963, 425 (defining merger as covering not just instances where possible expressional choices are limited but including efficiency considerations also.)
[1049] See Chapter 4, subsection 4.1-2 (discussing differing views and jurisprudence regarding whether scènes à faire and the merger doctrine are defences against infringement or fall beyond the scope of copyright.)
[1050] Supra 964.

input of the brief often numeric commands falls simultaneously as both *de minimis* expression and, pursuant to Article 9.2 of TRIPS, a non-protected method of operation. Absent substantive original reprogramming, which will in all probability be comparatively rare given that it contradicts the reason for the use of the toolkit in the first place, a combination of these factors will therefore at the textual level frequently render models created with the aid of an AI framework ineligible for copyright protection.

At the non-literal level, the AI open knowledge commons, grounded in the free circulation of scholarly literature and open source libraries of code, acts as a marked inhibitor on the originality of pre-trained AI architectures. Thus, pursuant to copyright doctrines, which withhold protection from stock expression, many of the more well-known base model topologies lack the originality of expression required to attract copyright. Moreover, at the level of the data and task-specific model where the architecture is unique, it is the result of the optimisation process – determined by the interplay of the model's hyperparameters with the training data in combination *ex post* with the environmental constraints of its deployment. Thus, as expression dictated not by the free and creative choices of the author but by data and exogenous business considerations, the core of the AI models will frequently be ineligible for copyright protection unless qualitatively substantive independent creative additions, or restructuring of the base program, are made by a human author. In the final analysist therefore, while the process of training a ML model may be time-consuming and, as a consequence, highly capital intensive, nevertheless as a process highly reliant on data and industrial standards, it frequently leaves no room for the autonomous original choices that copyright is designed to protect. From a normative perspective, as will be discussed below, this opens the door to the question of whether policy intervention aimed at recalibrating copyright regulation of software is desirable or not, and if so, what form it should take.

7. Compilations and the European Database Directive: Software and Model Parameters

This inquiry now shifts away from broad copyright issues concerning textual and non-textual expression in the ML models to focus on the well-defined legal concept of compilations and collective works, including the uniquely European regime of *sui generis* database rights. In particular, this subpart will focus not just on the software but the important variables in the model known as parameters. It is the parameters that encapsulate the learning from the training data and therefore it is these variables that in many ways embody the core value of the model. From an incentives perspective, given that the parameters play a key role in ensuring accurate predictions, an evaluation of whether they are protected by copyright, or in Europe the *sui generis* database right, is indispensable. This subpart will start with an explanation of the technical role played by parameters. Second, the investigation will move briefly to whether copyright inheres in parameters either as stand-alone works or as a compilation. After this, the applicability of the European *sui generis* database right to both the software and the parametric components of AI models will be examined.[1051]

7.1 The Role of Parameters in Machine Learning

Parameters are numeric values integral to the core calculations undertaken by many eager learning models. Automatically generated in the training phase from the interaction between the training data and the optimisation process, they are in essence a numeric embodiment of what is learned by the model during training.[1052]

Different terminologies for parameters are used according to the type of model. In linear and logistic regression algorithms, the parameters are referred to as "coefficients". In common with many models they act to weight the value of the input data, thus making a pivotal contribution to the process whereby the model is able to make an accurate prediction. Similarly, in a decision tree and neural network, the "weights" are values in the model which determine the calculations and output responses given from each node (sometimes referred to as "firing").[1053] By contrast, the "support vectors" in an SVM model are not created and calculated as such; rather, they are selected from the ingested training corpus as the datapoints that best define the dividing hyperplane, thus determining how accurately the trained model is able to classify unseen data.

Taking the parameters in an ANN as an example, they act to weigh the calculation taking place at each node to ensure the model can accurately solve problems based on unseen data. Dependent on the model, artificial neurons comprise mathematical calculations of differing levels of complexity. Despite this, all share the same basic principles:[1054]

i) A connecting link or weighted synapse which "weights" the value of the incoming data;
ii) A summation function which calculates the value of the weighted data inputs with the aid of a different type of computational weight known as a bias;
iii) An activation function which defines whether or how the weighted inputs are transformed into an output.

[1051] Some commentators have observed that an AI model is closer to the concept of a database than a computer program. See, Begoña Gonzalez Otero, 'Machine Learning Models under the copyright microscope: is EU Copyright fit for purpose?' Max Planck Institute for Innovation & Competition Research Paper No 21-10, 23 n.127.
[1052] Optimisation is the process of updating the model attributes (weights, learning rate etc) to minimise inaccurate predictions. Regularisation is the act of constraining a model so it doesn't overfit to the training data. (Géron, op. cit. 920, 27) In the interests of simplicity, the term optimisation will be used to refer to both optimisation and regularisation.
[1053] Not all decision tree models will incorporate weights. See Ali Behrouz et al, 'Fast optimization of weighted sparse decision trees for use in optimal treatment regimes and optimal policy design' (2022) CEUR Workshop Proceedings 3318.
[1054] Simon Haykin, *Neural Networks and Learning Machines* (Pearson 2009) 10.

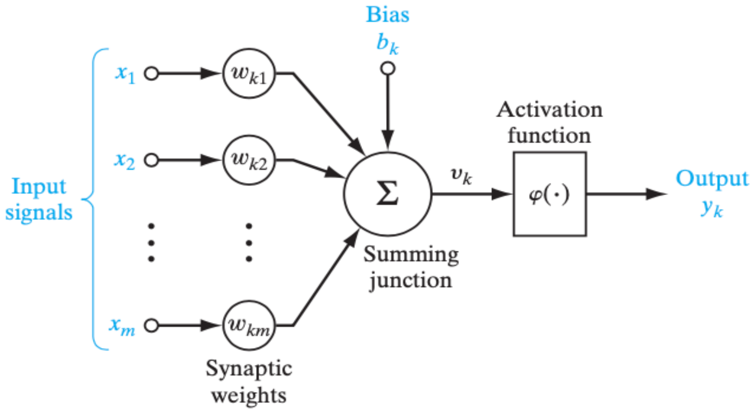

Fig. 5-7. A Typical Node / Neuron in a Neural Network[1055]

Having been assigned an initial value, often owing to well-known heuristics,[1056] weights are then updated automatically using regularisation and optimisation functions[1057] coupled with different training methodologies.[1058] Based on the error rate – the difference between the actual and the desired output – the optimisation functions repeatedly update the parameters accordingly as the software "learns" from the data the optimal settings. Large neural networks may contain billions of weights and, importantly from a protection perspective, they can be extracted and exported for use as the independent starting point for training new models in a process known as "transfer learning".[1059] This is particularly common where training data may be computationally expensive, time-consuming, or where data in the target domain is hard to locate.[1060] Given the high levels of investment that can accrue in their final determination, as the embodiment of what the model has learnt, whether copyright or *sui generis* protection may be afforded parameters is therefore a question of significant jurisprudential import.

7.2 Parameters and copyright protection

One early legal commentator on deep learning sets out that neural network weights could be afforded copyright protection based on two distinct lines of argumentation.[1061] The first being that human involvement is implied in selecting the environment within which parameters are created by the

[1055] Ibid.
[1056] Meenal Narkhede et al, 'A review on weight initialization strategies for neural networks' (2022) 55 Artificial Intelligence Review 1 291.
[1057] According to Google's Machine Learning Glossary regularisation is any mechanism which reduces overfitting. Popular optimisation functions for neural networks include Gradient Descent, Stochastic Gradient Descent, Adam etc.
[1058] For example, the error rate of a model can be updated after each training sample, a batch of training examples, or at the end of the process.
[1059] For example, word embedding models (mathematical representation of the meaning of a word so that words of the same meaning are grouped closer together) such as Word2vec are a particularly common feature of transfer learning in order to speed up the process of creating project specific word embeddings.
[1060] Behnam Neyshabur et al, 'What is being transferred in transfer learning?' (2020) NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems 44 512.
[1061] Donald L Wenskay, 'Intellectual Property Protection for Neural Networks' (1990) 3 Neural Networks 2 229, 235.

software, and the second reliant on the fact that, as a compilation, they constitute a type of copyright protected database.

Moving to answer Wenskay's first point: while parameters are unquestionably calculated as a result of human intervention upstream in selecting the data, defining the model, and managing the optimisation and regularisation process, as a computer-calculated variable they will unquestionably fall beyond the scope of copyright protection. Not only are individual values viz. mathematical calculations not conceptually protected by copyright as a matter of long-held doctrine,[1062] furthermore no free and original choice is exercised on the part of the human programmer in their creation, as they are derived automatically by the optimisation function's response to differing iterations of training data.[1063] Further, even in those common law jurisdictions which, in contravention of the Berne Convention, have introduced computer-generated copyright provisions, as mathematical variables viz. individual datapoints they will almost certainly remain beyond the scope of protection owing to the inherently ineligible nature of their subject matter.

Turning to the second contention, that a structured set of parameters constitutes a copyright protected collection of works, will find a different legal response dependent on the jurisdiction. In order to be considered a compilation under US copyright law, protected works must be compiled from "pre-existing materials"[1064] and arranged in an original and creative manner.[1065] Similarly, copyright in a database in the EU inheres in the original "selection or arrangement" of the elements of the database, excluding protection for the contents themselves.[1066] Although it could be argued the support vectors in an SVM model are pre-existing, this is not the case for weights in an ANN, decision tree, random forest, or the co-efficients in logistic and linear regression, as they are calculated and updated as part of the training phase. While the dynamic creation of parameters in most models appears to align with

[1062] See Bernt Hugenholtz, 'Data property: Unwelcome Guest in the House of IP' in *Better Regulation for Copyright: Academics meet Policy Makers* (The Greens, European Parliament 2017) 65, 69 (explaining that a datapoint is not protected by copyright law as it falls the idea side of the idea-expression dichotomy.) Josef Drexl, 'Designing Competitive Markets for Industrial Data - Between Propertisation and Access' (2016) Max Planck Institute for Innovation & Competition Research Paper No. 16-13, §45 (referencing Article 3.2 of the Database Directive's implicit exclusion of data from copyright protection.) See also *SAS v WPL* 66 ("figures or mathematical concepts which, considered in isolation, are not, as such, an intellectual creation of the author".) cf. see supra 966.
[1063] Concurring with the view that weights are non-protectable, see Patrick Ehinger and Oliver Stiemerling, 'Die urheberrechtliche Schutzfähigkeit von Künstlicher Intelligenz am Beispiel von Neuronalen Netzen' (2018) 34 Computer und Recht 12 761, §57.
[1064] §101 US Copyright Act: "A 'compilation' is a work formed by the collection and assembling of preexisting materials or of data that are selected, coordinated, or arranged in such a way that the resulting work as a whole constitutes an original work of authorship."
[1065] *Feist* 358.
[1066] See Article 3.1 of the Database Directive: "In accordance with this Directive, databases which, by reason of the selection or arrangement of their contents, constitute the author's own intellectual creation shall be protected as such by copyright. *No other criteria shall be applied to determine their eligibility for that protection*." (Emphasis added). Pursuant to *Football Dataco* the CJEU has ruled that the contents of the database are irrelevant to their protection pursuant to the Directive because "the concepts of 'selection' and of 'arrangement' within the meaning of Article 3.1 of Directive 96/9 refer respectively to the selection and the arrangement of data, through which the author of the database gives the database its structure. *By contrast, those concepts do not extend to the creation of the data contained in that database.*" (Emphasis added).

the definition of a database compilation in Union law, given the pre-existing materials requirement of §101 of the American Copyright Act, whether the same can be said in the US is less certain.[1067]

Moving beyond definitional concerns, as required by copyright law in both Europe and the US, as a rule the structure and organisation of the parameters in the model are not the result of human choice.[1068] Instead, they are automatically saved by the computer software in a file that ensures the data remains interoperable within its originating programming environment.[1069] As a compilation of variables, their structural selection and arrangement as a rule is determined by the AI framework in conjunction with the operating system. Lacking any human intervention, they therefore fail to reach the protection threshold required by copyright law.[1070] Even in instances where a programmer is involved in the selection and arrangement of the parameters, as explored in the previous chapter, as purely functional elements of the software, even weights in aggregate would appear to be rendered unprotected by copyright.[1071]

7.3 Sui Generis Database Rights: AI models and parameters

Having established that parameters, both at an individual level or as a compilation, likely fail to meet the requisite criteria for copyright protection in both Europe and the US, we shall now turn to the specifics of the European *sui generis* database regime. Independent of the operativity of copyright, as set out in Article 7(1) of the Database Directive, the maker of the database is protected from the unauthorised "extraction and or re-utilisation of the whole or of a substantial part, evaluated qualitatively and or quantitively of the contents of that database."[1072]

This sub-section will proceed along two paths. The first part examines whether an AI model as a whole, being a collection of sub-programs and parametric values, may enjoy *sui generis* database protection. After this, the *sui generis* database status of parameters as independent elements within the software will be evaluated to understand whether they would fall within the reach of the Directive.

[1067] See *The House Report on the Copyright Act of 1976* (House Judiciary Committee 1976) stating: "A "compilation" results from a process of selecting, bringing together, organizing, and arranging *previously existing material* of all kinds, regardless of whether the individual items in the material have been or ever could have been subject to copyright." (Emphasis added).
[1068] Database Directive. Article 3.1 " 'database' shall mean a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means."
[1069] Von Lewinski & Walter supra 316, §9.3.6 (stating that the application of the "arrangement" criterion to electronic databases "would not be sensible" as it is determined by the operating system). cf. Whilst in theory a programmer in theory may write their own dictionary to store the parameters in a unique and novel manner external to the model, given the commonality of dictionaries for storing objects (e.g. state_dict in PyTorch) this would seem a rare occurrence.
[1070] It could be argued therefore in the common law countries that have introduced a computer-generated copyright works provision that compilations of parameters by a computer may fall under such rules. However, any such suggestion is *reductio ad absurdum* as it would afford copyright protection to any and all public domain information digitised by simple fact of it having been saved by a computer and arranged within its memory.
[1071] See Peter Henderson and Mark Lemley, 'The Mirage of Artificial Intelligence Terms of Use Restrictions' (2024) pre-print arXiv:2412.07066, 29 (asserting that in the US weights are likely to fall under §102(b) owing to their functional nature.)
[1072] Article 7.1, Database Directive.

As to the first question, a teleological interpretation of the Directive warrants a response in the negative. As is discussed below, while in theory a technical rendering of the law may afford *sui generis* database protection to AI models, the intention of the Commission in passing the Directive nevertheless suggests this was not its aim. A careful reading of the Proposal reveals that the Software Directive and the Database Directive formed part of an intentionally two-fold complimentary legislative package.[1073] Thus, it can be deduced that, by implication, the Commission's aim in passing both Directives was to protect software as logically and operatively separate from the contents of a database.

Moreover, the policy purpose undergirding the introduction of the Directive emanates from an economic desire to support the creation of databases and enrich European information markets. As the Proposal makes clear "the availability of up-to date [sic] comprehensive sources of information, and the ability to store and manipulate large quantities of data are key factors in today's competitive business environment."[1074] An AI model however fulfils none of the genuine criteria for a database. Assuming, as required by Article 10.1, a ML model in its completed form can be recognisably determined,[1075] self-evidently it is not a compilation of reliable and accurate data that has been "collected and catalogued"[1076] for reference and analysis purposes; it is a functional machine whose purpose is to make predictions and classifications on data. Moreover, contrary to the value of a database and the grounds for its protection by the Directive, when it is deployed commercially – even if licensees have permission to utilise the software itself – its underlying code is by design frequently shielded from "extraction and/or re-utilisation" by firewalls and other technological protection measures.[1077] Where access is *by design* off limits, it is therefore from a teleological perspective clearly antithetical to the policy goal of the Directive. The Directive's stated aim is to stimulate reference markets while protecting the investment made in a database from the excessive extraction and reutilisation of its contents.[1078] Thus, where by intention such activities are not possible, any claims arising for protection under the Directive should fail.[1079]

[1073] §5.3.3. Proposal for a Council Directive on the Legal Protection of Databases. Com (92) Final – SYN 393. 1.2 ("The choice of copyright protection will complete a coherent package taken in conjunction with the Council Directive on the legal protection of computer programs (91/250/EEC) since the contents of the database and the program which stores and manages the materials are difficult to separate.")
[1074] Id. §1.2.
[1075] Article 10.1, Database Directive. "The right provided for in Article 7 shall run from the date of completion of the making of the database." Protection is only triggered therefore once the database has been completed, and therefore the model in the training phase is excluded. Ascertaining the point at which *sui generis* protection may arise given the dynamic nature of many AI models is unclear. See supra 1043 showing that some trained models are continually being updated even after deployment.
[1076] Supra 1073. §1.3.
[1077] Article 7.1 Database Directive.
[1078] Id. See Recitals 9-12.
[1079] Note Jean-Marc Deltorn, 'The elusive intellectual property protection of trained machine learning models: a European perspective' in Ryan Abbott (ed) *Research Handbook on Intellectual Property and Artificial Intelligence* (Elgar 2022) (although by implication based on differing reasoning (i.e. AI models created from single-source databases), Deltorn states "The *sui generis* right thus only applies to databases that contain data that is publicly accessible.") 84.

Moving from purposive policy interpretations to a legal analysis of the text of the Directive, whether *sui generis* protection extends to AI software and the model parameters therein is somewhat less clear. Although some scholarly commentators do not discount that *sui generis* database protection may inhere in a computer program, the qualifying criteria as outlined by the Directive certainly leaves such an assertion open to question.[1080]

One qualifying criterion for protection under the *sui generis* database regime is that of informational independence. Pursuant to established case law,[1081] Union law requires that the "independent work" criterion of Article 1(2) is interpreted to mean that a database must comprise of "independent materials" and that the "material retains *autonomous* informative value" once extracted from the database.[1082] Thus, even when the works are accessible independently and severally, they must retain the same informative value as *in situ*. Quaedvlieg argues that all technical and functional subject matter requires *inter*dependency of its components in order to operate, and as a consequence software fails to demonstrate *in*dependence as required by the Directive. Accordingly, software will therefore be excluded from protection as a *sui generis* database.[1083] Westkamp and Derclaye similarly highlight the *inter*dependent nature of software sub-programs, pointing to the fact that if one element of the software is removed, then the program will fail to operate correctly.[1084] As a result, Westkamp concludes that as dependent elements, contrary to the requirements of Article 1(2) of the Directive, the "purpose to individually inform is absent."[1085]

It should nevertheless be recalled that while this may be true of conventional software, in the case of deep learning models, a model may still work with some minor reduction in connectivity between neurons and layers.[1086] All the same, the feedforward nature of most neural networks, and the bi-directional nature of recurrent neural networks, serves to illustrate the interdependency of the sub-programs / layers of the model as input data is designed to flow across the whole software model, and like traditional software any qualitatively substantive removal of elements of the code will either cause it to fail to operate, or output a poor prediction.[1087]

[1080] Id. 96. Estelle Derclaye, *The Legal Protection of Databases: A Comparative Analysis* (Elgar 2008) 67 ff. Drexl, op. cit. 985, §18 (without elucidation, stating that "Simple/linear ML models do not qualify for database *sui generis* protection. However, complex, dynamic ML models may need to be differently assessed.")
[1081] *Football Dataco,* 26. Case C-444/02 *Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE* [2004] ECR I-10549 33 (hereinafter *OPAP*).
[1082] Case C-490/14 *Freistaat Bayern v Verlag Erstbauer GmbH* [2015] ECLI:EU:C:2015:735 23 (Emphasis added and quotation marks removed).
[1083] AA Quaedvlieg, 'Onafhankelijk, geordend en toegankelijk: het object van het databankenrecht in de richtlijn' (2000) 9 Informatierecht / AMI 177, 182.
[1084] Derclaye, supra 1080, 67. Guido Westkamp, 'EU Database protection for information uses under an intellectual property scheme: has the time arrived for a flexible assessment of the European Database Directive?' (11th Fordham Intellectual Property and Policy Conference, 2003).
[1085] Westkamp, supra 1084, 3.
[1086] Andy Johnson-Laird, 'Neural networks: The next intellectual property nightmare?' (1990) 7 Comp Lawyer 3 7, 13.
[1087] cf. Drexl, supra 1080, ibid. (with no clear elaboration, stating why more complex models may qualify for protection stating "[s]imple/linear ML models do not qualify for database sui generis protection. However, complex, dynamic ML models may need to be differently assessed.")

Applying the Directive's independent work criterion specifically to the question of parameters, the consensus in the literature is that they would fail to meet the autonomous informative value requirement, whereby as Leistner states, every element of a database should "hold the same self-contained informational content when accessed separately as each would have when viewed together".[1088] This position is based on the fact that an individual parameter is meaningless, as it is only in context when interlinked to the software and the other weights and variables that their information value is realised.[1089] To underline this, as Drexl observes, parameters in aggregate removed from the model renders the software essentially useless, incapable of making accurate predictions. Thus, their informational value is only concretised *in situ*.[1090]

A contrario, Deltorn suggests that the support vectors of one type of SVM model may be protected.[1091] Envisaging those SVM models that retain their support vectors when saved,[1092] Deltorn posits that as elements not calculated but extracted and isolated from the training data, they may qualify as having independent informational value.[1093] Nevertheless, this position is less than persuasive for the reasons stated above. Namely, that in isolation an individual support vector is no more than a single value in n-dimensional space and it is only in combination with others that a support vector's meaning and informational value – in the form of the model's hyperplane boundary – becomes apparent.

Moving to the object of the investment protected by the *sui generis* regime "either the obtaining, verification or presentation of the contents" of the database must be implicated.[1094] In order to qualify for protection, these qualifying acts must however be distinct and separate from the non-qualifying investment made in the creation of the works contained within the database.[1095] As established by Article 7.1 of the Directive, the investment in these three activities must be the result of "qualitatively and/or quantitatively a substantial investment." While the CJEU has yet to decide on the issue, given that AI frameworks provide in one place numerous pre-written programs to choose from, it seems improbable that investment in importing the software elements of the model into the local working environment would reach a threshold deemed sufficiently substantial to qualify for protection.[1096] This

[1088] M Leistner, 'The legal protection of telephone directories relating to the new database maker's right' (2000) 31 IIC-International Review of Industrial Property and Copyright Law 950, 956. See also Deltorn, supra 1079, 98. Drexl, supra 1080, ibid. Ehinger and Stiemerling, supra 1063, §60-61.
[1089] Ehinger and Stiemerling, ibid.
[1090] Drexl, supra 1080, ibid.
[1091] Deltorn, supra 1079, 98.
[1092] cf. Rather than save the support vectors of a trained model a data scientist may instead rely on an on-the-fly calculation of the hyperplane based on the information derived from the training phase.
[1093] Note, pursuant to *Case* C-338/02 *Fixtures Marketing Ltd v Svenska Spel AB* [2004] ECR I-10497 24 (hereinafter *Svenska Spel*); Case C-203/02 *British Horse Racing Board and Others v William Hill Organization Ltd* [2004] ECR I-10415 31(hereinafter *BHB v WH*); Case C-46/02 *Fixtures Marketing Ltd v Oy Veikkaus Ab* [2004] ECR I-10396 34 (hereinafter *FM v Veikkaus*); *OPAP* 40, Article 7.1 of the Directive that the *sui generis* right does not protect the creation of the contents of the database.
[1094] Article 7.1.
[1095] *BHB v WH*, 31. *Svenska Spel*, 29. *FM v Veikkaus,* 34. *OPAP,* 45.
[1096] Article 7.1 requires that "there has been qualitatively and / or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents" of a database.

is for the reason that the act of obtaining sub-programs from the toolkit would require no more than writing a line or two of code. (see Figs. 3 & 4) However, unless a spin-off derived from a primary activity of the developer,[1097] where functions are sourced elsewhere (providing their acquisition can be classified as obtaining and not creation), if a qualitatively and / or quantitatively substantial investment has occurred, the model will reach the requisite threshold for protection.[1098]

In terms of the qualifying act of presentation, which implicates the structuring and display of the contents,[1099] the presentational display of the ML interface will normally be determined by the framework or other pre-written software, and therefore in such cases clearly involves no direct investment in the presentation of the contents of the model. With regards to the structuring and verification of the model (i.e. ensuring the accuracy of the hyperparameter settings and parameters[1100]), both activities appear principally related to the act of model creation which is disqualified from protection.[1101] From the standpoint of the Directive, parameters and sub-programs are not arranged, monitored, enhanced, added to, or replaced to ensure marketability but in order for the model to operate. Their presence and arrangement in the work are integral to the program's functioning and therefore appear best delimited as creational acts. Moreover, it should be recalled that the task and data-specific architectural hyperparameters of the software are "verified" during the training phase (or in the case of some SVM model parameters selected from training data) at the same time as the model is being created, thus clearly indicating that they are creational actions which are excluded from the scope of the *sui generis* right.[1102]

In sum, AI models are therefore unlikely to qualify for *sui generis* database protection owing to a combination of different factors. Parameters are created by training the model on data and, lacking informational autonomy, they fall beyond the scope of the Directive. Likewise, in terms of the software, the substantial investment in building the model resides not in the obtaining, verification, or presentation of the sub-programs but in their fine-tuning which is, without question, a creative act required so that the model functions. The process of optimisation involves creating a software model, not with a view to verifying its informational contents being made available as a referenceable database, but as required so that the model can make accurate predictions. This is a non-protected act of fabrication where something new is arranged and structured according the functional demands of

[1097] Pursuant to *BHB v WH, OPAP* etc, databases which are by-products of a principal activity rarely attract *sui generis* protection as the investment in obtaining and verification of the data pertains to creation during the initial prime activity and not the creation of the database.
[1098] Derclaye, supra 1080, 75 ff. (listing some of the European cases where courts have ruled on substantiality in the context of the *sui generis* database right.)
[1099] Walter and Von Lewinski, supra 316 §9.7.11.
[1100] A C Beunen, *Protection for Databases: The European Database Directive and its Effects in the Netherlands, France and the United Kingdom* (Wolf Legal Publishers 2007) 134.
[1101] Supra 1095.
[1102] Derclaye, supra 1080, 197 (observing that pursuant to *BHB v WH,* and *Svenska Spel*, where the acts of obtaining, verifying and presenting occur at the same time as the creation of the database the contents are excluded from protection.)

the software model and therefore rightly should fall beyond the scope of protection of the *sui generis* right.

8. Do AI models require consolidated protection by copyright law?

Given the limited instances in which copyright and *sui generis* database regimes afford protection to AI software, whether further consolidation of copyright is needed to safeguard the economic interests of investors and creators is a normative question of global significance.[1103] This final subpart will evaluate whether further copyright or para-copyright legislative protections are required in order safeguard industrial investment in this vitally important sector of the digital economy.

As justification for further regulatory regulation is frequently grounded in commercial imperatives, the starting point for any normative discussion should be economic. Considering the huge investment in AI that has taken place over the last 10-15 years, it appears that a lack of copyright protection has certainly not hampered growth, with private investment in AI reaching $252.3 billion in 2024 alone.[1104] Thus, by implication, extra-copyright legal regimes that exist to address market harms appear to be functioning adequately.

Extra-copyright protection for AI investors *inter alia* include trade secrets, patents, licences, unfair competition law, computer misuse and trespass legislation, conduct-based torts, as well as the protection of know-how through confidentiality clauses in employee and partnership contracts.[1105] In addition to legal interventions, numerous technological protection mechanisms may be called upon to deter free riding.[1106]

As Drexl highlights, the existence of manifold regimes designed to create artificial exclusivity raises the question of whether further regulatory intervention could be detrimental to public welfare.[1107] By implication, a plethora of IP and non-IP forms of protection should serve to make policymakers aware of the adverse impacts and deadweight loss that may arise from excessive monopoly control over

[1103] For discussion of patent and / or trade secret protection of AI software, inter alia see Deltorn, op. cit. 1079, Drexl, op. cit. 985. Ehinger & Stiemerling supra 1063. Otero supra 1051. Lee et al supra 986. Jasper Siems, 'Protecting Deep Learning: Could the New EU-Trade Secrets Directive Be an Option for the Legal Protection of Artificial Neural Networks?' in Woodrow Barfield and Ugo Pagallo (eds) *Algorithmic Governance and Governance of Algorithms: Legal and Ethical Challenges* (Springer 2021) 137. Woodrow Barfield (ed), *Research Handbook on the Law of Artificial Intelligence* (Edward Elgar Publishing 2018) Part IV.
[1104] *Artificial Intelligence Index Report 2025* (Stanford University 2025) 218.
[1105] Jacqueline Lipton, 'IP's Problem Child: Shifting the Paradigms for Software Protection' (2006) 58 Hastings Law Journal 205 (arguing that copyright software law should be scaled back as a combination of contract, TPMs and trade secret law provide sufficient protection.) A number of leading US academics have from the start opposed copyright protection for software: Pamela Samuelson, 'CONTU Revisited: The Case Against Copyright for Computer Programs in Machine Readable Form' (1984) Duke Law Journal 663. Pamela Samuelson et al, supra 761.
[1106] In addition to TPMs, AI source code is frequently inaccessible to competitors and innovators because it sits in the cloud "as a service". Acronyms for such services include SaaS (software as a service), (IaaS) (infrastructure as a service) and (PaaS) (platform as a service). e.g. Amazon Web Services, IBM Watson and Microsoft's Azure AI etc.
[1107] Drexl, op. cit. 985 §16 (underlining that over-protection of AI could create negative welfare effects.)

access to algorithms and the ideas which inhere in AI software. In a world dominated by superplatforms and powerful IoT market incumbents, any normative actions should be sensitive not only to the needs of private investors but the public interest that dwells in competition, innovation, and vibrant technology markets.

Not only legal but technological barriers may also hamper access to ideas and certain categories of expression which, for public policy reasons, copyright otherwise renders unprotected. The ability for locally hosted AI programs, APIs, and the cloud to work in unison means that vendors now have the technological means to prevent lawful access to code in ways unimaginable forty years ago when software was first conceived as a literary right under copyright law.[1108] Backed up by mandatory anti-circumvention provisions in international treaties[1109] and other computer trespass laws, private ordering through the application of technical protection measures often creates an inviolable multi-layered defence for technology companies. Furthermore, unlike any inconstant legal intervention, it is one that investors are able to control themselves end-to-end.

Taking the case of AI software in driverless cars as one example of the current application of TPMs, as a matter of course vehicles are fitted with encrypted neural network chips.[1110] The requirement for fast response times means that currently much of the time-critical data processing takes place locally in the vehicles. Encryption is required for safety reasons to ensure that neither the software nor the signals being sent and received by the vehicle (e.g. wireless data exchanges with the vehicle manufacturer etc) are subject to interference from bad actors. Furthermore, vehicle-to-cloud connectivity is becoming increasingly sophisticated,[1111] with experts predicting that by the late 2020s some of the time-sensitive core functions performed by driverless vehicles will be controlled from cloud-based architectures.[1112]

As we see in this example, from a data integrity perspective both consumers and corporations unquestionably benefit from the application of TPMs. All the same, it cannot be denied that the application of a "triple lock" of technological protectionist measures will also create negative welfare effects as they indiscriminately prevent not just unauthorised access but, from a copyright perspective, entirely legitimate and lawful activities also.

[1108] See Lothar Determann & David Nimmer, 'Software Copyright in the Cloud' in Christopher Yoo et al (eds), *Regulating the Cloud: Policy for Computing Infrastructure* (MIT Press 2015) 235 (outlining how the cloud is rewriting copyright norms and frustrating user privileges.)
[1109] Article 11 WCT. Article 18 WPPT.
[1110] 'Under The Hood of Tesla's AI Chip That Takes the Driverless Battle to NVIDIA's Home Turf' (*Analytics India Magazine*, 26 April 2019) https://analyticsindiamag.com/under-the-hood-of-teslas-ai-chip-that-takes-the-driverless-battle-to-nvidias-home-turf/
[1111] Edgar Cortés Gallardo Medina et al, 'Object detection, distributed cloud computing and parallelization techniques for autonomous driving systems'(2021) 11 Applied Sciences 7 2925.
[1112] 'Top 4 Car Features That Rely on Vehicle-to-Cloud Connectivity' (*Autocrypt,* 16 February 2022). https://autocrypt.io/4-car-features-rely-on-vehicle-to-cloud-connectivity/

The first tier of the triple technological lock that face legitimate actors wishing to perform statutorily authorised activities is the concealment of unprotected expression and function in the machine code. A second barrier to access is the widespread use of TPMs which include both copy-control and access-control mechanisms.[1113] The third layer of protection that has come to the fore as network latency has reduced is the remote, cloud-based provision of software technologies. e.g. SaaS, PaaS, AIaaS, etc.[1114] Accordingly, the underlying software is frequently inaccessible and off-limits to licensees, as the program is utilised remotely on the providers' infrastructure via an API. The cumulative effect of these multiple layers of protection is that innovators are effectively excluded from accessing the unprotected ideas and processes embodied in software, notwithstanding that such access is expressly permitted by the legislator for reasons of innovation policy.[1115]

Owing to the multiple forms of legal and extra-legal protection measures that exist, the regulatory question arises as to whether further protection for software in the form of copyright is required at all.[1116] Whereas Ginsburg argued in 1994 it was too soon to retreat from copyright protection for computer programs,[1117] in the intervening decades the ubiquity of open-source software, widespread reliance on private ordering, combined with the application of ever more sophisticated technological security measures indicate that copyright protection not just for AI but for any form of software may now be superfluous. As shown, corporations wishing to protect their investments have at their disposal a wide array of effective extra-copyright legal and technical mechanisms providing ample safeguard.[1118]

Indeed, it can be envisaged that the outright removal of copyright protection for software, may bring with it tangible benefits. Firstly, the courts would no longer be distracted by the complex and time-consuming doctrinal challenges that software protection still continues to present.[1119] Furthermore, as Lipton argues, from the standpoint of jurisprudence the interplay between licences and copyright law often muddies the juridical waters. This leads to higher costs and "decreases the efficiency of the legal process… and potentially disadvantages litigants who may not have the financial means to produce expert witnesses to support their claims and defenses."[1120] As frequently highlighted in the literature, the over-protection by not just copyright but any form of IP creates negative externalities as

[1113] See Chapters 2 and 3 for discussion on TPMs in the relevant jurisdictions.
[1114] Supra 1106.
[1115] Despite express provisions in the Software Directive (Article 5.1) and the US Copyright Act (§117 (c)) authorising error correction, copyright and other IP rights are being used to frustrate consumers and businesses from repairing software embedded machinery. See Aaron Perzanowski, *The Right to Repair* (CUP 2022) 110 ff.
[1116] Lipton, supra 1105 (stating that "large-scale copyright protection of code is simply not necessary anymore.") 247. See also *Economic Impact of IP on Investment in AI* (Intellectual Property Office Research Paper 2022) (it is interesting to observe that amongst SMEs involved in AI, amongst differing forms of IP, patents had the highest perceived value and understanding.) 5.
[1117] Jane Ginsburg, 'Four Reasons and a Paradox: The Manifest Superiority of Copyright over Sui Generis Protection of Computer Software' (1994) 94 Colum L Rev 2559, 2560.
[1118] See supra 953.
[1119] See Lemley and Samuelson, supra 735 (discussing the Supreme Court's avoidance of engaging with copyright software doctrines by simply treating Google's actions as hypothetically fair use).
[1120] Lipton, supra 1105, 248.

downstream technological advancement and innovation can be stymied.[1121] Additionally, in removing copyright from the legal equation entirely, benefits may accrue as companies focus on more certain protection strategies than copyright itself can provide.[1122]

Nevertheless, it may also be reasoned that there are sound public policy reasons why copyright law should not retreat from the regulation of AI software but instead refocus. As Slowinski argues, copyright protection for AI software may be "desirable not so much for the protection but for the exceptions and limitations included in copyright law."[1123] While to date in public fora there has been comparatively little policy discourse concerning the weak protection of AI software by copyright law,[1124] nonetheless it has induced a palpable shift in the international politics of digital trade as states seek to defend the market incumbent position of their own home-grown AI champions.[1125] As will be discussed in depth in the next chapter, notably the governments of Japan and the United States have been working strategically via Free Trade Agreements and the WTO to foreclose access not just to copyright protectable expression but to the public domain ideas and methods of operation which are encoded into software algorithms. From the viewpoint of copyright law, by seeking to frustrate competition through preventing access to ideas, these two export-oriented AI superstates are incontrovertibly seeking to cement their hegemonic position in highly lucrative sectors of the technology market.[1126] It is in this context that bodies of laws such as copyright can, if well-crafted, facilitate access to important information crucial for competition and innovation, thus serving an important role in nurturing economic and technological development globally.

Conclusion

As shown, AI models are afforded weak protection by copyright law, and in Europe, also unlikely to qualify for *sui generis* database right protection. As discussed in the next chapter, the absence or at best uneven protection afforded investors who have expended resource in training AI models has created a spill-over effect in international trade law. Through targetted copyright interventions, in tandem with other innovation-oriented legal and extra-legal measures, it is submitted that rethinking copyright as it applies to machine learning can bring positive welfare effects not just for net AI exporter nations, but for countries at all stages of development.

[1121] Perzanowski, supra 1115. See also Drexl, op. cit. 985 §16.
[1122] Drexl (discussing non-formal IP means of protecting AI investment such as conduct-based tort, unfair competition laws, contracts and TPMs.) ibid.
[1123] Peter Slowinski, 'Rethinking Software Protection' 361, in Lee et al, supra 986.
[1124] It is noteworthy that despite as discussed herein the most valuable source code elements of AI models being unlikely to attract copyright protection, in its 2020 consultation 'Artificial Intelligence Call for Views: Copyright and Related Rights' the UK IPO stated "We are unaware of specific problems with protection for AI software. Copyright appears to provide adequate protection for software, allowing its creators to benefit financially from their work."
[1125] See Maurizio Borghi and Benjamin White, 'Data extractivism and public access to algorithms' in Maurizio Borghi & Roger Brownsword (eds), *Law, Regulation and Governance in the Information Society: Informational Rights and Informational Wrongs* (Routledge 2022).
[1126] Most obviously e-commerce, robotics, driverless vehicle and the Internet of Things.

Chapter 6: Artificial Intelligence and Trade Law: How Exceptions to Copyright Can Ameliorate AI Geocentrism

Introduction

Considering the origins of copyright in the book trade over 300 years ago, it is perhaps surprising to many that it has evolved into the important regulator of data-driven innovation and technology markets that it is today. Copyright is a near ever-present feature of the AI development and innovation cycle. From the creation of the AI models themselves, the training data, to the object of the data analysis, the contours of copyright will frequently affect both incentives to invest in AI and market competition.

It is evident that in these early days of AI adoption, a handful of companies, primarily located in North America and East Asia, wield excessive market power[1127] in areas of the economy as diverse as e-commerce, search, social media, and autonomous vehicles.[1128] A new era dominated by AI "super monopolies" that concentrate rent and disproportionate influence in the hands of a select band of actors and states demands new thinking and new rules. Owing primarily to TRIPS, the near global reach of copyright law recommends it as one potentially effective avenue for market intervention by the authorities. This may serve to not only incentivise local technology companies, but also mitigate the acute societal challenges posed by the excessive market disequilibrium that has emerged.

Building on the preceding chapters of the thesis, this final section will proceed as follows: Part I will outline the industrial strategy of export-oriented AI superstates in the form of a new source code discipline which has become a notable feature of Preferential Trade Agreements (PTA) and e-commerce negotiations at the WTO since 2017.[1129] After this, how these trade law developments rewrite the norms of copyright will be briefly discussed. Part II will detail how derogations to copyright law may be developed at a domestic and supra-national level to create a right of access to ML model code, thus, harnessing intellectual property law to support a more efficient distribution of AI technological advances in the interests of societal welfare and equitable development.

[1127] The term "excessive market power" is used to refer to a situation where a company or a group of companies have established their policies such as price, product, linkages, market position, portfolios etc to a significant extent absent severe competitive constraints. The term excessive is used to denote that market power has become problematic from an economic perspective and potentially from a societal perspective also, including instances where it can act as a lever for political influence. See *The Evolving Concept of Market Power in the Digital Economy* (OECD 2022). On the political dimensions of market power see Lina Khan and Sandeep Vaheesan, 'Market Power and Inequality: The Antitrust Counterrevolution and its Discontents' (2017) 11 Harvard Law & Policy Review 235.

[1128] In North America and Europe technology markets are dominated by GAFAM and in China, Baidu, Alibaba and Tencent (BAT). Driverless vehicle markets tend to be dominated by US, China and Japan, and robotics exports are led by Japanese firms such as Mitsubishi, Denso, Kawasaki, Fanuc etc. (See Chapter 1 n.172).

[1129] The term e-commerce is used to cover all types of technology dependent activities including search, social media platforms, digital platforms, hardware and software etc.

This chapter approaches the issue of geocentricism in technology markets, with a particular emphasis on the effect that AI legal developments can have on second-comer states and the large number of developing countries in the Global South. Not only do these countries represent the majority of the global population, but many are characterised by weak institutional capacity and low levels of investment in technological infrastructure and R&D. While arguably less important for more developed late-comer nations such as those in Europe, given the internal structural challenges faced by developing and least-developed nations, an international approach that facilitates knowledge transfer via AI-related limitations and exceptions is more certain to foster positive development outcomes than if these nations were left to navigate these challenges independently.

Section One

1. The Genesis of Source Code and Algorithm Disciplines in Digital Trade Law

Precipitated by the growing importance of digital trade in the Global North, in September 1998 the General Council of the WTO established the "Work Programme on e-Commerce" with a stated aim of exploring the relationship between existing international treaties and digital trade.[1130] With no mandate to draft treaty language, institutional deadlock around the Doha Agenda, and concerned with clandestine technology transfer, it was not long before Japan and the US began to take digital trade matters into their own hands through bilateral and plurilateral trade agreements.[1131]

The inception into preferential trade agreements (PTAs) of electronic commerce terms can be traced back to as early as the year 2000.[1132] By 2023, at least 433 trade agreements can be identified as containing one or more digital trade provision.[1133] One particularly notable milestone was the 2016 Japan-Mongolia Economic Partnership Agreement, which marked the first time a PTA expressly incorporated a source code provision.[1134] With the exception of software required for critical infrastructure, the Agreement prohibits access to forms of "mass-market software" as "a condition of the import, distribution, sale, or use of such software or of products containing such software."[1135] Intertwined with these prohibitions, Article 9.10 additionally sets out that there should be no restrictions placed on the physical location of the parties' computing facilities. Thus, outlawing data localisation requirements which would otherwise prevent the export from Mongolia to Japan of

[1130] https://docs.wto.org/dol2fe/Pages/SS/directdoc.aspx?filename=q:/WT/L/274.pdf&Open=True
[1131] Hosuk Lee-Makiyama, *Briefing Note: AI and Trade Policy* (2018) 6 ff. (referencing the procurement laws linked to market access of India, Colombia, Indonesia and China which allow these "governments to misappropriate commercial algorithms and source codes".)
[1132] Article 7. US-Jordan Free Trade Agreement (24/10/2000).
[1133] Source: TAPED database. University of Lucerne. https://www.unilu.ch/en/faculties/faculty-of-law/professorships/burri-mira/research/taped/
[1134] Agreement between Japan and Mongolia for an Economic Partnership (7/6/2016) Article 9.11.
[1135] Id. Article 9.11.1.

personal and non-personal data for extraction, computational analysis, and data commodification purposes.

Since the Japan-Mongolia PTA, both data localisation and source code provisions have become more common in international trade agreements[1136] – most powerfully in two mega-regional agreements: the 2018 Comprehensive and Progressive Agreement for Trans-Pacific Partnership (CPTPP) and the 2020 United States-Mexico-Canada Agreement (USMCA).[1137] The USMCA is particularly noteworthy for introducing, for the first time in trade law, provisions that restrict access not only to software but also to the underlying algorithms. According to Article 19.16.1: "No Party shall require the transfer of, or access to, a source code of software owned by a person of another Party, *or to an algorithm expressed in that source code*, as a condition for the import, distribution, sale or use of that software, or of products containing that software, in its territory."[1138] Derogations to the provision include judicial proceedings and making them available to a regulatory body for a "specific investigation, inspection, examination, enforcement action … subject to safeguards against unauthorized disclosure."[1139]

Although less common than the source code discipline, specific algorithm provisions may further be found in the United States and Japan Digital Trade Agreement (2020),[1140] UK-Japan Comprehensive Economic Partnership Agreement (2021),[1141] as well as the UK-Singapore Digital Economy Agreement (2022).[1142]

Beyond PTAs, the future landscape of digital trade is likely to be significantly influenced by the inclusion of both the source code and algorithm discipline in the Joint Statement Initiative,[1143] a plurilateral initiative launched in 2017 at the WTO, spearheaded by the United States[1144] and Japan.[1145] Having grown from initially 71 participants to 91 countries, as of 2024, while not part of formal World Trade Organisation proceedings, negotiations on the JSI involve the majority of Member States and are therefore highly likely to have some impact on future developments at the

[1136] For example, Free Trade Agreement between the European Union and New Zealand 2024/866 (25/3/24) Article 12.11. Agreement between the European Union and Japan for an Economic Partnership (27/12/2018) Article 8.73. Agreement on Digital Trade between the European Union and the Republic of Singapore 2025/0009 (8/4/25) Article 11.
[1137] Respectively, prohibitions on access to source code: Articles 14.17 and 19.6; and data localisation: Articles 14.11-13 and Articles 19.11-12.
[1138] (Emphasis added).
[1139] Article 19.16.2.
[1140] Article 17.1.
[1141] Article 8.73.2 (a).
[1142] Article 8.61-k.
[1143] See leaked JSI text (2020) 48. The source code and algorithm discipline provision was jointly proposed by Japan, US, UK, Canada, Mexico, Ukraine, Peru and Taiwan. https://www.bilaterals.org/IMG/pdf/wto_plurilateral_ecommerce_draft_consolidated_text.pdf
[1144] As of October 2023, while according to one American WTO trade envoy the US has "not to reverse a position … not to change [its] position" it has nonetheless withdrawn its attribution on source code and data localisation provisions in the JSI draft agreement under negotiation. See: https://www.youtube.com/watch?v=xU62epV3LW8.
[1145] Jane Kelsey, 'How a TPP-Style E-commerce Outcome in the WTO would Endanger the Development Dimension of the GATS Acquis (and Potentially the WTO)' (2018) 21 Journal of International Economic Law 2, 273 (outlining the e-commerce focussed strategy of the US, Japan and their supporters at the 11th Ministerial Conference in Buenos Aires.)

WTO. Seeking to reach agreement on issues as diverse as electronic commerce, trust in digital trade, telecommunications, and digital market access, arguably a large part of the negotiations lack immediate economic resonance for those in the developing world.[1146] As one trade envoy has put it, these new digital rules reflect less the needs of the Global South than being designed to "further strengthen the e-commerce behemoths like Google, Facebook, Amazon, Microsoft, Apple, Alibaba, and Baidu… to monetise data on an unprecedented scale."[1147]

2. PTAs, the JSI and AI Superstates: Rewriting the Norms of Copyright Law?

Irion and Makiyama-Lee attribute the emergence of specific source code and algorithm disciplines to a concern that information revealed in the national certification process tied to market entry is being disseminated *sub rosa* to domestic competitors and state-run enterprises.[1148] The general absence of copyright protection for many AI models may therefore constitute a plausible factor influencing the development of such provisions. [1149] This is because *ex post* action in the courts against unauthorised recipients of the software *a priori* may lack a robust and clear legal focus that copyright protection would otherwise provide. Thus, trying to outlaw access to the software and the concepts contained therein, state to state negotiations and agreements seek to close the legal gap. A further potential factor may be, as we see emerging in Europe, the concern that national legislators impose broad algorithmic transparency obligations. This could erode the competitive advantage that exporters of AI technologies possess and, in terms of intellectual property law, potentially prevent the software from being protected by either trade secrets or unfair competition law.

Irrespective of the source code discipline's origin, the scope of the treaty language is observably inexact. As commented on by Borghi and White, the term "algorithm expressed in source code" has no circulation in information technology law.[1150] From the viewpoint of copyright jurisprudence, "access to" source code and algorithms fall beyond the scope of exclusive rights where the restricted act of reproduction is not implicated.[1151] Furthermore, a prohibition against accessing the software's underlying concepts in the form of its algorithms contradicts one of copyright's core tenets: the idea-expression dichotomy.[1152] Attesting to its foundational significance, alongside TRIPS,[1153] *inter alia* it

[1146] https://www.wto.org/english/tratop_e/ecom_e/joint_statement_e.htm
[1147] D Ravi Kunth, 'Pluri-accord on e-com could reinforce lawlessness in digital trade' (*Third World Network*, 14/8/2019).
[1148] Kristina Irion, 'Algorithms Off-limits?' (2022) Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 1561, n.36 and n.5. Lee-Makiyama, supra 1131, ibid.
[1149] See Chapter 5.
[1150] Borghi and White, supra 1125, 114 (referring to source code discipline wording in the USMCA etc).
[1151] Id. 116.
[1152] Id. 115 ff.
[1153] Article 9.2.

is enshrined in the US Copyright Act[1154] and the European Computer Programs Directive,[1155] both of which expressly exclude ideas, procedures, and principles from the scope of proprietary entitlements.

Thus, although for sound innovation policy reasons the idea-expression dichotomy ensures that concepts and methods of operation remain in the public domain, this most cardinal of copyright doctrines is being systematically rendered null and void by the operation of international trade law. Notwithstanding the Development Agenda and the broad encouragement of technology transfer in TRIPS,[1156] as a consequence of the CPTPP, already the opportunity for some less advanced economies to acquire valuable knowledge and establish better purchase in the AI economy is being undermined. Further, adoption into any instrument emerging from the JSI would have a negative knock-on innovation effect on the majority of states worldwide.

It should furthermore be recalled that the point of market entry is of unique significance for digital technologies, as valuable source code will often be located on servers based extra-territorially, protected from even licit access by the operation of anti-circumvention and anti-trespass legislation.[1157] Thus, while government import regulations and their accompanying procedures could provide a unique means for the unprotected ideas and methods of operation which inhere in the software to be studied and analysed in the interests of economic development, trade law devised by a handful of powerful and technologically developed states is being utilised to frustrate new venues for knowledge transfer.

3. Copyright and its Potential to Alleviate AI Geocentrism

Needless to say, it is no coincidence that the United States and Japan – as architects of the source code discipline – are key players in many ML-dependent areas of the economy, such as e-commerce, search, social media markets, generative AI, robotics, and IoT markets.[1158] According to Deutsche Bank, the combined market capitalisation of the leading American technology corporations – Tesla, Nvidia, Google, Amazon, Meta, Apple and Microsoft – reached $13.1 trillion in Q1 2024, exceeding the combined market worth of all companies listed on global stock exchanges, bar that of the US.[1159] Similarly, IOT export-oriented conglomerates like Nissan and Toyota are pioneers of driverless

[1154] §102(b).
[1155] Article 1.2, Article 3 and Recitals 11, 14 and 16. See also Chapter 4, subsection 3.1.
[1156] Article 66.2.
[1157] For example, copyright laws, not to mention open licences, will frequently permit reproductions for reasons such as reverse engineering, interoperability, back-up etc.
[1158] See Chapter 1.
[1159] http://www.aastocks.com/en/usq/quote/stock-news-content.aspx?symbol=AAPL&id=NOW.1326810

vehicles.[1160] Robotics markets are similarly dominated by Japanese firms with seven of the world's top ten robotics firms based in Japan.[1161]

With respect to the contribution a general purpose technology like AI could make to development, in sharp contrast to the US and Japan, according to UNCTAD the majority of developing countries are characterised by low levels of ML skills, compounded further by low opportunities.[1162] Consequently, absent interventions, developing nations are unlikely to benefit from the transformative effects of Industry 4.0 technologies.[1163] As outlined in the opening chapter of this thesis, for the Global South, the challenges surrounding participation in the AI economy are further exacerbated by a lack of infrastructure and investment, the back-shoring of Industry 4.0 manufacturing to developed countries, and a dramatically reduced foreign investment footprint for digital companies compared to that of more traditional industries.[1164] Attesting to the extent of "algorithmic colonisation", as just one example, Birhane observes that in the e-commerce sphere much of the African continent's "digital infrastructure and ecosystem is controlled and managed by Western monopoly powers such as Facebook, Google, Uber, and Netflix."[1165]

As Korinek and Stiglitz observe, irrespective of a country's current level of development, in the long term it is hard to envisage the emergence of a competitive AI economy.[1166] As discussed in Chapter 1, the reason for this is attributable to a number of interrelated factors: network effects, economies of scope and scale, and a heavy reliance on data. As a result of these intertwined factors, once a company has reached a tipping point in a market, competing with them on even terms becomes highly challenging if not impossible. As noted by Barwise and Watkins, such is the level of dislocation in many e-commerce markets that rather than investing in head-on competition, a competitor is better off eclipsing the market as a whole by developing larger adjacent markets.[1167]

In view of the fact that information technologies have such a marked propensity to induce excessive market power, new innovative forms of regulation are required.[1168] As has been well noted in the literature, traditional competition law tools are often inadequate and have been too slow to adapt to the challenges of digitalisation.[1169] Anti-trust law also fail to address systemic market-wide problems

[1160] 'Japan to allow Level 4 autonomous vehicles on its highways by April 2023' (*Electronomous*).
[1161] Chapter 1 n.69.
[1162] *Industry 4.0 for Inclusive Development* (UNCTAD 2022).
[1163] Id. Fig.7.
[1164] Chapter 1, subsection 3, 1.7.4.
[1165] Abeba Birhane, 'Algorithmic colonization of Africa' (2020) 17 SCRIPTed 2, 389.
[1166] Korinek & Stiglitz, op. cit. 118.
[1167] Barwise & Watkins, supra 133 ibid.
[1168] Anton Korinek and Jai Vipra 'Concentrating Intelligence: Scaling Laws and Market Structure in Generative AI' Institute for New Economic Thinking (Feb 2024). See Fig.1 (showing the high level of market concentration in foundation model markets.)
[1169] See Crémer et al, supra 134, 48 (outlining that because many online services are free, analysis of market share is often inappropriate) See *The Market Power of Platforms and Networks* (Bundeskartellamt 2016) (exploring how the competition law regime should evolve to deal with digital economy antitrust enforcement). David Sparks Evans, Allan Fels, Catherine Tucker (eds), *The Evolution of Antitrust in the Digital Era: Essays on Competition Policy Vol I & II* (Competition Policy International 2020).

effectively as it approaches issues of dominance of an individual company on a case-by-case basis. Moreover, when dealing with transnationals, competition law frequently struggles to establish effective remedies owing to the limited national reach of the authorities.

By contrast, thanks to the WTO and TRIPS, copyright as a tool of regulation has near global effect. Furthermore, as it percolates through all elements of the AI production and deployment cycle, it is a highly apposite tool of regulation. It is proposed, therefore, that an evolution of copyright law's limitations and exceptions could act as one important strand of policy intervention to ameliorate the excessive concentration that is a common feature of information technology markets.

Having made the case for new legal thinking to address the lack of competition in AI markets, the remainder of this chapter now turns to the modalities of a novel software exception that, combined with the introduction of TDM exceptions, could serve as an important tool for knowledge transfer and enhanced market competition. First, the relationship between copyright law and trade law will be discussed. Second, the characteristics and features of the proposed exception are described. Finally, where the software made available pursuant to the novel exception is subject to copyright, its compliance with a traditional approach to the three-step test is discussed, alongside more flexible interpretations.

Section Two

1. Framing of a New Form of AI Copyright Exception[1170]

Having become, in the late 20th century, a pivotal regulator of how information is lawfully used in digital markets, the need to situate intellectual property laws alongside other technology-relevant bodies of law is justifiably more urgent than ever. Whilst intellectual property may tentatively act as a limitation on the exercise of competition law,[1171] its interplay with other IP technology-relevant bodies of law is arguably harder to grasp. The many public interest challenges that digitalisation and datafication pose have, according to some scholars, precipitated a sense of crisis.[1172] Efroni argues that this doctrinal unease has notably contributed to an influx of academic voices calling for a positive integration of human rights perspectives into intellectual property jurisprudence.[1173]

[1170] In the interests of readability, the term "exception" is used to describe the proposal even though copyright relevant reproductions will not be made in all instances, due to the public domain and openly licensed nature of much AI software.
[1171] For an overview of the development between IPR and competition law see Jonathan Turner, *Intellectual Property and EU Competition Law* (OUP 2010) 2-39.
[1172] Zohar Efroni, *Access Rights: The Future of Digital Copyright Law* (OUP 2011) passim. See also Edward Geller, 'Beyond the Copyright Crisis: Principles for Change' (2008) 55 J Copyright Soc'y USA 165, 179.
[1173] Id. 178.

With regards to copyright law, this discourse commonly manifests itself in a call for the redefining of limitations and exceptions as "user rights".[1174] Central to this trend is a reliance on fundamental rights including freedom of information,[1175] freedom of expression,[1176] as well as constitutional law.[1177] As Borghi highlights, beyond theoretical discourse, in a shift driven by the enactment of the Charter of Fundamental Rights of the European Union, the CJEU now explicitly recognises limitations and exceptions, not as mere derogations, but as independent user rights.[1178]

It is from this tradition, the need to appropriately and proportionally give "fair balance" to the rights of creators with wider societal welfare, that a proposal for a new form of access right to trained AI software is most cogently situated.[1179] A novel AI software exception that permits second comers to learn from, and potentially replicate, the models of AI incumbents has the potential to ameliorate the significant economic and territorial concentration that defines a number of core ML-dependent markets.[1180] Through the introduction of TDM exceptions and the type of novel AI forward engineering exception outlined below, the developmental benefits of this new general purpose technology have the potential to be more equitably spread than otherwise is likely to be the case absent strong regulatory interventions.

Before proceeding further, it is first worthwhile stating that in terms of trade law, even where a country is bound by the source code discipline, the introduction of a novel AI software provision in national copyright law is unlikely to contravene its international obligations. This is because it is not a "*condition for* the import, distribution, sale or use of that software". Thus, rather than acting as an obligatory disclosure requirement by the state linked to the import of the software, it functions as a user right that *may* be enjoyed by anyone that has acquired or has access to it.

Additionally, free from trade law constraints, a new form of software exception that provides a right of access to trained AI software and its parameters is also unlikely to engage proprietary copyright entitlements. This is because, as discussed in Chapter 5, much of the software is either open source, a product of the knowledge commons, or – in the case of parameters – a machine-generated array of

[1174] For example, see Niva Elkin-Koren, 'The New Frontiers of User Rights' (2016) 32 Am U Int'l L Rev 1 (discussing the recasting of exceptions by courts as user rights in Canada and Israel). C J Craig 'Globalizing user rights-talk: On copyright limits and rhetorical risks' (2017) Rev 33 Am U Int'l L 1 (discussing the growing prominence of user rights as a term). Christophe Geiger et al 'Limitations to Copyright in the Digital Age, Safeguards for User's Rights, Creativity and Author's Remuneration Interests' in Andrej Savin & Jan Trzaskowski (eds), *Research Handbook on EU Internet Law* (Edward Elgar 2023) 149.
[1175] Marcella Favale, 'The right of access in digital copyright: right of the owner or right of the user?' (2012) 15 The Journal of World Intellectual Property 1 1, 11 (discussing the right of access particularly in the context of TPMs.)
[1176] See Christophe Geiger, 'Copyright as an Access Right, Securing Cultural Participation Through the Protection of Creators' Interests' in R Giblin, K Weatherall (eds), *What if We Could Reimagine Copyright?* (ANU Press 2017) 73, 84 ff.
[1177] Id. 80 ff.
[1178] See Borghi, supra 314 (starting with *Padawan SL v Sociedad General de Autores y Editores de España* (SGAE), tracing the treatment of limitations and exceptions by the CJEU from derogations to autonomous statements of fundamental rights.) 267-270.
[1179] Id. 269-70 (discussing the EU legislative principle of "fair balance" and how the CJEU has moved from an endogenous assessment of the directives to a balancing of different fundamental rights.)
[1180] For example, online advertising, driverless vehicles, robotics, logistics, and social media.

unprotectable numerical values. Nevertheless, where elements of the program do implicate copyright, as a matter of course any exception that provides access to a third party involving reproduction of the work will require the national law of a TRIPS / Berne signatory to comply with the three-step test.[1181]

The need to protect society from the destructive forces unleashed by high levels of asymmetry in information technology markets underscores an urgent requirement for normative advances in the interpretation of the three-step test. Such advancements must recognise the need for a proportionate trade-off between larger policy goals and the interests of individual copyright owners. From the standpoint of jurisprudence, an approach that recognises an inherent margin for discretion within the three-step test will therefore be a prerequisite.[1182] As discussed in more detail below, this author accordingly supports the view put forth by a number of leading European scholars that "[w]hen correctly applied, the three-step test requires a comprehensive overall assessment rather than the step-by-step application" that hitherto has been the norm.[1183] *De lege ferenda,* no single step should be prioritised, thus providing leeway for a more proportionate balance between the interests of individual authors and the greater interests of society at large.

It is of note that leeway for economic development, and by implication the equitable formation of markets, is not without sound foundation in the international treaties. Notably, Part I Article 8 of TRIPS allows signatories to "adopt measures necessary… to promote the public interest in sectors of vital importance to their socio-economic and technological development". Likewise, recognising the importance of technology in the support of development, Article 66 obliges Member States to provide incentives to encourage knowledge transfer to the least-developed countries. Furthermore, to foster international policy support, subject to Recommendation 7 of the "45 Adopted Recommendations under the WIPO Development Agenda (2007)", the World Intellectual Property Organisation is committed to "help countries deal with intellectual property-related anti-competitive practices".

2. Principles of a Novel AI Provision Providing a Right of Access to AI Model Software

We now turn to the principles underlying a novel form of copyright limitation and exception.[1184] By permitting access to AI software that sits behind technological fences, its aim is to induce technology transfer and innovation in a manner that aligns with technological realities, by establishing an access right to the trained models of entities identified by authorities as possessing excessive market

[1181] Art 13 of TRIPS, and Art 9(2) of the Berne Convention.
[1182] Daniël Jongsma, 'The nature and content of the three-step test in EU copyright law: A reappraisal' in *The Routledge Handbook of EU Copyright Law* (Routledge 2021) 338 n.23-4 (citing flexible scholarly reinterpretations of the three-step test.)
[1183] 'Declaration on a Balanced Interpretation of the "Three-Step Test" in Copyright Law' (2008) 39 International Review of Intellectual Property and Competition Law 6 707.
[1184] See supra 1170.

power.[1185] We do not provide a draft model law. Instead, the basic principles for creating a new form of statute-based regulation are outlined, with the intention of stimulating discussion and contributing to the details of any prospective implementation. This author recognises that these goals may not best be suited to concretisation solely in copyright law given the uneven, if not unprotected, nature of core elements of many AI models. Consequently, extra-copyright intervention across a number of differing legal regimes may be desirable.[1186]

1. *Obligations permitting software access limited to designated market incumbents*

Reflective of the fact the policy goal of a new form of software exception is to ameliorate the negative consequences that arise from information technologies propensity to create "winner takes all" markets,[1187] the right to access the software should apply solely to market incumbents in AI-dependent markets that have either tipped, or the authorities designate as being characterised by a structural or systemic lack of competition.[1188] Thus, only those AI models deployed in markets designated as uncompetitive fall under the novel exception regime proposed, not all.

Evincing of the legal notion of proportionality, as embodied in the Digital Services Act's hierarchy of online platforms obligations, a tiered approach to accessing AI software recognises the particular impact on the economy and society that severe market asymmetries engender.

2. *Obligations for software disclosure linked to the incumbent's exercise of relevant data analytics exceptions*

Obligations in return for the conferral of rights is not an uncommon concept in the field of intellectual property. In patent law, protection is conditional on an obligation to disclose the inventions that have been registered.[1189] In terms of unregistered rights, Art. 8.1 of the Proposal for a Council Directive on the Legal Protection of Databases is of particular note as, although absent from the enacted Directive, it sought to create an obligation on licensors of single-source databases, requiring that their databases be made available on "fair and non-discriminatory grounds."[1190]

[1185] See supra 1127. It is interesting to observe that much of the early literature referring to access rights arose in response to concerns that digital technology and piracy would undermine the legitimate interests of authors. Situated in discussion around TPMs, an access right was conceived, not as right for users to obtain access, but for authors to control access. For example, Simon Olswang, 'Access right: an evolutionary path for copyright into the digital era?' (1995) 17 EIPR 5, 215. Jane Ginsburg, 'From having copies to experiencing works: the development of an access right in US copyright law' (2003) 50 J Copyright Soc'y USA 113. Cf. see Marcella Favale 1175, 14-16.
[1186] For example, in Europe a new raft of legislation provides access to algorithms in certain instances such as the AI Act, Digital Markets Regulation and the Digital Services Regulation.
[1187] See Korinek & Stiglitz, op. cit. 118, and Crémer, supra 1341169.
[1188] See Chapter 1 discussing a right of access to algorithmic software and data under the Digital Services Act, Digital Markets Act and AI Act. The Data Act (2023/2854) also creates a right of access to data for public authorities in the instance of a public emergency.
[1189] Article 83. Disclosure of Invention. European Patent Convention. Entry into force 7/10/1977.
[1190] Proposal for a Council Directive on the Legal Protection of Databases. Article 8.1 and Recitals 31-5.

Rightsholders identified by the authorities as holding excessive market power may be reluctant to provide access to their software, necessitating the imposition of sanctions for non-compliance. Thus, the ability to continue to enjoy the exercise of any permitted acts facilitating data analysis could be directly predicated on the ability of beneficiaries of the novel software exception to access the trained model. Where such derogations are not present in national law, or incumbents' models are created reliant on extra-territorial exceptions, if access is not forthcoming, other remedies will be required.[1191] Beyond court orders mandating access to the AI models, in extremis, the termination of sale of any product or service that relies on the exercise of those exceptions could also be contemplated.

3. *Object of the software disclosure*

A market incumbent identified by the authorities should, upon request from a third party, be required to give access to trained AI model software to permit its reproduction and subsequent analysis. Additionally, the parameters deployed in the market or markets designated as uncompetitive should also be disclosed. As discussed in Chapter 5, the core elements of a trained model will often not attract proprietary copyright as they are a combination of open source software, commonplace expression, and in the case of the parameters, data-written code lacking the personal creative stamp of a human author. Accordingly, where copyright is absent, the authorities will be able to design a remedy free of the three step-test.

Where a novel AI software exception provides access to copyright protected materials, consistent with the idea-expression dichotomy, the independent programming of a model solely reliant on ideas, functions, and public domain subject matter present in the accessed software will not implicate exclusive rights. In other words, as reliant solely on public domain elements, the coded expression in the derived software will not implicate the proprietary entitlements of the original work.[1192] Thus, on first blush, the application of the three-step test will not be required.

At a technical level, owing to the inherently procedural nature of much of the copying undertaken by computer technologies, accessing the model in digital form will however where copyright inheres in the code engage the exclusive right of reproduction as a precursor to being able to read and analyse the incumbent's model program. Where such copies constitute transient acts, a number of jurisdictions have exempted temporary copies required for the automated processing of copyright works from the reach of proprietary entitlements.[1193] In the absence of relevant exemptions in national law, or where

[1191] See Stability AI defence submitted 27/3/24 in *Getty et al v Stability AI* EWHC 3090 which implicitly relies on fair use as Stability's Stable Diffusion model was developed in the US. §5.1.
[1192] See Chapter 4 outlining different copyright exceptions and doctrines which permit the unauthorised reproduction of software code. e.g. reverse engineering, merger, scènes à faire, the idea-expression dichotomy etc.
[1193] For example, Article 5.1, Information Society Directive. The picture is more complicated in the United States where derogations for temporary copies fixed in a tangible medium are exempted as a consequence of a mix of statute and case law. See Maria Pallante, 'The next

the act of accessing the works for analysis falls beyond their intended scope, akin to the logic underpinning the lawful reproduction of published patent specifications,[1194] deeming the procedural copying necessary for software access and review as a violation of the three-step test would – from a teleological perspective – undermine the innovation policy objectives that the proposed provision seeks to advance.

From the standpoint of jurisprudence, it should be observed that traditionally copyright has implicitly regulated reproduction in the context of its dissemination of a work.[1195] Thus, in principle at least, leaving the act of accessing a work for the purposes of reading beyond the reach of author's entitlements.[1196] Pursuant to this line of argument, normatively, legally-mandated digital access regimes[1197] should therefore not contravene the three-step test. *De lege ferenda,* reading and the learning this engenders, are uses which ought to fall beyond copyright's (digital) exploitation rights and, for innovation policy reasons, intermediate copies that facilitate these acts are activities that the law should actively incentivise.[1198]

Such an approach focusing on tangible software subject matter alone is not without its shortcomings. As outlined in Chapter 1, the ability to compete in AI markets will require *inter alia* access to large volumes of the appropriate kind of data (veracity), combined with the availability of a robust scalable technical infrastructure. For organisations or countries that lack the skills necessary to utilise the accessed subject matter, the additional acquisition of know-how is particularly vital. In common with other applied technologies, less formal know-how garnered from trial and error, experience, conference attendance, online communities, etc, also comprise an important input to model design. At the very least, reflective of technology transfer obligations under TRIPS,[1199] and accompanying an access right to tangible information assets, the authorities overseeing the novel software exception should also encourage the proportionate transfer of valuable tacit know-how.[1200] Notably, for

---

great copyright act' (2012) 36 Colum JL & Arts 315, 325 ff. (in addition to fair use, outlining some of the statutory responses to the automated copying undertaken by computers.)

[1194] Dean Alderucci, 'Copyright Protection for Patents: Some Surprising Implications for Artificial Intelligence' (2021) 123 W Va L Rev 791, 831 (outlining the different legal bases for exempting patent specifications from copyright protection in Switzerland, US, Canada and Australia.)

[1195] C-360/13 *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others* [2014] ECLI:EU:C:2014:1195. Abraham Drassinower, 'Authorship as Public Address: On the Specificity of Copyright vis-a-vis Patent and Trade-Mark' (2008) Mich St Law Review 199 (referring to copyright as "an exclusive right of public presentation".) Matthew Sag, supra 190, 1626. See also, infra 1196, Ginsburg, Lemley and Casey.
(in which the European Court ruled that temporary copies required to access a work online "forms part of … viewing") 61.

[1196] Favale, supra 1175 (commenting that "… TPMs seem[] to expand traditional copyright protection, by creating an additional right for the owner: 'The right of access' ") 1. Jane Ginsburg, 'Copyright and control over new technologies of dissemination.' (2001) 101Columbia Law Review 1613, 1632 (explaining that traditional rights in copyright do not dictate access and use to a work owing to the doctrine of exhaustion.) Mark Lemley and Bryan Casey, 'Fair Learning' (2021) 99 Tex L Rev 743 (situating unregulated acts such as reading, listening and exhaustion under the idea-expression dichotomy as important unregulated legal incentives to allow others to create further works.)

[1197] See examples of laws that mandate access to copyright works n.1186.

[1198] As far as this author is aware there have not been any questions raised concerning copyright relevant reproductions and the three-step test in the context of mandated access to software by authorities in the context of the Digital Services Act, Digital Market Act, and EU AI Act.

[1199] Article 66.

[1200] Where know-how is formally protected it is usually not protected by copyright but by contract or trade secrets. See Richard Li-dar Wang, 'Ancillary Orders of Compulsory Licensing and Their Compatibility with the TRIPS Agreement' in Reto M Hilty, Kung-Chung Liu, countries that are currently subject to software and algorithm disciplines, the transfer of intangible knowledge assets falls beyond the scope of PTA obligations as currently they are restricted to solely the transfer of the software and "the algorithm expressed in that source code".[1201]

4. *Requirement for remuneration*

In recognition of the R&D investments undertaken by dominant players, a fair, reasonable, non-discriminatory licence subject to appropriate remuneration should be entered into by organisations wishing to utilise the intellectual labour of the market incumbent. As the aim of the novel exception is to primarily stimulate competition in sectors of the economy which are designated as exhibiting market failure, remuneration set by the authorities should be at levels that do not act as a barrier to new competition.

Remuneration could be lower or removed for publicly funded organisations, charities, scale-ups, start-ups, and SMEs in order to promote future market and societal innovation. Consideration should also be given to future application of any knowledge acquired through the exception.[1202] For example, in terms of market proximity, lower rates for use of the knowledge in the dominant incumbent's core market – as well as adjacent markets as defined by the relevant authority – may be appropriate to stimulate more competition, with higher rates being more applicable where the technology is deployed in new and unrelated sectors of the economy.[1203]

5. *Appointment of a competent authority*

Accompanying the introduction of a novel software exception, governmental oversight of both the legal and technological aspects of its oversight will be required. Considering the principles outlined above, it will be essential to address decisions concerning *inter alia*: designation of excessive market power and the software implicated, the terms and conditions of a licence, remuneration, the presence or absence of copyright, the management of know-how transfer, mechanisms for dispute resolution, the appropriate currency for the parameters made available,[1204] and issues arising from non-compliance.

Compulsory Licensing: Practical Experiences and Ways Forward (Springer 2015) (proposing that in terms of trade secrets the requirement to divulge know-how can be justified in line with Art 10 *bis* 2 of the Paris Convention for the Protection of Industrial Property.)

[1201] For example, see Article 19.16.1 of the USMCA.

[1202] See Article 9.4 of the Data Act which limits the remuneration for data supplied to SMEs or not-for-profit organisations to below cost.

[1203] A similar point is made in P Samuelson et al, supra 761, 2308, 2418-9 (proposing a market proximity approach to legal regulation and renumeration in software markets.)

[1204] Some model parameters will be updated extremely frequently and therefore determining the age and frequency of the parameters provided to licensees will require a decision by the competent authority.

From a technical perspective, an optimal-access infrastructure will have to be established. Whether this involves creating a repository of code and all accompanying literature under the control of the competent authority, or reliance on the incumbent's own infrastructure, will require consideration and potentially significant public investment if the former option is concluded to be optimal.

It may also be beneficial for the authorities to oversee interrelated innovation-focused activities. This may include the deposit of publicly funded AI software in an open repository,[1205] or in the European Union, undertaking complementary responsibilities for competent authorities forthcoming from the Data Act,[1206] Data Governance Act,[1207] AI Act, etc. In this manner, a number of machine learning innovation-oriented systems could be linked under one umbrella with the effect of reducing transaction costs for government and third parties alike.

6. *Endogenous to copyright law the explicit promotion of competition and public welfare by the authorities*

Neither the amelioration of market concentration in AI markets, nor the safeguarding of the private interests of authors, should be regarded as an end in themselves. The benefit of addressing the systemic lack of competition in information markets will profit not just a wider cohort of companies, but the public at large. With more actors incentivised, greater innovation will be possible across a wide number of markets and sectors of society. Moreover, transfer of technology know-how can act to spur AI innovation, not just in the discrete sectors of the domestic economy, but potentially transnationally, thus benefiting countries at all levels of development.

By explicitly framing the goal of a novel copyright exception to be one of competition and public welfare, it serves to recognise the scope and broad societal impact that copyright has in the datafication age. No longer is copyright an abstruse field of law; since the advent of mass-market reproduction technologies, and the emergence of the internet, its effect on the everyday life of people has been profound. In view of this, an explicit expression by the authorities of the goals of a novel AI software exception has a number of benefits. As we see in Art 30-4 of the JCA, the increasingly high demands placed on copyright in the digital age prompted the explicit codification of its objective as one that enables the use of a work for the purposes of enjoyment. As a consequence of this, the

[1205] For example, as a literary work, governments could include publicly funded trained software as a subject of mandatory publication pursuant to a secondary publishing law.
[1206] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 OJ L2023/2854 22.12.2023.
[1207] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 OJ L152/1.

authorities have been provided with an effective jurisprudential tool for guiding future law-making.[1208]

Those voices who seek to overstretch copyright law, and undermine the balance at its heart, are often reluctant to accept public interest interventions. Thus, a further benefit that can be derived from defining competition and public welfare as the goal of a novel AI software exception is that it may be easier for the legislature to resist counterpressure to over-protect and further extend existing information monopolies.

To sum up, the successful exercise of a novel AI decompilation exception is predicated first and foremost on a right of access. As the designated AI models belonging to market incumbents will be protected by a myriad of network security protocols, without a right of access to the software / parameters as a precursor to analysis and forward-engineering, the public policy goals of the novel exception will not be realised. Augmenting this, in order to allow analysis, an exemption to the right of reproduction will also be required.

While a growing body of European digital laws allow some access to AI software and algorithms,[1209] importantly a number of access-enabling mechanisms are already embedded in copyright law.[1210] The most analogous access regime to the one proposed is that of TPM circumvention in countries where statute tasks an authority to intervene to grant a beneficiary of an exception full access to a copyright work. Nevertheless, the very high levels of economic and geographic concentration that we see in AI sectors of the economy necessitate a higher calibration in approach. Therefore, under the novel software provision proposed, it would be necessary not only for the authorities to identify areas where market power is excessive, but to guarantee access to the AI software and parameters in return for fair levels of remuneration.

3. A Novel AI Software Exception and the Role of the Three Step Test

3.1 Cumulative Interpretations of the Three-Step Test

Whereas copyright and related rights in the digital environment unquestionably serve some actors well, whether it supports public welfare as a whole is more open to question. In the case of AI, the

[1208] See Chapter 3, section one, subsection 1.3.
[1209] Supra 1188.
[1210] Christophe Geiger and Bernd Justin Jütte, 'Copyright as an Access Right: Concretizing Positive Obligations for Rightholders to Ensure the Exercise of User Rights' (2024) 73 GRUR international 11, 1019 (citing in Europe, TPM circumvention procedures, contractual overridability, and obligations to contract in Art 15 of the AVMSD Directive and Art 17(4) of the CDSM Directive as examples in the copyright acquis which create obligations on rightsholders to allow the exercise of rights.)

natural monopoly effects of information technologies[1211] risk economic hegemony, not just by superstar firms, but by superstar countries who derive direct benefit from the presence of local AI agglomeration economies. Accordingly, as asserted by Korinek and Stiglitz, the geographic imbalance in the distribution of the fruits of machine learning will act to impede equitable economic development across the globe.[1212] Given the turning point for society that AI indubitably represents, a novel software exception that mandates access and provides the means for greater competition has the potential to reinvigorate existing markets and drive competition across the digital economy.

Bound by *inter alia* TRIPS and the Berne Convention, as with all statutory flexibilities, a new software exception implicating the reproduction of protected subject matter must comply with the three-step test.[1213] However, as shown, oftentimes a trained AI model will *not be* subject to proprietorial copyright protection. Where this is the case, free from the uncertain confines of international treaties, signatories will be at liberty to mandate access to trained model software and model parameters in line with their own public policy needs.

This subpart examines the proposed novel AI provision in the light of the TRIPS / Berne three step test (3ST). As stated above, this author's view is that access as an automated step to reading and viewing should normatively fall beyond the reach of authorial entitlements. Be that as it may, as all forms of technical reproduction involve reproduction of some sort, compliance by the authorities with the three-step test is required. A traditional approach to the 3ST will commend that each of its prongs is evaluated cumulatively.[1214] Devoid of intrinsic normative principles as to which policy goals are particularly meritorious, or how the test should be balanced with other bodies of law,[1215] as widely observed, its focus on private rights has the propensity to hinder development and welfare-enhancing policy innovation.[1216]

<u>First Step</u>

According to Gervais, the requirement under the first prong of the three-step test that any exception is limited to "certain special cases" is largely redundant.[1217] This is because, in order to comply with the

[1211] See Chapter 1, section 3 (outlining six factors which underpin a lack of competition in many AI markets.)
[1212] Korinek and Stiglitz, op. cit. 118, 15.
[1213] In addition to TRIPS and the Berne Convention, the three-step test is also incorporated into the WCT (1996), WPPT(1996), Beijing Treaty on Audiovisual Performances (2012) and the Marrakesh Treaty (2013).
[1214] Sam Ricketson, *The Berne Convention for the Protection of Literary and Artistic Works: 1886-1986* (Kluwer 1987) 482.
[1215] Jerome Reichman and Ruth Okediji, 'When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale' (2012) 96 Minnesota Law Review 4 1362, 1390.
[1216] For example, see Christophe Geiger, 'The Social Function of Intellectual Property Rights' (2013) Max Planck Institute for Intellectual Property and Competition Law Research Paper No.13-06 (setting out that the social function inherent in all laws means that rights must always be "put in context with other rights of equal value and with collective interests".) Jerome Reichman, 'The limits of "limitations and exceptions" in copyright law' in Ruth Okediji (ed), *Copyright Law in an Age of Limitations and Exceptions* (CUP 2017) (arguing for a contextualisation of copyright within a broader set of equally valuable bodies of law.)
[1217] See Daniel Gervais, 'Towards a New Core International Copyright Norm: The Reverse Three-Step Test' (2005) 9 Marquette Intellectual Property Law Review 1, 15 ff. Jongsma, supra 1182, 349 (observing that frequently the CJEU has entirely eschewed any analysis of the first step.)

subsequent two factors, any exception must be circumscribed in its design and application. Ricketson[1218] and Senftleben[1219] specify that any exception must be justified by a rational public policy need. In terms of the proposed novel software exception, on the one side framed by legitimate regulatory considerations concerning a lack of competition in designated AI markets, it is targeted narrowly at those actors who wield excessive market power as expressly determined by the authorities. On the other, the beneficiaries are limited to licensed technology companies and other organisations with the capacity to understand and adapt the discrete elements of the accessed AI software model that are protected by copyright. Considering the well-defined and narrow scope of the proposed exception it therefore readily aligns with the notion of a special case as prescribed by the TRIPS / Berne three-step test.

Second Step

Compliance with the second criterion is uncontestably more problematic. As observed by Geiger and others, owing to its exclusive focus on economic rights, when considered in isolation, the second prong would preclude "law-makers from taking into account any interests other than the private economic interests of right-holders."[1220] Without doubt, where the designated software is protected by copyright, its use by third parties in an incumbent's primary markets will conflict head-on with the normal exploitation of the work.[1221] Thus, in concordance with the traditional cumulative interpretation of the 3ST, any exception which unreasonably limits a market, or permits competition in ways that conflict with how "right holders normally extract economic value from that right to the work",[1222] is non-compliant with the second step.[1223]

Third Step

Precluding *unreasonable prejudice* to the legitimate interests of the rightsholder, academic interpretations of the three-step test regard the third prong as carrying comparatively more interpretive weight than the second, owing to the fact it incites an analysis of the justification for the exception.[1224] Similar reasoning has also been clearly applied by the courts.[1225] Further, as the French language

[1218] Ricketson, supra 1214, ibid.
[1219] Martin Senftleben, *Copyright, Limitations and the Three-Step Test* (Kluwer Law 2004) 258.
[1220] Christophe Geiger et al, 'Towards a Balanced Interpretation of the 'Three-Step Test' for Copyright Exceptions' (2008) 4 EIPR 489, 490.
[1221] In terms of the second step and the question of using software in ancillary markets, in *Google LLC v Oracle America Inc,* 141 S Ct 1163 (2021) 1207. Google's use of 11,500 lines of declaratory code was ruled as not creating a market harm inter alia on the grounds that it was deployed in a different market to that of Oracle. Thus, indicating that market harms may only arise in primary markets.
[1222] *Report of the WTO Panel* ("WTO Panel Decision") dated June 15 2000, WT/DS160/R. §6.183.
[1223] Id. 183-5. See also Gervais, supra 1217, 17.
[1224] Christophe Geiger, Daniel Gervais, and Martin Senftleben, 'The three-step test revisited: How to use the test's flexibility in national copyright law' (2013) 29 America University International Law Review 581, 604. See supra 1217, 18 (Gervais arguing that the third step is a "key that allows legislators to establish a balance".) Senftleben, supra 1219, 132 (describing the third prong as an "adjusting tool" without which a balance between public and private interests cannot be achieved.)
[1225] *England And Wales Cricket Board Ltd & Anor v Tixdaq Ltd & Anor* [2016] EWHC 575 (Ch) 92 (opining that the third step qualifies the second step: "… it is not sufficient for an exception not to apply that there is some conflict with the copyright owner's legitimate interests,

version makes particularly transparent – *préjudice injustifié* – the copyright owner is not entitled to control *all* uses of their work. Thus, allowing for *justifiable* public policy considerations to come into play when balanced with the legitimate interests of the copyright owner.

In terms of the application of the third factor to the proposed novel exception, it is firmly justified on both social and economic public policy grounds. Furthermore, considering that any market incumbent *ex post* identified as wielding excessive market power by the authorities have *already* established a dominant position, a targeted intervention aimed at discrete proprietorially protected sub-elements of an incumbent's model software is unlikely to unreasonably prejudice market conditions.[1226] In fact, it is hard to conceive that an "unreasonable loss of income"[1227] will arise from the reproduction of limited amounts of protected code when market dominance stems overwhelmingly from extra-copyright factors including privileged access to data, network effects, and economies of scope and scale.

Furthermore, it should be recalled that the effect of the three-step test in protecting the legitimate interests of a rightsholder should not extend beyond those elements of the software which are subject to proprietorial protection. Thus, the highly valuable parameters, open source components, and other programs derived from the knowledge commons will fall beyond the reach of the TRIPS / Berne three-step test. Considering the multitude of factors that contribute to an incumbent's dominant position in a market, as in *Google v Oracle,* even if the isolated elements of the software protected by copyright were to be reproduced in their entirety by second comers, it seems improbable that such actions will impose an unreasonable economic burden on the copyright owner.[1228] This is almost certainly the case in core markets where a corporation has already established dominance. More detailed analysis by the authorities on a case-by-case basis will, however, be required where protected software elements are deployed in adjacent and new markets.[1229]

The compensated nature of the novel software exception may also contribute as a factor weighing in favour of 3ST compliance. Although owing to the test's second prong it is not deterministic,[1230] it is

including the copyright owner's normal exploitation of the work. Rather, the exception can apply unless those interests are unreasonably prejudiced."

[1226] The term dominance is used to indicate excessive market power as designated by the authorities, rather than inferring an analysis of dominance as typically applied in European competition law.

[1227] WTO Panel Decision §6.229.

[1228] *Google LLC v Oracle America Inc* 141 S Ct 1163 (2021). The Supreme Court finding that Google's use of only a small percentage of the Oracle platform software favoured fair use.

[1229] Cf. Samuelson, supra 1203.

[1230] Senftleben, supra 1219, 80 ("if the legitimate interests of the author are prejudiced … to such an extent that a reasonable level even cannot be reached by providing for the payment of equitable remuneration, the limitation is still impermissible.")

posited that the exception's requirement for remuneration can nevertheless play a role in mitigating unreasonable prejudice and economic loss.[1231]

In sum, whether a novel software exception would be compatible with the international treaties – pursuant to the traditional cumulative analysis of the three-step test – is a moot point. As discussed, the second prong's exclusive focus on an author's economic interests may well preclude the adoption of the novel software exception as proposed. However, as examined in the next subpart, greater levels of certainty and regulatory leeway can be found to support such an exception in interpretations which balance and situate the three-step test within a wider legal and public policy system.[1232]

3.2 Flexible Normative Interpretations of the Three-Step Test

More flexible interpretations of the 3ST have been proposed by a number of scholars. Geller argues that the three-step test should not regulate either definitional limitations endogenous to copyright law (e.g. the idea-expression dichotomy, exhaustion etc) or constitutional limitations including fundamental rights such as freedom of expression and freedom of information.[1233] In a similar vein, Bently and Aplin posit that as a mandatory obligation provided by Berne, the quotation exception is a definitional limitation excluded from the application of the three-step test.[1234] Of more direct relevance to a novel software exception, however, is the well-known declaration on a "Balanced Interpretation of the 'Three step-test' in Copyright Law" put forth by a group of European scholars.[1235] Addressing the negative impact on public policy that a narrow interpretation of the TRIPS / Berne 3ST encourages, the declaration's approach is conceptually simple in that it rejects the traditional cumulative interpretation that has hitherto been the norm.

Acknowledging that the second step in particular can prevent the operation of legitimate public policy interventions, the declaration sets forth that the "three steps are to be considered together and as a whole in a comprehensive overall assessment."[1236] As a result, even were the prospective software exception is found to be incompatible with the second prong, if based on justifiable and proportionately drawn public policy grounds, a holistic balancing of all three factors could still render it compliant with international treaties.

[1231] Senftleben, ibid. See also WTO Panel Decision §6.95 where the panel states "TRIPS Agreement allows Members to substitute a compulsory licence for an exclusive right ... or determine other conditions *provided that they are not prejudicial to the right holder's right to obtain an equitable remuneration*."

[1232] It is of note that Victor Hugo, the founding father of the Berne Convention, also argued for the pre-emption of the public interest over the rights of authors (Opening Speech at the International Literary Conference (17/6/1878)) ("Si l'un des deux droits, le droit de l'écrivain et le droit de l'esprit humain, devait être sacrifié, ce serait, certes, le droit de l'écrivain, *car l'intérêt public est notre préoccupation unique, et tous, je le déclare, doivent passer avant nous.*" (Emphasis added).

[1233] Paul Edward Geller, 'A German Approach to Fair Use: Test Cases for TRIPS Criteria for Copyright Limitations' (2009) 57 Journal of the Copyright Society of the USA, 553.

[1234] Tanya Aplin and Lionel Bentley, *Global Mandatory Fair Use* (CUP 2020) Chapter 3.

[1235] Geiger et al, supra 1220. For further balanced interpretations of the three-step test see Jongsma, supra 1182, n.21-22.

[1236] Geiger et al, id. 492.

Applying the Declaration's interpretation of the three-step test, as discussed above, the circumscribed nature of the rightsholders and beneficiaries of the novel software exception clearly constitute a special case under the first prong. The limited economic harm, a requirement for remuneration, and the public welfare benefits of well-functioning AI markets also act as justifiable public policy considerations under the third step to limit the scope of the copyright owner's exclusive entitlements. Therefore, it is posited that, in unison, these could legitimately counterbalance the potential conflict with the normal exploitation of the work by the incumbent under the second prong.

A flexible approach to the three-step test that would support the adoption of a novel AI software exception may find purchase even in unexpected quarters. As pointed out by one commentator, the European Court's permissive approach to the order of the three steps suggests that it may well endorse the balanced interpretation articulated in the Declaration.[1237]

Moving from the specificities of a novel software exception to jurisprudence more widely, additional benefits arising from a redefinition of the three-step test can readily be envisaged. In the tradition of legal realism, the 3ST has the facility to act as a positive tool of law making, bringing societal interests and public policy firmly into the decision-making process of the courts. Extricated from a cumulative interpretation, it becomes more than a solipsistic end in itself, focused one-sidedly on the proprietary entitlements of private interests. As a consequence of bringing the interests of copyright owners and the public at large more into line, it will serve better to reinforce the prime goal of copyright law: the incentivisation of further creativity. Furthermore, from a regulatory standpoint, it will better ensure that copyright no longer creates friction or sits in contradiction with other equally important bodies of law.

Conclusion

As argued in this thesis, the stark imbalance in AI markets demands a radical policy response. As discussed, faced with trade law developments encroaching on the public domain, exceptions to copyright is one area where state interventions could help ameliorate an endemic lack of competition. A legal realist approach to the three-step test provides authorities the means to address severe social and economic harms in a manner which the traditional approach fails to guarantee. By recognising that copyright law is not a narrow goal existing in its own right, and normatively should find its place within a hierarchy of policy considerations, the three-step test can serve the interests of not just private economic actors but humanity more broadly.

[1237] Paulien Wymeersch, 'EU Copyright Exceptions and Limitations and the Three-Step Test: One Step Forward, Two Steps Back' (2023) 72 GRUR International 7 631, 641.

Chapter 7: Conclusion

This research is grounded in the belief that without radical policy intervention, AI will further entrench deep disparities in the digital economy – divisions which traverse the traditional Global North and South paradigm that underpins much of development economics and political theory. As indicated by the title, this thesis has examined both sides of the copyright bargain – exclusive rights and their derogations – to evaluate how well copyright supports and incentivises AI. As implied by Samuelson's observation that copyright is the only law capable of "bring[ing] generative AI systems to their knees"[1238], the scope of authorial entitlements encroaches into every corner of machine learning. As a general purpose technology deployed across all sectors of the economy, the contours of copyright are therefore vitally important, influencing both incentives to invest as well as competition in digitalised markets.

Based on a comparative and black letter law doctrinal approach, it is concluded that wider and more flexible limitations and exceptions to support AI can contribute to heightened competition and growth. Countries wishing to support their AI economy, have much to learn from not only the US and the EU, but also Japan, who after over a decade of deliberation sought to support its research-intensive digital economy through the medium of a qualified general exception.

In terms of exclusive rights, it is submitted that the weak copyright protection afforded many AI models has spurred new developments in international trade law. It has influenced the introduction of a source code and algorithm discipline into a number of PTAs, and interrelated to this, then become a significant focus of the ongoing discussions on digital trade at the WTO.

The importance of AI as a general purpose technology, coupled with the intense market disequilibrium that powerful trading nations are actively entrenching through trade law, underscores the need for new thinking. This is why this author proposes that in order to promote development, nations not only introduce open norms modelled on those of Japan and the US, but adopt a new legal framework for approaching reverse-engineering targeted at market incumbents.

In more detail, first, it was determined that the exercise of TDM limitations and exceptions take vastly different forms across the four jurisdictions that are the subject of this thesis. As discussed in Chapter 2, putting the question of the lawfulness of cross-border collaborations aside, basic research-oriented non-commercial ML is well facilitated in the EU copyright framework, safeguarding it from both contractual and technological override. In contrast, commercial AI receives limited support. Entirely

[1238] 'Generative AI meets copyright law' (2023) https://www.youtube.com/watch?v=6sDGIrVO6mo outlawed in the UK unless it falls within the confines of the temporary copying exception, unique in Union law, the EU has to all intents and purposes left the exercise of commercial AI involving protected works to the volition of rightsholders via a rights reservation mechanism. As a result, it was concluded that not only does copyright law render Europe a comparatively hostile market to invest in AI, but the limitations placed on commercial access to works by the rights reservation mechanism will serve to increase the risk of bias and poor model predictions. Something which runs contrary to both the public interest and, paradoxically, the widely espoused objectives of the newly implemented AI Act.

In contrast to Europe, the Japanese legislature has strongly supported its own AI industries. Article 30-4 creates an open norm, facilitating both digital and analogue uses of authorial works for "non-enjoyment" purposes. In doing so, it serves to enunciate in codified form the telos of copyright itself. Furthermore, unlike the United States which has primarily relied piece-meal on fair use rulings to support TDM, Japan has systematically employed legislation in combination with authentic interpretation to not only permit copyright-reliant AI but safeguard it from contractual and technological override. Thus, not just as internal scope limitations acting as a juridical end in itself, but as an expression of wider economic and policy interests, in Japan, AI exceptions take the form of not just derogations from authorial privileges but full-blown user rights.

Having determined the asymmetrical nature of ML exceptions across several of the world's major economies, the analysis turned to the role of exclusive rights in protecting AI computer programs – an aspect of copyright law that has received little attention in the scholarly literature. Much has changed since 1994 when TRIPS rendered copyright protection of software mandatory across all WTO member states. The ubiquity of openly licensed AI toolkits, a heavy reliance on heuristics and the knowledge commons, combined with the fact that the most economically valuable elements of ML code are generated by the data rather than human authorship, together, suggest copyright protection of ML software is significantly less common than for other types of computer program.

Despite this, likely attributable to rightsholders securing adequate protections through the application of technological protection measures, licensing, and a reliance on extra-copyright statutory protections, as of yet there have been no significant calls for further safeguards in the copyright framework to protect AI software.

Conversely, driven by the flimsy copyright protection afforded AI software, robust safeguards for a company's investments in AI have been forthcoming in trade law. As examined in Chapter 6, the emerging software and algorithm disciplines in preferential trade agreements seek to foreclose as a condition of import, access to the very ideas and functionality that underpin ML models. From a jurisprudential perspective, this runs counter to one of copyright's cornerstone principles – the idea-expression dichotomy – designed to leave ideas and methods of operation beyond the reach of proprietorial entitlements in order to encourage further creativity and societal advancement.

Information technology markets tend to lack competition. It is of no surprise therefore that markets reliant on AI, such as e-commerce, IoT and robotics are dominated internationally by a handful of incumbents emanating from a very select number of countries. As has been set out in this thesis, the pervasive presence of copyright law across the AI lifecycle recommends it as one important policy lever to alleviate intense market disequilibrium. Moreover, it is one that could have global effect if effectuated through international treaties.

In order to increase competition, informed by the "patent bargain" and the growing number of obligations in EU regulations to disclose data to the authorities, a new form of limitation and exception to copyright law has been proposed. Accordingly, beneficiaries of ML exceptions defined as dominant in a national market are required to share their trained model software and parameters with third parties. By spreading knowledge and engendering subsequent technological innovation, such an approach serves to advance the underlying goals of copyright itself. The thin or entire lack of proprietary copyright protection afforded to many AI programs, further acts to expedite the intervention. From the standpoint of the three-step test, even where proprietary subject matter is present in the model, this author contends that a flexible interpretation that provides space for wider regulatory concerns beyond that of just copyright alone, will permit the adoption of the novel AI software exception proposed into national law.

The very real economic significance of machine learning, destined to further reinforce existing market inequalities, underscores the necessity of a multi-pronged regulatory response. This author asserts that owing to the prominence of copyright across the AI ecosystem, policymakers should acknowledge more comprehensively than they have done to date, copyright's potential as a key tool to foster R&D, competition and positive economic change. Whilst protecting the legitimate interests of rightsholders is the primary rationale underpinning the copyright framework, the thin defence it affords ML software compared to private ordering and digital locks warrants a re-evaluation of incentive rationales. Thus, this author submits the focus of policymakers should shift from private exclusive rights toward the role of limitations and exceptions in promoting competition and innovation.

In terms of market regulation, Japan is of particular note. It has dedicated significant governmental resources to codifying a highly flexible AI-enabling copyright exception. It is this strategic attention to copyright and its influence on the development of AI markets, both domestically and internationally, that likely explains the key role Japan has played in advancing the emergent software and algorithm disciplines in international law.

In sum, the adoption of AI by the world's largest technology and manufacturing companies, frequently predicated as it is on monopolistic or quasi-monopolistic access to data, is ineluctably reinforcing deep-seated economic inequalities across the globe. To the extent that copyright law influences the shape of the AI economy, its design matters a great deal. To support equitable development in an area of the economy strongly characterised by natural monopolies, new competition focussed copyright norms must emerge.

Central to this, are the incentives provided by technology and machine learning focussed limitations and exceptions in national and international law. It is submitted that L&Es that allow both commercial and non-commercial data analytics and model training irrespective of hindrances from contract and technological protection measures are a prerequisite of any competitive economy. Moreover, by acknowledging that information technologies such as AI have a tendency to foster natural monopolies, it allows policymakers to envisage new forms of user privileges, such as that put forward in this thesis.

Faced by unprecedented concentrations of wealth in the hands of a few companies predominantly based in North America and East Asia, it behoves policymakers and the international community to develop new thinking internal to copyright law. L&Es that encourage AI, while ameliorating excessive monopoly power through a new copyright bargain such as that proposed herein, present a persuasive regulatory lever consistent with broader IP traditions and the equitable development aspirations of both governments and their citizenry.

Recommendations for Legislators

1. Owing to the importance of AI as a tool of research as well as a significant potential contributor to economic growth, policy makers should (re)evaluate how best limitations and exceptions can facilitate the further adoption of machine learning. Whether countries already have introduced AI-related L&Es, or have yet to adopt them, governments should ensure that both commercial and non-commercial ML is possible as well as ensure that neither contractual provisions, reservation of rights nor technological protection measures unnecessarily hinder the use of copyright works for AI purposes.

2. Based on the research findings, a hybrid approach to text and data mining exceptions is most recommended. The flexibility offered by an open norm as operative in Japan and the

US that allows for the reproduction, adaptation and communication to the public of copyrighted works for a wide range of potential acts is optimal. Not least because of the time it takes to introduce new exceptions which requires users of copyright works to wait significant periods of time during which research and economy-contributing acts remain unlawful. Japan's approach to technological protection measures, which allows copyright users to circumvent TPMs unliterally where it prevents AI is the most innovation-friendly of the jurisdictions analysed. In terms of contract override, authentic interpretation as practiced in Japan, does not confer the legal certainty of an explicit provision in statute as is in Europe. Thus, combining an open norm while allowing for unilateral circumvention of digital locks as is the case Japan, with an explicit codified provision rendering contractual override of the exception null and void is optimal.

3. Supplementing a text and data mining exception with a novel reverse engineering provision that provides access to the software and model parameters belonging to players designated by the authorities as dominant in a market is recommended. This will support innovation and vitality in digital markets characterised by their lack of competition.

4. It is proposed that both these provisions should be introduced in international treaties. Firstly, the international nature of the AI oligopoly speaks to an international solution to ensure a consistency of approach. Secondly, the global importance of AI as a general purpose technology demands that countries at all stages of development should have the potential to benefit from it. The lack of resource, institutional capacity and the underdeveloped regulatory infrastructures, common in parts of the Global South, underscore the desirability of incorporating such provisions into international treaties. Furthermore, an international framework would immediately yield benefits for researchers, businesses and citizens in jurisdictions where the treaties are viewed as self-executing, obviating the need to wait for enabling legislation at the domestic level.

Contribution to Knowledge

1. *Across chapters two and three, comparative analysis of limitations and exceptions that support AI in four key jurisdictions, including in chapter three an in-depth exploration of a novel open norm that was introduced into Japanese copyright law in 2018.*

2. *An examination of the extent of protection afforded machine learning models by current copyright software law.* Addressing important gaps in the literature, this is the first cross-jurisdictional and in-depth study undertaken that applies copyright doctrines to standard AI model production workflows. By focussing in chapter five on the extent to which exclusive rights protect trained AI models, it informs both policy discussions within copyright law and recent developments in trade law concerning software and algorithm disciplines.

3. *Informed by patent law and growing EU algorithmic intervention, chapter six proposes an AI-specific reverse engineering provision designed to support technology transfer, competition and the development agenda.* The emergence of superstar firms, AI superstates and AI nationalism is a prominent feature of the early 21st century. AI technologies are further entrenching the excessive geographical and market concentration of these companies. Without radical policy interventions, ideally at the international level, this trend will exacerbate the large gap between AI incumbents who often preside over winner-takes-all markets and everyone else – a new paradigm which transcends the established Global North-South divide.

4. *An analysis of how copyright applies to a wide array of machine learning models.* The academic literature overwhelmingly focuses on neural network models, and one relatively small subset of the these – namely generative AI models. Based on a wide reading of the technical literature, this study draws conclusions on an array of models such as support vector machines (SVMs), linear and logistic regression models, decision trees, random forests, including models that require no training such as K-Nearest Neighbours (KNN).

5. *The proposal of a framework for courts to utilise when determining whether new exclusive rights inhere in a trained AI model* (Annex 1). Whereas copyright will subsist in some, if not many of the foundational routines that are made available in AI toolkits, whether exclusive rights inhere in the trained model is more open to question. This thesis proposes a framework to help the courts follow appropriate lines of inquiry to establish whether and / or where copyright subsists in the trained ML model.

6. This thesis sets out a *holistic legal analysis of AI* that considers the role that both limitations and exceptions to copyright and exclusive rights play in shaping innovation. Furthermore, it proposes a taxonomy anchored in legal doctrine for the evaluating the copyright status of computer-generated works.

Summation

The rise of superstar multinationals and AI superstar states confined to a handful of territories in the Global North represents an economic challenge of historically unparalleled proportions. This new era demands a different approach to market regulation and requires a markedly different kind of analysis and response. Multiple policy levers will be required to address extreme levels of market concentration and power. In terms of the role of copyright law, it is posited a novel AI software exception, giving access to not only the valuable public domain elements of a machine learning model but, where necessary, also the protected software code, can act as one of a number of policy interventions to support more balanced and equitable data-driven innovation around the globe.

Most notably in the European Union, the authorities have only just started to acknowledge and develop new forms of law to deal with the market-destructive forces that information technologies have unleashed.[1239] Nonetheless, AI, as a technology so often reliant on the copyright works of others, evinces that the worth of intellectual property law may also be effectively harnessed to maximise consumer welfare. The tendency for information technologies to create natural monopolies and induce severe inequalities of income requires not just tinkering at the edges but new thinking. In this context, it is posited that through the exercise of L&Es, such as TDM exceptions and the novel AI software provision proposed herein, copyright as a prime regulator of information and non-personal data can bolster competition interventions in the public interest.

[1239] Supra 1188.

Bibliography

Case Law

Court of Justice of the European Union

Airfield NV and Canal Digitaal BV v Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (Sabam) and Airfield NV v Agicoa Belgium BVBA (C-431/09 and C-432/09) [2011] ECR I-09363

Art & Allposters International BV v Stichting Pictoright (C-419/13) [2014] ECLI:EU:C: 2014:2214

Bezpečnostní softwarová asociace - Svaz softwarové ochrany v Ministerstvo kultury (C-393/09) [2010] ECR I-13971 ("BSA")

Blue Air Aviation SA v UCMR – ADA Asociaţia pentru Drepturi de Autor a Compozitorilor and Uniunea Producătorilor de Fonograme din România (UPFR) v Societatea Naţională de Transport Feroviar de Călători (SNTFC) 'CFR Călători' SA (C-775/21 and C-826/21) [2023] ECLI:EU:C:2023:307

British Horse Racing Board and Others v William Hill Organization Ltd (C-203/02) [2004] ECR I-10415

Cofemel – Sociedade de Vestuário SA v G-Star Raw CV (C-683/17) [2011] ECLI:EU:C: 2011:648

Fixtures Marketing Ltd v Organismos prognostikon agonon podosfairou AE (Case C-444/02) [2004] ECR I-10549 ("OPAP")

Fixtures Marketing Ltd v Oy Veikkaus Ab (C-46/02) [2004] ECR I-10396

Fixtures Marketing Ltd v Svenska Spel AB (C-338/02) [2004] ECR I-10497

Football Association Premier League Ltd and Others v QC Leisure and Others and Karen Murphy v Media Protection Services Ltd (C-403/08 and C-429/08) [2011] ECR I-9083 ("FAPL")

Football Dataco Ltd and Others v Yahoo! UK Ltd and Others (C-604/10) [2012] ECLI:EU:C: 2012:115

Freistaat Bayern v Verlag Erstbauer GmbH (C-490/14) [2015] ECLI:EU:C:2015:735

Funke Medien NRW GmbH v Bundesrepublik Deutschland I (C469/17) [2019] ECLI:EU:C: 2019:623

Infopaq International A/S v Danske Dagblades Forening (C-5/08) [2009] ECR I-06569 ("Infopaq I")

ITV Broadcasting Ltd and Others v TVCatchUp Ltd (C-607/11) [2013] ECLI:EU:C:2013:147

Land Nordrhein-Westfalen v Dirk Renckhoff (C-161/17) [2018] ECLI:EU:C:2018:634

Levola Hengelo BV v Smilde Foods BV (C-310/17) [2018] ECLI:EU:C:2018:618

Mediakabel BV v Commissariaat voor de Media (C-89/04) [2005] ECR I-4909

Ochranný svaz autorský pro práva k dílům hudebním o.s. v Léčebné lázně Mariánské Lázně a.s. (C-351/12 28) [2014] ECLI:EU:C:2014:110

Eva-Maria Painer v Standard Verlags GmbH and Others (C-145/10) [2012] ECR I-12533

Pelham GmbH and Others v Ralf Hütter and Florian Schneider-Esleben (C-476/17) [2019] ECLI:EU:C:2019:624

Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others (C-360/13) [2014] ECLI:EU:C:2014:1195

Reha Training Gesellschaft für Sport- und Unfallrehabilitation mbH v Gesellschaft für musikalische Aufführungs- und mechanische Vervielfältigungsrechte eV (GEMA) (C-117/15) [2016] ECLI:EU:C: 2016:379

Ryanair Ltd v PR Aviation BN (C-30/12) [2015] ECLI:EU:C:2015:10

SAS Institute Inc v World Programming Ltd Case (C-406/10) [2012] ECLI:EU:C:2012:259

SI and Brompton Bicycle Ltd v Chedech / Get2Get (C-833/18) [2020] ECLI:EU:C:2020:461

Sociedad General de Auditores y Editores de España (SGAE) v Rafael Hoteles SL (C-306/05) [2006] ECLI:EU:C:2006:764

Società Consortile Fonografici (SCF) v Marco Del Corso (C-135/10) [2011] ECLI:EU:C: 2011:431

Marc Soulier and Sara Doke v Premier ministre and Ministre de la Culture et de la Communication (C-301/15) [2016] ECLI:EU:C:2016:878

Vereniging Openbare Bibliotheken v Stichting Leenrecht (C-174/15) [2016] ECLI:EU:C:2016:856

Japan

Contact Angle Program Case. 接触角計算プログラム事件 Intellectual Property High Court, Heisei 28/4/27, April 27, 2016, Hanji 2321/85

Cybers Case. サイボーズ事件 Tokyo District Court, Heisei 14 /9/5, September 14, 2002, Hanji 1811/127

*Disk Publisher Control Case.*ディスクパブリッシャーソフト事件 Tokyo District Court, Heisei 24/12/18, December 18, 2012, WA 5771

Manufacturing Program Case. JR-CAD プログラム事件 Tokyo District Court, Heisei 13, 2001 WA 17306

Mobile Phone Contents Case. 携帯電話コンテンツ事件 Intellectual Property High Court, Heisei 23/2/28 February 23, 2011, Heisei 22 NE 10051

National Space Development Agency of Japan Program Case. 宇宙開発事業団プログラム事件 Intellectual Property High Court, Heisei 18/12/26, December 26, 2006, Hanji 2019/92

New Masuda Ashi Case. NEW 増田足事件 Tokyo District Court, Heisei 23/1/28, January 23, 2011, Hanji 2133/114

Sekisan Kun Case. 積算くん事件 Osaka District Court, Heisei 10, 1998, WA 13577

System Science ① Case. システムサイエンス①事件 Tokyo High Court, Heisei 1/6/20, June 20, 1989, Hanji 1322/138

Trainline Design Program Case. 電車設計用プログラム事件 Tokyo District Court, Heisei 15/1/31 January 31, 2003, Hanji 1820/127

UK

Anacon Corp Ltd v Environmental Research Technology Ltd [1994] FSR 659 (Ch)

Boosey v Whight [1900] 1 122 (Ch)

Cantor Fitzgerald International v Tradition (UK)Ltd [2000] RPC 95 (Ch)

Costa v Dissociadid Ltd [2022] 1934 (EWHC IPEC)

Cummins v Bond [1927] 1 167 (Ch)

Donoghue v Allied Newspapers Ltd [1938] 106 (Ch)

England And Wales Cricket Board Ltd & Anor v Tixdaq Ltd & Anor [2016] EWHC 575 (Ch)

Express Newspapers plc v Liverpool Daily Post & Echo [1985] 1 WLR 1089 (Ch)

Exxon Corp v Exxon Insurance Consultants International Ltd [1982] RPC 69 (CA)

Getty Images v Stability AI [2025] EWHC 2863 (Ch)

Jefferys v Boosey [1854] IV 815/869 (HLC)

Jockey Club v Rahim (22 July 1983) (Ch)

John Richardson Computers Ltd v Flanders [1993] FSR 497 (Ch)

Ibcos Computers v Barclays Mercantile Highland Finance [1994] FSR 275 (Ch)

Kirk v Fleming (1928-1935) MacCC 44 (Ch)

Navitaire Inc v Easyjet Airline Company and Bulletproof Technologies Inc [2004] EWHC 1725 (Ch)

Noah v Shuba [1991] FSR 14 (Ch)

Nova Productions Limited v Mazooma Games Limited & Others [2007] EWCA Civ 219 (CA)

Pike v Nicholas [1869] LR5 Ch App 251 (CA)

THJ Systems Limited & Anor v Daniel Sheridan & Anor [2023] EWCA Civ 1354 (CA)

United States

Andersen et al v Stability AI et al 3:23-cv-00201 (ND Cal 2023)

Andrien v So. Ocean Cty Chamber of Commerce 927 F 2d 132 (3d Cir 1991)

Apple Computer Inc v Microsoft Corp 35 F 3d 1435 (9th Cir 1994)

Atari Games Corp v Nintendo of America Inc 975 F 2d 832 (Fed Cir 1992)

Atari Games Corp v Nintendo of America Inc WL 214886 (ND Cal 1993)

Atari Inc v North American Philips Consumer Electronics Corp 672 F 2d 607 (7th Cir 1982)

Authors Guild v Google Inc No 05 Civ 8136 (SDNY 2005)

Authors Guild Inc v Google Inc, No 05-CV-8136-DC (SDNY 2009) (Amended Settlement Agreement)

Authors Guild v Google Inc No 05 Civ 8136 (DC) 770 F Supp 2d 666 (SDNY 2011)

Authors Guild Inc v Google Inc 954 F Supp 2d 282 (SDNY 2013)

Authors Guild v HathiTrust 755 F 3d 87 (2d Cir 2014)

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

Baker v Selden 101 US 99 (1880)

Bartz v Anthropic PBC No 3:24-cv-05417-WHA, 2025 WL (ND Cal 2025)

Bowers v Baystate Techs Inc 320 F 3d 1317, 1326 (Fed Cir 2003)

Burk v Johnson 146 F 209 213 (8th Cir 1906)

Cain v Universal Pictures Co 47 F Supp 1013 (9th Cir 1942)

Campbell v Acuff-Rose 510 US 569 (S Ct 1994)

Cariou v Prince 714 F 3d 694 (2d Cir 2013)

CCC Information Inc v Maclean Hunter Market Reports Inc 44 F 3d 61 (2d Cir 1994)

CDN Inc v Kapes 197 F3d 1256 (9th Cir 1999)

Chamberlain Grp Inc v Skylink Techs Inc 381 F 3d 1178 1203 (Fed Cir 2004)

Computer Associates International Inc v Altai Inc 982 F 2d 693 (2d Cir 1992)

Computer Entertainment v Connectix (9th Cir 2000)

Design Data Corporation v Unigate Enterprise Inc 847 F 3d 1169 (9th Cir 2017)

J Doe 1 et al v GitHub, Microsoft Corp, OpenAI et al 4:2022cv06823 (ND Cal 2022)

Getty Images (US) Inc v Stability AI Inc 1:23-cv-00135 (D Del 2023)

Google LLC v Oracle Inc 750 F 3d 1339 (Fed Cir 2014)

Golan v Holder 565 US 302 (S Ct 2012)

Feist Publications Inc v Rural Tel Serv Co 499 US 340 (S Ct 1991)

Naruto v Slater 888 F 3d 418 (9th Cir 2018)

Field v Google Inc 412 F Supp 2d 1106 (D Nev 2006)

Forest Park Pictures v Universal Television Network Inc 683 F 3d 424 (2d Cir 2012)

Fox Broad Co v Dish Network LLC 160 F Supp 3d 1139 1178–79 (CD Cal 2015)

Google LLC v Oracle Am Inc 141 1183 (S Ct 2021)

Gottschalk v Benson 409 US 63 (S Ct 1972)

Harper & Row v Nation Enterprises 471 US 539 (S Ct 1985)

Kadrey et al v Meta Platforms Inc No 23-cv-03417-VC (N D Cal 2025)

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

Lexmark Int'l Inc v Static Control Components Inc 387 F 3d 522 (6th Cir 2004)

Lindsay v The Wrecked and Abandoned Vessel RMS Titanic 52 USPQ 2d 1609 (SDNY 1999)

Lotus Development Corporation v Borland International Inc 49 F 3d (1st Cir 1995)

Lowry's Reports v Legg Mason Inc 186 F Supp 2d 592, 593, 594-55 (D Md 2002)

MDY Indus LLC v Blizzard Entm't Inc 629 F 3d 928 (9th Cir 2010)

McGraw-Hill Cos Inc v Google Inc No 05 Civ 8881 (SDNY 2005)

New York Times v Microsoft No. 1:23-cv-11195 (SDNY 2023)

Nichols v Universal Pictures Co 45 F 2d 119 (2d Cir 1930)

NXIVM Corp v The Ross Institute, 364 F 3d 471 (2d Cir 2004)

Parker v Flook 43 US 584 (S Ct 1978)

Plains Cotton Co-op v Goodpasture Computer 807 F 2d 1256 (5th Cir 1987)

Rearden LLC v Walt Disney Co No. 17-cv-04006-JST (ND Cal 2018)

Ritchie v Williams 395 F 3d 283 (6th Cir 2005)

Sega Enterprises Ltd v Accolade Inc 977 F 2d 1510 (9th Cir 1992)

Seltzer v Sunbrock, 22 F. Supp 621 (SD Cal 1938)

Sony Computer Entertainment v Connectix 203 F 3d 596 (9th Cir 2000)

Sony Corp of America v Universal City Studios Inc 464 US 417 (S Ct 1984)

Stephen Thaler v Shira Perlmutter Register of Copyrights and Director of the United States Copyright Office, et al. 22-1564 (BAH) (DC Cir 2023)

Tilghman v Proctor 102 US 707 (S Ct 1881)

Torah Soft Ltd v Drosnin 136 F Supp 2d 276 (SDNY 2001)

Trade Mark Cases 100 US 82 (S Ct 1879)

A V ex rel Vanderhye v iParadigms LLC 62 F 3d 630 (4th Cir 2009)

Universal City Studios v Corley (formerly *Reimerdes*) 273 F 3d 429 (2d Cir 2001)

Walker v University Books 602 F 2d 859 (9th Cir 1979)

Wrench LLC v Taco Bell Corp 256 F 3d 446 (6th Cir 2001)

X Corp v Bright Data Ltd No. C 23-03698 WHA (ND Cal 2024)

Whelan Associates, Inc v Jaslow Dental Laboratory Inc 797 F 2d 1222 (3d Cir 1986)

White-Smith Music Publishing Company v Apollo Company, 209 US 1 (S Ct 1908)

Others:

China: *Beijing Feilin Law Firm v Baidu Corporation,* Min Chu, 2019*,* Beijing Internet Court (239)

Li v Liu, Min Chu, 2023, Beijing Internet Court (11279)

Shenzhen Tencent Computer System Co Ltd v Shanghai Yingxun Technology Co Ltd, Minchu, 2019, Nanshan District People's Court (14010)

Germany: *LAION v Robert Kneschke,* Hamburg District Court 310 O.22723 (September 27 2024)

GEMA v OpenAI, Munich Regional Court 42 O 14139/24 (November 11 2025)

Netherlands: *Zonen Endstra v Nieuw Amsterdam* (2008) ECLI:NL:HR:2008: BC2153, HR 30.05.2008, NJ 2008

Conference Proceedings

Batina L and others, 'Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel' (USENIX Security Symposium, Santa Clara, August 2019)

Behrouz A and others, 'Fast Optimization of Weighted Sparse Decision Trees for Use in Optimal Treatment Regimes and Optimal Policy Design' (Advances in Interpretable Machine Learning AIMLAI, Atlanta, October 2022)

Ciani J, 'Learning From Monkeys: Authorship Issues Arising From AI Technology' (Advances in Interpretable Machine Learning AIMLAI, Atlanta, October 2022)

Davis R, 'Intellectual Property and Software - the Assumptions Are Broken.' (World Intellectual Property Organisation, Stanford University, March 1991)

Filippova E, 'Empirical Evidence and Economic Implications of Blockchain as a General Purpose Technology' (IEEE Technology & Engineering Management Conference (TEMSCON), Atlanta, June 2019)

Goldblum M and others, 'Adversarial Attacks on Machine Learning Systems for High-Frequency Trading' (ACM International Conference on AI in Finance, online, November 2021)

Goodfellow I and others, 'Generative Adversarial Nets' (Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, December 2014)

Gunderson JP and Gunderson LF, ‘Intelligence≠ Autonomy≠capability’ (Performance Metrics for Intelligent Systems. Gaithersberg, August 2004)

Hannak A and others, ‘Measuring Price Discrimination and Steering on E-commerce Web Sites’ (Internet Measurement Conference, Vancouver, November 2014)

Hua W, Zhang Z and Suh GE, ‘Reverse Engineering Convolutional Neural Networks Through Side-channel Information Leaks’ (55th ACM/ESDA/IEEE Design Automation Conference (DAC), San Francisco, June 2018)

Innes M and others, ‘On Machine Learning and Programming Languages’ (Proceedings of SysML Conference – Association for Computing Machinery (ACM), Stanford University, February 2018)

Irion K, ‘Algorithms Off-Limits?’ (Proceedings of 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT), Seoul, June 2022)

Jaszi P, ‘Public Interest Exceptions in Copyright: A Comparative and International Perspective’ (Correcting Course: Rebalancing Copyright for Libraries in the National and International Arenas, New York, May 2005)

Korinek A and Ng DX, ‘The macroeconomics of superstars’ (Proceedings of the 5th International Monetary Fund Statistical Forum, Washington, November 2017)

Lavesson N and Davidsson P, ‘Quantifying the Impact of Learning Algorithm Parameter Tuning.’ (Association for the Advancement of Artificial Intelligence Conference, Boston, July 2006)

Lee C H and others, ‘Calculating Feature Weights in Naive Bayes with Kullback-Leibler Measure.’ (IEEE 11th International Conference on Data Mining, Vancouver, December 2011)

Lerner A and Alonso G, ‘Data Flow Architectures for Data Processing on Modern Hardware’ (IEEE International Conference on Data Engineering, Utrecht, May 2024)

Li J and Lewis HW, ‘Fuzzy Clustering Algorithms – Review of the Applications’ (IEEE International Conference on Smart Cloud, New York, November 2016)

Mantovani RG and others, ‘Hyper-parameter Tuning of a Decision Tree Induction Algorithm’ (5th Brazilian Conference on Intelligent System – BRACIS, Recife, October 2016)

Mladenić D and others, ‘Feature Selection Using Linear Classifier Weights: Interaction with Classification Models’ (27th ACM/SIGIR International Symposium on Information Retrieval, Sheffield, October 2004)

Mohamed AR, Hinton G and Dahl G, ‘Deep Belief Networks for Phone Recognition.’ (Neural Information Processing Systems (NIPS) Workshop on Deep Learning for Speech Recognition and Related Applications, Vancouver, December 2009)

Pezeshkpour P and others,‘An Empirical Comparison of Instance Attribution Methods for NLP’ (Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, online, June 2021)

Raina R and others, ‘Large-Scale Deep Unsupervised Learning Using Graphics Processors’ (The 26th Annual International Conference on Machine Learning Held in Conjunction With the 2007 International Conference on Inductive Logic Programming, Montreal, June 2009)

Reith RN and others, ‘Efficiently Stealing Your Machine Learning Models’ (Workshop on Privacy in the Electronic Society, London, November 2019)

Shi S and others, ‘Benchmarking State-of-the-Art Deep Learning Software Tools’ (7th International Conference on Cloud Computing and Big Data (CCBD), Macau, November 2016)

Shokri R and others, ‘Membership Inference Attacks Against Machine Learning Models’ (IEEE Symposium on Security and Privacy, San Jose, May 2017)

Singh J and others, ‘Responsibility & Machine Learning: Part of a Process’ (Microsoft Cloud Computing Research Centre 3rd Annual Symposium, Cambridge, September 2016)

Smilkov D and others, ‘Tensorflow.JS: Machine Learning for the Web and Beyond’ (Proceedings of the 2nd Systems Modelling Language Conference, Stanford University, March – April 2019)

Song C and others, ‘Machine Learning Models That Remember Too Much’ (ACM SIGSAC Conference on Computer and Communications Security, Dallas, October – November 2017)

Tramèr F and others, ‘Stealing Machine Learning Models via Prediction APIs’ (25th USENIX Security Symposium, Austin, August 2016)

Vardi MY, ‘What is an Algorithm?’ (14th International Congress of Logic, Methodology and Philosophy of Science, Nancy, July 2011)

Wang B and Gong NZ, ‘Stealing Hyperparameters in Machine Learning’ (IEEE Symposium on Security and Privacy, San Francisco, May 2018)

Westkamp G, ‘EU Database Protection for Information Uses Under an Intellectual Property Scheme: Has the Time Arrived for a Flexible Assessment of the European Database Directive?’ (Fordham International Intellectual Property Conference, New York, April 2003)

Governmental Publications

Japan

ＡＩと著作権に関する考え方について「文化審議会著作権分科会２０２４」| ‘On the Approach to AI and Copyright’ (Copyright Subcommittee of the Council for Cultural Affairs 2024)

Society 5.0 「内閣府２０１５」| ‘Society 5.0’ (Cabinet Office 2015)

新たな知財制度上の課題に関する研究会 報告書 | ‘Report from the Research Group on New Issues in IP Law’ (Ministry of Economy Trade and Industry Feb 2022) Annex 2, 36-45.

知的財産推進計画 2007「知的財産戦略本部２００７」| ‘Intellectual Property Promotion Plan 2007’ (IPSHQ 2007)

知的財産推進計画 2015「知的財産戦略本部２０１５」| ‘Intellectual Property Promotion Plan 2015’ (IPSHQ 2015)

著作権法における権利制限規定の柔軟性が及ぼす 効果と影響等に関する調査研究「青山社中株式会社 ２０１７」 | ‘Research Study on the Effects and Impacts of Flexible Copyright Exceptions’ (Aoyama Shachū Co Ltd, March 2017)

著作権法の一部を改正する法律 概要説明資料「文化庁２００９」| ‘Overview of the Act for Partial Amendment of the Copyright Act’ (Agency for Cultural Affairs 2009)

著作権法の一部を改正する法律 概要説明資料「文化庁２０１８」| ‘An Overview of the Act for Partial Amendment of the Copyright Act’ (Agency for Cultural Affairs 2018)

デジタル・ネット時代における知財制度の在り方について <検討の視点>「知的財産戦略本部２００７」| ‘On the Ideal Form of Intellectual Property Systems in a Digital and Networked Era - Perspectives for Consideration’ (IPSHQ 2007)

デジタル・ネット時代における知財制度の在り方について「知的財産戦略本部２００８」| ‘On the Ideal Form of Intellectual Property Systems in a Digital and Networked Era’ **(**IPSHQ 2008)

日本再興戦略　改訂 2014 | ‘Japan Revitalisation Strategy 2014’ (Cabinet Office 24/6/2014)

文化審議会著作権分科会 報告書　平成２３年 | ‘Report of the Copyright Subcommittee of the Council for Cultural Affairs’ (2011)

文化審議会著作権分科会 報告書　平成２９年 | ‘Report of the Copyright Subcommittee of the Council for Cultural Affairs’ (2017)

United Kingdom

Exceptions to Copyright: Research (Intellectual Property Office 2014)
Intellectual Property and Innovation: White Paper Part III (United Kingdom 1986)

Hargreaves I, ‘Digital Opportunity: A Review of Intellectual Property and Growth. An Independent Report by Professor Ian Hargreaves’ (2011)

Kretschmer M and others, ‘The Relationship Between Copyright and Contract Law (Strategic Advisory Board for Intellectual Property Policy’ (SABIP 2010)

‘Legal Protection for Software-Related Innovation. A Report of the Public Forum Held at the United Kingdom Patent Office on 19 October 1994’ (Patent Office 1994)

Department of Trade, *Reform of the Law Relating to Copyright, Designs and Performers’ Protection: A Consultative Document* (Cmd 8302, 1981)

Whitford Committee, *Copyright and Designs Law: Report of the Committee to Consider the Law on Copyright and Designs* (Whitford Committee 1977)

United States

Office of Technology Assessment, 'Computer Software and Intellectual Property – Background Paper' (Office of Technology Assessment 1990)

Fagnan K and others, 'Data and Models: A Framework for Advancing AI in Science' (USDOE Office of Science 2019)

Library of Congress, 'Final Report of the National Commission on New Technological Uses of Copyrighted Works' (CONTU) (Library of Congress, 1978)

Office of Technology Assessment, 'Finding a Balance: Computer Software, Intellectual Property, and the Challenge of Technological Change' (Office of Technology Assessment 1992)

Office of Technology Assessment, 'Intellectual Property Rights in an Age of Electronics and Information United States. (Office of Technology Assessment 1986)

US International Trade Commission, 'Potential Impact on the US Economy and Industries of the GATT Uruguay Round Agreements Investigation No. 332-353' Volume I (US International Trade Commission 1994)

Other

Bundeskartellamt, 'The Market Power of Platforms and Networks' (Bundeskartellamt 2016)

Legislation and Legal Counsel, 'Uses of Copyrighted Materials for Machine Learning' (Government of Israel 2022)

Journal Articles and Working Papers

Aaronson SA and Leblond P, 'Another Digital Divide: The Rise of Data Realms and Its Implications for the WTO' (2018) 21 J Intl Econ L 245

Akl A and others, ‘Optimizing Deep Neural Networks Hyperparameter Positions and Values’ (2019) 37 Journal of Intelligent & Fuzzy Systems 6665

Alderucci D, ‘Copyrights Protection for Patents: Some Surprising Implications for Artificial Intelligence’ (2020) 123 WVL 791

Allamanis M and others, ‘A Survey of Machine Learning for Big Code and Naturalness’ (2018) 51 Association for Computing Machinery Computing Surveys (CSUR) 1

Amir R, Evstigneev I and Gama A, ‘Oligopoly With Network Effects: Firm-specific Versus Single Network’ (2021) 71 Economic Theory 1203

Amsden AH, ‘Diffusion of Development: The Late-industrializing Model and Greater East Asia’ (1991) 81 The American Economic Review 282

Ard B, ‘Taking Access Seriously’ (2020) 8 Tex A & M Law Review 225

Asay CD, ‘Intellectual Property Law Hybridization’ (2016) 87 UCLR 65

Asay CD, Sloan A and Sobczak D, ‘Is Transformative Use Eating the World’ (2020) 61 BCL Rev 905

Ateniese G and others, ‘Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers’ (2015) 10 International Journal of Security and Networks 137

Autor D and others, ‘The Fall of the Labor Share and the Rise of Superstar Firms’ (2020) 135 The Quarterly Journal of Economics 645

Averitt NW and Lande RH, ‘Consumer Sovereignty: a Unified Theory of Antitrust and Consumer Protection Law’ (1996) 65 ALJ 713

Avila Pinto R, ‘Digital Sovereignty or Digital Colonialism’ (2018) 27 SUR International Journal on Human Rights 15

Azmeh S, Foster C and Echavarri J, ‘The International Trade Regime and the Quest for Free Digital Trade’ (2020) 22 International Studies Review 671

Band J, 'The Protectability of Application Program Interfaces: Oracle America V. Google' (2015) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2186628> accessed 21 May 2026

Banzhaf III JF, 'When a Computer Needs a Lawyer' (1966) 71 Dick LR 240

Barkan DM, 'Software Litigation in the Year 2000: the Effect of Object-oriented Design Methodologies on Traditional Software Jurisprudence' (1992) 7 High Tech Law Journal 315

Barnes DW, 'Congestible Intellectual Property and Impure Public Goods' (2010) 9 Northwestern Journal of Technology and Intellectual Property XVII

Barocas S and Selbst AD, 'Big Data's Disparate Impact' (2016) 104 Cal L Rev 671

Barraclough E,'Beware the Open Source Software Risks' (2008) 176 MIP 20

Batistič S and Van Der Laken P, 'History, Evolution and Future of Big Data and Analytics: A Bibliometric Analysis of Its Relationship to Performance in Organizations' (2019) 30 British Journal of Management 229

Bedre-Defolie Ö and Nitsche R, 'When Do Markets Tip? An Overview and Some Insights for Policy' (2020) 11 JECL & Pract 610

Beldiman D, 'Utilitarian Information Works – Is Originality the Proper Lens' (2010) 14 Marquette Intellectual Property Law Review I

Bessen J, 'Information Technology and Industry Concentration' (2020) 63 JLE 531

Birhane A,'Algorithmic Colonization of Africa' (2020) 17 2 Scripted 389

Black AP, 'Object-Oriented Programming: Some History, and Challenges for the Next Fifty Years' (2013) 231 Information and Computation 3

Bohannan C, 'Copyright Preemption of Contracts' (2007) 67 Mary L Rev 616

Borghi M, 'Chasing Copyright Infringement in the Streaming Landscape' (2011) 42 IIC 3

Borghi M and Karapapa S,‘Contractual Restrictions on Lawful Use of Information: Sole-Source Databases Protected by the Back Door?’ (2015) 37 EIPR 505

Borghi M and Karapapa S, ‘Non-Display Uses of Copyright Works: Google Books and Beyond’ (2011) 1 QMJIP 21

Bown C, ‘How the United States Marched the Semiconductor Industry Into Its Trade War With China’ (2020) East Asian Economic Review 24 4

Boyle J, ‘Fencing Off Ideas: Enclosure & the Disappearance of the Public Domain’ (2002) 131 Daedalus 13

Braga A and Logan RK, ‘AI and the Singularity: A Fallacy or a Great Opportunity?’ (2019) 10 Information 2

Bresnahan TF and Trajtenberg M, ‘General Purpose Technologies ‘Engines of Growth’?’ (1995) 65 Journal of Econometrics 83

Breyer S, ‘The Uneasy Case for Copyright: A Study of Copyright in Books, Photocopies, and Computer Programs’ (1970) 84 HLR 281

Bridy A, ‘Coding Creativity: Copyright and the Artificially Intelligent Author’ (2012) 5 Stanford Technology Law Review 1

Bridy A,‘The Evolution of Authorship: Work Made by Code’ (2015) 39 Columbia Journal of Law & the Arts 395

Brkan M, ‘Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond’ (2019) 27 IJL & IT 91

Buitinck L and others, ‘API Design for Machine Learning Software: Experiences From the Scikit-learn Project’ (2013) < https://arxiv.org/pdf/1309.0238> accessed 21 May 2026

Burk D, ‘Algorithmic Fair Use’ (2019) 86 University Chicago Law Review 283

Butler T, ‘Can a Computer Be an Author - Copyright Aspects of Artificial Intelligence’ (1982) 4 University of California Law San Francisco Communications & Entertainment Law Journal 707

Cabay J and Lambrecht M, 'Remix Prohibited: How Rigid EU Copyright Laws Inhibit Creativity' (2015) 10 JIPLP 359

Carroll MW, 'Copyright and the Progress of Science: Why Text and Data Mining is Lawful' (2019) 53 Uc Davis L Rev 893

Casella B and Formenti L, 'FDI in the Digital Economy: a Shift to Asset-light International Footprints' (2019) 25 Transnational Corporations 1

Castelfranchi C, 'Alan Turing's "Computing Machinery and Intelligence" ' (2013) 32 Topoi 293

Chiou T, 'Copyright Lessons on Machine Learning: What Impact on Algorithmic Art' (2019) 10 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 398

Chisum DS, 'The Patentability of Algorithms' (1985) 47 U Pitt L Rev 959

Clifford RD, 'Intellectual Property in the Era of the Creative Computer Program: Will the True Creator Please Stand Up' (1996) 71 Tul L Rev 1675

Cohen JE and Lemley MA, 'Patent Scope and Innovation in the Software Industry' (2001) 89 Calif L Rev 1

Colombo F and others, 'Algorithmic Composition of Melodies with Deep Recurrent Neural Networks' (2016) < https://arxiv.org/pdf/1606.07251> accessed 21 May 2026

Cortes Gallardo Medina E and others, 'Object Detection, Distributed Cloud Computing and Parallelization Techniques for Autonomous Driving Systems' (2021) 11 Applied Sciences 2925

Craig CI, 'Globalizing User Rights-Talk: On Copyright Limits and Rhetorical Risks' (2017) 33 Am U Int'l L Rev 1

Cronin C, 'Virtual Music Scores, Copyright and the Promotion of a Marginalized Technology' (2004) 28 Columbia Journal of Law & the Arts 1

Cuéllar M-F, 'A Simpler World: On Pruning Risks and Harvesting Fruits in a Orchard of Whispering Algorithms' (2017) 51 University of California Davis Law Review 27

Czako Z and others, ‘AutomaticAI – a Hybrid Approach for Automatic Artificial Intelligence Algorithm Selection and Hyperparameter Tuning’ (2021) 182 Expert Systems with Applications 115

Dachs B and others, ‘Bringing It All Back Home? Backshoring of Manufacturing Activities and the Adoption of Industry 4.0 Technologies’ (2019) 54 6 Journal of World Business 101017

Davidson DM, ‘Protecting Computer Software: a Comprehensive Analysis’ (1983) Arizona St L J 611

Davies CR, ‘An Evolutionary Step in Intellectual Property Rights – Artificial Intelligence and Intellectual Property’ (2011) 27 CLS Rev 601

De Angelis D, ‘Un’analisi Della Ricezione Italiana Delle Eccezioni E Limitazioni Obbligatorie Introdotte Dalla Direttiva 2019/790/ue Nel Campo Delle Attività Digitali Di Insegnamento, Ricerca E Accesso Al Patrimonio Culturale’ (2022) N.4 Il Diritto Di Autore

de Cock Buning M, ‘Autonomous Intelligent Systems as Creative Agents Under the EU Framework for Intellectual Property’ (2016) 7 European Journal of Risk Regulation 310

Deltorn JM, ‘Deep Creations: Intellectual Property and the Automata’ (2017) 4 Frontiers in Digital Humanities 3

Demers J, ‘Sound-Alikes, Law, and Style’ (2014) 83 University of Missouri-Kansas City School of Law 303

Denicola RC, ‘Ex Machina: Copyright Protection for Computer Generated Works’ (2016) 69 Rutgers UL Rev 251

Derclaye E, ‘Infopaq International a/s V Danske Dagblades Forening (C-5/08): Wonderful or Worrisome? The Impact of the ECJ Ruling in Infopaq on UK Copyright Law’ (2010) EIPR 247

Derclaye E, ‘Assessing the Impact and Reception of the Court of Justice of the European Union Case Law on UK Copyright Law: What Does the Future Hold?’ (2014) 240 Rev Intern Dr Auteur 5

Dickenson J and others, ‘Creative Machines: Ownership of Copyright in Content Created by Artificial Intelligence Applications’ (2017) 39 EIPR 457

Dornis T, ‘Artificial Creativity: Emergent Works and the Void in Current IP Doctrine’ (2019) 22 Yale Journal of Law & Technology 1

Dosi G and Stiglitz JE, ‘The Role of Intellectual Property Rights in the Development Process, With Some Lessons from Developed Countries: An Introduction’ (2014) 1 Intellectual Property Rights: Legal and Economic Challenges for Development 1

Drassinower A, ‘Authorship as Public Address: on the Specificity of Copyright Vis-a-Vis Patent and Trademark’ (2008) Michigan State Law Review 199

Drexl J, ‘Designing Competitive Markets for Industrial Data’ (2017) 8 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 257

Drexl J and others, ‘Artificial Intelligence and Intellectual Property Law - Position Statement of the Max Planck Institute for Innovation and Competition of 9 April 2021 on the Current Debate’ (2021) Max Planck Institute for Innovation & Competition Research Paper No. 21-10

Drexl J and others, ‘Technical Aspects of Artificial Intelligence: An Understanding from an Intellectual Property Law Perspective’ (2019) Max Planck Institute for Innovation & Competition Research Paper No. 19-13

Dreyfuss R and Frankel S, ‘From Incentive to Commodity to Asset: How International Law is Reconceptualizing Intellectual Property’ (2014) 36 Mich J Intl L 557

Ducato R and Strowel AM, ‘Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out’ (2021) 43 EIPR 5

Durand C and Milberg W, ‘Intellectual Monopoly in Global Value Chains’ (2020) 27 Review of International Political Economy 404

Dusollier S, ‘Electrifying the Fence: The Legal Protection of Technological Measures for Protecting Copyright’ (1999) 6 EIPR 285

Dutt D and others, ‘Trade Liberalization and Fiscal Stability in Developing Countries: What Does the Evidence Tell Us?’ (2020) 11 Global Policy 375

Earl PE and Mandeville T, ‘The Competitive Process in the Age of the Internet’ (2009) 27 Prometheus 195

Ehinger P, ‘Die Urheberrechtliche Schutzfähigkeit Von Künstlicher Intelligenz Am Beispiel Von Neuronalen Netzen’ (2018) 34 Computer Und Recht 761

Eberle E, ‘The Methodology of Comparative Law’ (2011) 16 Roger Williams University Law Review 1

Elkin-Koren N, ‘The New Frontiers of User Rights’ (2016) 32 Am U Int'l L Rev 1

Elkin-Koren N and Netanel NW, ‘Transplanting Fair Use Across the Globe: A Case Study Testing the Credibility of US Opposition’ (2020) 72 Hastings L J 1121

Englund SR, ‘Idea, Process, or Protected Expression?: Determining the Scope of Copyright Protection of the Structure of Computer Programs’ (1990) 88 Mi L 866

Ennis S and Fletcher A, ‘The Goldilocks Dilemma: When is Digital Regulation “just right”?’ (2020) 11 JECL & Pract 551

Fang Z and others, ‘A Comparative Study of Heterogeneous Ensemble-Learning Techniques for Landslide Susceptibility Mapping’ (2021) 35 International Journal of Geographical Information Science 321

Favale M, ‘The Right of Access in Digital Copyright: Right of the Owner or Right of the User?’ (2012) 15 JWIP 1

Fjelland R, ‘Why General Artificial Intelligence Will Not Be Realized’ (2020) 7 Humanities and Social Sciences Communications 1

Flynn S and others, ‘Implementing User Rights for Research in the Field of Artificial Intelligence: A Call for International Action’ (2020) Intellectual Property and Technology, Law, & Security Program Research Paper Series 2020-12

Flynn S and Palmedo M, ‘The User Rights Database: Measuring the Impact of Copyright Balance’ (2019) Working Papers 42

Foss-Solbrekk K, 'Three Routes to Protecting AI Systems and Their Algorithms Under IP Law: The Good, the Bad and the Ugly' (2021) 16 JIPLP 247

Foster C and Azmeh S, 'Latecomer Economies and National Digital Policy: An Industrial Policy Perspective' (2020) 56 The Journal of Development Studies 1247

Franceschelli G and Musolesi M, 'Copyright in Generative Deep Learning' (2022) 4 Data & Policy E17

Günter Frankenberg, 'Critical, 'Comparisons: Re-thinking Comparative Law' (1985) 26 Harv Int'l LJ 411

Gehlsampath P, 'Regulating the Digital Economy: Are We Moving Towards a 'win-win'or a 'lose-lose' MERIT Working Papers 2018-005

Geiger C, 'The Missing Goal Scorers in the Artificial Intelligence Team: of Big Data, the Fundamental Right to Research and the Failed Text and Data Mining Limitations in the CDSM Directive' (2021) Intellectual Property and Technology, Law, & Security Program Research Paper Series 383

Geiger C and others, 'Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU' (2019) Propiedad Intelectual Y Mercado Único Digital Europeo 27

Geiger C, Gervais D and Senftleben M, 'Three-step Test Revisited: How to Use the Test's Flexibility in National Copyright Law' (2013) 29 Am U Int'l L Rev 581

Geiger C, Griffiths J and Hilty RM, 'Towards a Balanced Interpretation of the "Three-Step Test" in Copyright Law' (2008) EIPR 489

Geiger C and Jütte BJ, 'Copyright as an Access Right: Concretizing Positive Obligations for Rightholders to Ensure the Exercise of User Rights' (2024) 73 GRUR Int 1019

Geller PE, 'A German Approach to Fair Use: Test Cases for Trips Criteria for Copyright Limitations' (2009) 57 J Copyright Soc'y USA 553

Geller PE, 'Beyond the Copyright Crisis: Principles for Change' (2007) 55 J Copyright Soc'y USA 165

Geradin D, ‘What is a Digital Gatekeeper? Which Platforms Should Be Captured by the EC Proposal for a Digital Market Act?’ (2021) <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3788152> accessed 21 May 2026

Gerschewski S, ‘Do Local Firms Benefit from Foreign Direct Investment? An Analysis of Spillover Effects in Developing Countries’ (2013) 9 Asian Social Science 67

Gervais D, ‘Exploring the Interfaces Between Big Data and Intellectual Property Law’ (2019) 10 Journal of Intellectual Property, Information Technology and E-Commerce Law 3

Gervais D, ‘Feist Goes Global: A Comparative Analysis of the Notion of Originality in Copyright Law’ (2001) 49 J Copyright Soc'y USA 949

Gervais D, ‘The Derivative Right, or Why Copyright Law Protects Foxes Better Than Hedgehogs’ (2012) 15 Vanderbilt Journal of Entertainment & Technology Law 785

Gervais D, ‘The Machine as Author’ (2019) 105 Iowa L Rev 2053

Ghafele R and Gibert B, ‘A Counterfactual Impact Analysis of Fair Use Policy on Copyright Related Industries in Singapore’ (2014) 3 Laws 327

Ginsburg JC, ‘Essay: From Having Copies to Experiencing Works: the Development of an Access Right in US Copyright Law’ (2002) 50 J Copyright Soc'y USA 113

Ginsburg JC, ‘Four Reasons and a Paradox: the Manifest Superiority of Copyright Over Sui Generis Protection of Computer Software’ (1994) 94 CLR 2559

Ginsburg JC, ‘The Concept of Authorship in Comparative Copyright Law’ (2002) 52 De Paul L Rev 1063

Ginsburg JC and Budiardjo LA, ‘Authors and Machines’ (2019) 34 Berkeley Tech LJ 343

Givon M and others, ‘Software Piracy: Estimation of Lost Sales and the Impact on Software Diffusion’ (1995) 59 Journal of Marketing 29

Goldstein P, ‘Copyright and Its Substitutes’ (1997) 45 J Copyright Soc'y USA 151

Goldstein P, ‘Fair Use in Context’ (2007) 31 The Columbia Journal of Law & the Arts 433

Gordon E, ‘The Very Idea: Why Copyright Law is an Inappropriate Way to Protect Computer Programs’ (1998) 20 EIPR 10

Grimmelmann J, ‘Copyright for Literate Robots’ (2015) 101 Iowa L Rev 657

Grimmelmann J, ‘There's No Such Thing as a Computer-Authored Work - and It's a Good Thing, too’ (2015) 39 The Columbia Journal of Law & the Arts 403

Grimmelmann J and others, ‘Talkin’ Bout AI Generation: Copyright and the Generative-AI Supply Chain.’ (2025) 72 J Copyright Soc'y USA 251

Guadamuz A, ‘A Scanner Darkly: Copyright Liability and Exceptions in Artificial Intelligence Inputs and Outputs’ (2023) 2 GRUR Int 2024

Hart RJ, ‘Copyright and Computer Generated Works’ (1988) 40 ASLIB Proceedings 6

He T, ‘The sentimental fools and the fictitious authors: rethinking the copyright issues of AI-generated contents in China.’ (2019) APLR 27 2

Hedrick SF, ‘I ‘think’, Therefore I Create: Claiming Copyright in the Outputs of Algorithms’ (2018) 8 New York University Journal of Intellectual Property & Entertainment Law 2

Heim M and Nikolic I, ‘A FRAND Regime for Dominant Digital Platforms’ (2019) 10 Journal of Intellectual Property, Information Technology and E-Commerce Law 38

Henderson P and Lemley MA, ‘The Mirage of Artificial Intelligence Terms of Use Restrictions’ (2024) Princeton University Program in Law & Public Affairs Research Paper No. 2025-04

Henderson P and others, ‘Foundation Models and Fair Use’ (2023) 24 Journal of Machine Learning Research 1

Henderson-Aellers B and Edwards JM, ‘The Object-Oriented Systems Life Cycle’ (1990) 33 Communications of the Association for Computing Machinery 142

Hill RK, ‘What an Algorithm Is’ (2016) 29 Philosophy & Technology 35

Hinton GE, Osindero S and Teh YW, ‘A Fast Learning Algorithm for Deep Belief Nets’ (2006) 18 Neural Computation 1527

Holroyd C, ‘Technological Innovation and Building a ‘super Smart’society: Japan’s Vision of Society 5.0’ (2020) Journal of Asian Public Policy 1

Hu X and others, ‘Neural Network Model Extraction Attacks in Edge Devices by Hearing Architectural Hints’ (2019) < https://arxiv.org/pdf/1903.03916> accessed 20 May 2026

Huet J and Ginsburg JC, ‘Computer Programs in Europe: A Comparative Analysis of the 1991 EC Software Directive’ (1992) 30 The Columbia Journal of Transnational Law 327

Hughes J, ‘Size Matters (or Should) in Copyright Law’ (2005) 74 Fordham L Rev 575

Idei H, AI 生成物に関する知的財産権の現状と課題 ~ Society 5.0 の実現に向けて~「２０１８」68 情報の科学と技術 | ‘The Current Status and Challenges of Intellectual Property Rights Related to AI-Generated Products – Towards the Realisation of Society 5’ (2018) 68 The Journal of Information Science and Technology Association

Idei H, AI 創作の現状と著作権法を中心とする検討課題「２０２０」コピライト 60 13 | ‘Current State of AI-Generated Works and Issues Centred on Copyright Law’ (2020) Copyright 60 13

Idei H, 生成物の著作物性の判断基準とその判断手法に関する一考察「２０１８」パテント 71 | ‘A Study on Criteria and Methods for Determining Copyrightability of Generated Works’ (2018) Patent 71

Idei H, 2020 年、改めてロボット・AI 社会の知的財産制度を考える「２０２０」法の支配 197 | ‘Reconsidering Intellectual Property Systems in a Robots and AI Society’ (2020) Rule of Law 197

Ikemura S, 柔軟な権利制限規定と実務への影響 「２０１８」126 Business Law Journal | ‘Flexible Limitations on Rights and Their Impact on Practice’ (2018) Business Law Journal 126

Inguanez D, ‘A Refined Approach to Originality in EU Copyright Law in Light of the ECJ’s Recent Copyright / Design Cumulation Case Law’ (2020) 51 7 IIC 797

Johnson-Laird A, ‘Neural Networks: The Next Intellectual Property Nightmare?’ (1990) 7 Computer Law 7

Judge EF and Gervais DJ, ‘Of Silos and Constellations: Comparing Notions of Originality in Copyright Law’ (2010) 27 Cardozo Arts & Entertainment Law Journal 10

Jullien B and Sand-Zantman W, ‘The Economics of Platforms: A Theory Guide for Competition Policy’ (2021) 54 Information Economics and Policy 100880

Karjala DS, ‘Copyright, Computer Software, and the New Protectionism’ (1987) Jurimetrics 33

Karjala DS, ‘Japanese Courts Interpret the "Algorithm" Limitation on the Copyright Protection of Computer Programs’ (1991) 31 Jurimetrics 233

Karjala DS, ‘Protection of Computer Programs Under Japanese Copyright Law’ (1986) 8 EIPR 105

Karjala DS, ‘The Protection of Operating Software Under Japanese Copyright Law’ (1988) 29 Jurimetrics 43

Karjala DS, ‘The Relative Roles of Patent and Copyright in the Protection of Computer Programs’ (1998) 17 J Marshall J Computer & Info L 41

Karjala DS and Sugiyama K, ‘Fundamental Concepts in Japanese and American Copyright Law’ (1988) 36 AJCL 613

Karjiker S, ‘Copyright Protection of Computer Programs’ (2016) 133 SALJ 51

Kato M, Ueno T, Okumura K, 柔軟な権利制限規定による著作物の利用拡大と、これからの課題 上 「２０１９」４１ New Business Law | ‘Expansion of the Use of Copyrighted Works through Flexible Limitation Provisions and Future Challenges (Part I)’ (2019) 4 1 New Business Law

Kelsey J, ‘How a TPP-Style E-Commerce Outcome in the WTO Would Endanger the Development Dimension of the GATS Acquis (and Potentially the WTO)’ (2018) 21 JIEL 273

Khan LM, ‘Amazon's Antitrust Paradox’ (2016) 126 YLJ 710

Khan LM, ‘Sources of Tech Platform Power’ (2017) 2 Georgetown Law Technology Review 325

Kim L, ‘The Dynamics of Samsung's Technological Learning in Semiconductors’ (1997) 39 California Management Review 86

Klingler-Vidra R and Wade R, ‘Science and Technology Policies and the Middle-Income Trap: Lessons from Vietnam’ (2020) 56 The Journal of Development Studies 717

Konstantinov AV and Utkin LV, ‘Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines’ (2021) 222 Knowledge-Based Systems 106993

Kop M, ‘AI & Intellectual Property: Towards an Articulated Public Domain’ (2019) 28 Tex Intell Prop LJ 297

Korinek A and Vipra J, ‘Concentrating Intelligence: Scaling Laws and Market Structure in Generative AI’ (2024) Institute for New Economic Thinking. Working Paper 228.

Krizhevsky A, Sutskever I and Hinton GE, ‘ImageNet Classification with Deep Convolutional Neural Networks’ (2012) 25 Advances in Neural Information Processing Systems 1

Lancieri F and Sakowski PM, ‘Competition in Digital Markets: A Review of Expert Reports’ (2021) 26 Stanford Journal of Law, Business & Finance 65

Lashkari D, Bauer A and Boussard J, ‘Information Technology and Returns to Scale’ (2024) 114 American Economic Review 1769

Lee E, ‘Technological Fair Use’ (2009) 83 So Cal LR 797

Lee JY, ‘Artificial Intelligence Cases in China: Feilin v Baidu and Tencent Shenzhen v Shanghai Yingxin’ (2021) 7 China and WTO Review 211

Legrand P, ‘European Legal Systems Are Not Converging’ (1996) 45 International and Comparative Law Quarterly 52

Lehr D and Ohm P, ‘Playing with the Data: What Legal Scholars Should Learn About Machine Learning’ (2017) 51 University of California Davis Law Review 653

Leistner M, ‘The Legal Protection of Telephone Directories Relating to the New Database Maker's Right’ (2000) 31 IIC 95

Lemley MA, ‘Faith-Based Intellectual Property’ (2015) 62 University of California Los Angeles Law Review 1328

Lemley MA, ‘Intellectual Property and Shrinkwrap Licenses’ (1994) 68 So Cal L R 1239

Lemley MA and Casey B, ‘Fair Learning’ (2020) 99 4 Tex L Rev 743

Lemley MA and Samuelson P, ‘Interfaces and Interoperability After Google v Oracle' (2021) 100 Tex L Rev 1

Leppänen L, Tuulonen H and Sirén-Heikel S, ‘Automated Journalism as a Source of and a Diagnostic Device for Bias in Reporting’ (2020) 8 Media and Communication 39

Leval PN, ‘Campbell as Fair Use Blueprint’ (2015) 90 WLR 597

Leval PN, ‘Toward a Fair Use Standard’ (1990) 103 HLR 1105

Levendowski A, ‘How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem’ (2018) 93 WLR 579

Li Q, ‘Copyright protection paradigm in the generative AI era: originality, attribution, and infringement’ (2025) 16 Advances in Social Behavior Research 8

Liashchynskyi P and Liashchynskyi P. ‘Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS’ (2019) <https://arxiv.org/abs/1912.06059> accessed 21 May 2026

Lipton JD, ‘IP's Problem Child: Shifting the Paradigms for Software Protection’ (2006) 58 Hastings LJ 205

Litman J, ‘Copyright and Information Policy’ (1992) 55 L & C Prob 185

Liu J, ‘An Empirical Study of Transformative Use in Copyright Law’ (2019) 22 Stanford Technology Law Review 163

Lunney Jr GS, ‘Copyright Protection for ASIC Gate Configurations: PLDs, Custom and Semicustom Chips’ (1989) 42 SLR 163

Maeng K and others, ‘Network Effects of Connected and Autonomous Vehicles in South Korea: a Consumer Preference Approach’ (2021) 90 Research in Transportation Economics 100998

Malairaja C and Zawdie G, ‘The ‘Black Box’ Syndrome in Technology Transfer and the Challenge of Innovation in Developing Countries. The Case of International Joint Ventures in Malaysia’ (2004) 3 International Journal of Technology Management & Sustainable Development 233

Margoni T and Kretschmer M, ‘A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology’ (2022) 71 GRUR Int 685

Martinez R, ‘Artificial Intelligence: Distinguishing Between Types & Definitions’ (2019) 19 Nevada Law Journal 9

Matulionyte R, ‘Australian Copyright Law Impedes the Development of Artificial Intelligence: What Are the Options?’ (2021) 52 IIC 417

Matulionyte R, ‘Can AI Infringe Moral Rights of Authors and Should We Do Anything About It? An Australian Perspective’ (2023) 15 Law Innovation and Technology 124

Maudes J and others, ‘Random Feature Weights for Decision Tree Ensemble Construction’ (2012) 13 Information Fusion 20

Maxwell W, Bolger K and Zeggane T, ‘Un Regard Transatlantique sur les "Scènes à Faire" et Autres Banalités en Droit D'auteur Propriétés Intellectuelles’ (2009) 30 Propriétés Intellectuelles 35

McCutcheon J, ‘The Vanishing Author in Computer-Generated Works: A Critical Analysis of Recent Australian Case Law’ (2012) 36 MULR 915

Mcintyre S, ‘Trying to Agree on Three Articles of Law: The Idea / Expression Dichotomy in Chinese Copyright Law’ (2010) 1 Cybaris Intellectual Property Law Review 62

Menell PS, ‘An Analysis of the Scope of Copyright Protection for Application Programs’ (1989) 45 SLR 1045

Mezei P, ‘From Leonardo to the Next Rembrandt – the Need for AI-Pessimism in the Age of Algorithms’ (2020) 84 2 Archiv für Medienrecht und Medienwissenschaft 390

Mittelstadt BD and others, ‘The Ethics of Algorithms: Mapping the Debate’ (2016) 3 Big Data & Society 2

Mizuta I, 著作権行政をめぐる最新の動向について「２０１７」コピライト 679 57 | ‘The Newest Developments in Copyright Administration’ (2017) Copyright 679 57

Moon K, ‘Another Nail in the Coffin for Non-literal Software Copyright Infringement?’ (2015) 10 JIPLP 921

Mosca M, ‘On the Origins of the Concept of Natural Monopoly: Economies of Scale and Competition’ (2008) 15 The European Journal of the History of Economic Thought 317

Nakayama N, 著作権法改正の潮流 「２００９」コピライト 578 2 | Trends in Copyright Law Reform (2009) Copyright 578 2

Narkhede MV Bartakke PP and Sutaone MS, ‘A Review on Weight Initialization Strategies for Neural Networks’ (2022) 55 Artificial Intelligence Review 291

Netanel NW, ‘Copyright and a Democratic Civil Society’ (1996) 106 Yale Law Journal 283
Netanel NW. ‘Making Sense of Fair Use’ (2011) 15 Lewis & Clark Law Review 715

Neuhäusler P, ‘The Use of Patents and Informal Appropriation Mechanisms - Differences Between Sectors and Among Companies’ (2012) 32 Technovation 681

Newell A, ‘Response: The Models Are Broken, the Models Are Broken’ (1985) 47 U Pitt L Rev 1023

Newman C and others, ‘Technology Transfers, Foreign Investment and Productivity Spillovers’ (2015) 76 European Economic Review 168

Newman Jo, ‘New Lyrics for an Old Melody: The Idea / Expression Dichotomy in the Computer Age’ (1999) 17 Cardozo Arts & Entertainment Law Journal 691

Newman N, ‘Search, Antitrust, and the Economics of the Control of User Data’ (2014) 31 Yale J Reg 401

Neyshabur B, Sedghi H and Zhang C, ‘What is Being Transferred in Transfer Learning?’ (2020) 33 Advances in Neural Information Processing Systems 512

Nimmer D, Brown E and Frischling GN, ‘The Metamorphosis of Contract into Expand’ (1999) 87 CLR 17

Nuccio M and Guerzoni M, ‘Big Data: Hell or Heaven? Digital Platforms and Market Power in the Data-Driven Economy’ (2019) 23 Competition & Change 312

Oddi AS, ‘An Uneasier Case for Copyright Than for Patent Protection of Computer Programs’ (1993) 72 Neb LR 351

Okediji RL, ‘Legal Innovation in International Intellectual Property Relations: Revisiting Twenty-One Years of the Trips Agreement’ (2014) 36 University of Pennsylvania Journal of International Law 191

Olswang S, ‘Access Right: An Evolutionary Path for Copyright Into the Digital Era?’ (1995) 17 EIPR 215

Otero BG, ‘Machine Learning Models Under the Copyright Microscope: Is EU Copyright Fit for Purpose?’ (2021) 70 GRUR Int 1043

Pagani M, Miric M and El Sawy O, ‘The Octopus Effect: When and Who Platform Companies Acquire’ (2021) LSE Business Review 1

Pallante MA, ‘The Next Great Copyright Act’ (2012) 36 The Columbia Journal of Law & the Arts 315

Panzar JC and Willig RD, ‘Economies of Scope’ (1981) 71 The American Economic Review 268

Pedregosa F and others, ‘Scikit-learn: Machine Learning in Python’ (2011) 12 The Journal of Machine Learning Research 2825

Perel M and Elkin-Koren N, ‘Accountability in Algorithmic Copyright Enforcement’ (2015) 19 Stanford Technology Law Review 473

Phan AV, Le Nguyen M and Bui LT, ‘Feature Weighting and SVM Parameters Optimization Based on Genetic Algorithms for Classification Problems’ (2017) 46 Applied Intelligence 455

Pianykh OS and others, ‘Continuous Learning AI in Radiology: Implementation Principles and Early Applications’ (2020) 297 Radiology 6

Podgorelec V and others, ‘Decision Trees: An Overview and Their Use in Medicine’ (2002) 26 Journal of Medical Systems 445

Posner RA, ‘When is Parody Fair Use?’ (1992) 21 JLS 67

Prause M and Weigand J, ‘Industry 4.0 and Object-oriented Development: Incremental and Architectural Change’ (2016) 11 Journal of Technology Management & Innovation 104

Priel D, ‘Legal Realism, Legal Positivism, and Determinacy’ (2026) 71 Am J Juris 1 17

Probst P, Boulesteix A-L and Bischl B, ‘Tunability: Importance of Hyperparameters of Machine Learning Algorithms’ (2019) 20 Journal of Machine Learning Research 1

Quaedvlieg A, ‘Onafhankelijk, Geordend en Toegankelijk: Het Object Van Het Databankenrecht in De Richtlijn’ (2000) 9 Informatierecht / AMI

Quigley J, ‘Socialist Law and the Civil Law Tradition’ (1989) 37 AJCL 4 781

Ramalho A, ‘Will Robots Rule the (Artistic) World? A Proposed Model for the Legal Status of Creations by Artificial Intelligence Systems’ (2017) 21 1 J Internet L 21 12

Rao DS, ‘Neural Networks: Here, There, and Everywhere - An Examination of Available Intellectual Property Protection for Neural Networks in Europe and the United States’ (1996) 30 Geo Wash J Int'l L & Econ 509

Reichman JH, ‘Legal Hybrids Between the Patent and Copyright Paradigms’ (1994) 94 Columbia Journal of Law & the Arts 2432

Reichman JH and Okediji RL, ‘When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale’ (2012) 96 Minn L Rev 1362

Reichman JH and Samuelson P, ‘Intellectual Property Rights in Data’ (1997) 50 Vand L Rev 49

Ricketson S, ‘The 1992 Horace S Manges Lecture - People or Machines: The Berne Convention and the Changing Concept of Authorship’ (1991) 16 The Columbia Journal of Law & the Arts 1

Rigamonti CP, ‘Deconstructing Moral Rights’ (2006) 47 Harvard Int’l LJ 353

Riley KC, ‘Data Scraping as a Cause of Action: Limiting Use of the CFAA and Trespass in Online Copying Cases’ (2018) 29 Fordham Intellectual Property Media & Entertainment Law Journal 245

Robinson GH, ‘Protection of Intellectual Property in Neural Networks’ (1990) 7 Computer Law 17

Rodrik D, ‘What Do Trade Agreements Really Do?’ (2018) 32 Journal of Economic Perspectives 73

Rosati E, ‘Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and Its Role in the Development of AI Creativity’ (2019) 27 2 APLR 198

Rub GA, ‘Copyright Survives: Rethinking the Copyright-Contract Conflict’ (2017) 103 Va L 1141

Sag M, ‘Copyright and Copy-Reliant Technology’ (2009) 103 NWULR 1607

Sag M, ‘Copyright Safety for Generative AI’ (2023) 61 Hous LR 295

Sag M, ‘Fairness and Fair Use in Generative AI’ (2024) 92 Fordham L Rev 5 1887

Sag M, ‘The New Legal Landscape for Text Mining and Machine Learning’ (2019) 66 Copyright Soc'y USA 291

Sakai M, 著作権法における『創作性』概念に関する一考察「２０１２」7 研究報告電子化知的財産・社会基盤 EIP 1 | ‘A Study on the Concept of “Creativity” in Copyright Law’ (2012) 7 Information Processing Society of Japan’ (EIP) 1

Samuelson P, ‘Allocating Ownership Rights in Computer-Generated Works’ (1985) 47 U Pitt L Rev 1185

Samuelson P, ‘Benson Revisited: The Case Against Patent Protection for Algorithms and Other Computer Program-Related Inventions’ (1990) 39 Emory Law Journal 1025

Samuelson P, ‘Freedom to Tinker’ (2016) 17 Theo Inq L 562

Samuelson P, ‘Functionality and Expression in Computer Programs: Refining the Tests for Software Copyright Infringement’ (2016) 31 Berkeley Tech LJ 1215

Samuelson P, ‘Google Book Search and the future of books in cyberspace.’ (2009) 94 Minn L Rev 1308

Samuelson P, ‘Intellectual Property and the Digital Economy: Why the Anti-Circumvention Regulations Need to Be Revised’ (1999) 14 Berkeley Tech LJ 519

Samuelson P, ‘Legally Speaking: Should Program Algorithms Be Patented’ (1990) 33 Communications of the Association for Computing Machinery 23

Samuelson P, ‘Reconceptualizing Copyright's Merger Doctrine’ (2016) 63 Copyright Soc'y USA 417

Samuelson P, ‘Staking the Boundaries of Software Copyrights in the Shadow of Patents’ (2019) 71 FLR 243

Samuelson P, ‘The Uneasy Case for Software Copyrights Revisited’ (2010) 79 Geo Wash L Rev1746

Samuelson P, ‘Why Copyright Law Excludes Systems and Processes from the Scope of Its Protection’ (2006) 85 Tex L Rev 1921

Samuelson P and others, ‘Manifesto Concerning the Legal Protection of Computer Programs’ (1994) 94 The Columbia Journal of Law & the Arts 2308

Sawada M, 契約による権利制限規定のオーバーライド(2021) コピライト | ‘The Override of L&Es by Contract’ Copyright (2021) 727 61

Scherer MU, ‘Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies’ (2015) 29 Harv JL & Tech 353

Schwartz H, ‘Heuristics or Rules of Thumb’ (2010) Behavioral Finance: Investors, Corporations, and Markets 57

Schwartz HM, ‘Global Secular Stagnation and the Rise of Intellectual Property Monopoly’ (2022) 29 Review of International Political Economy 1448

Schwartz KS, Baim SA and Mayergorz S, ‘Software and Scènes à Faire’ (2009) 190 Copyright World

Schweitzer H and Welker R, ‘Competition Policy for the Digital Era’ (2019) 3 The Antitrust Chronicle 16

Sell SK, ‘Industry Strategies for Intellectual Property and Trade: The Quest for Trips, and Post-Trips Strategies’ (2002) 10 Cardozo International & Comparative Law Review 79

Senftleben M, ‘Compliance of National TDM Rules With International Copyright Law: An Overrated Nonissue?’ (2022) 53 IIC 1477

Senftleben M and others, ‘Towards a European Research Freedom Act: A Reform Agenda for Research Exceptions in the EU Copyright Acquis’ (2025) 56 IIC 1329

Sharir O, Peleg B and Shoham Y, ‘The Cost of Training NLP Models: A Concise Overview’ (2020) < https://arxiv.org/pdf/2004.08900> accessed 21 May 2026

Sobel BLW, ‘Artificial Intelligence's Fair Use Crisis’ (2017) 41 The Columbia Journal of Law & the Arts 45

Stern RH, ‘On Defining the Concept of Infringement of Intellectual Property Rights in Algorithms and Other Abstract Computer-Related Ideas’ (1995) 23 American Intellectual Property Law Association. Quarterly Journal 401

Suchman M and Mertz E, ‘Toward a new legal empiricism: empirical legal studies and new legal realism’ (2010) 6.1 Annual Review of Law and Social Science 555

Surblytė G, ‘Data-Driven Economy and Artificial Intelligence: Emerging Competition Law Issues’ (2017) 67 Wuw 120

Swinson J, ‘Copyright or Patent or Both: An Algorithmic Approach to Computer Software Protection’ (1991) 5 Harv JL & Tech 145

Talib MA and others, ‘A Systematic Literature Review on Hardware Implementation of Artificial Intelligence Algorithms’ (2021) 77 The Journal of Supercomputing 1897

Toosi A and others, ‘A Brief History of AI: How to Prevent Another Winter (A Critical Review)’ (2021) 16 Positron Emission Tomography (PET) Clinics 449

Toth AK, ‘Algorithmic Copyright Enforcement and AI: Issues and Potential Solutions Through the Lens of Text and Data Mining’ (2019) 13 Masaryk University Journal of Law and Technology 361

Turing AM, ‘Mind’ (1950) 59 Mind 433

Ueno T, 人工知能と機械学習をめぐる著作権法上の課題: 日本とヨーロッパにおける近時の動向 (特集 AI がもたらす知的財産法の変容と未来)「２０１９」91 法律時報 33 | ‘Copyright Law Issues Regarding Artificial Intelligence and Machine Learning: Recent Trends in Japan and Europe’ (Special Feature: The Transformation and Future of Intellectual Property Law that AI Brings) (2019) 91 Legal Times 33

Ueno T, 人工知能による“発明”と“創作”「２０１７」 2017 Japio Year Book | ‘ “Invention” and “Creation” by Artificial Intelligence’ (2017) 2017 Japio Year Book

Ueno T, 著作権法における権利制限規定の再検討、日本版、フェアユース の可能性「２００７」コピライト 47 560 | ‘Reconsidering Copyright Limitations and Exceptions in Copyright Law and the Possibility of a Japanese-Style Fair Use’ (2007) Kopiraito 47 560

Ueno T, 平成 30 年著作権法改正について「２０１８」年報知的財産法２０１８－１９日本評論社 | ‘On the 2018 Amendment to the Copyright Act’ (2018) Intellectual Property Law Annual Report 2018–19, Hyōronsha

Ueno T, ‘The Flexible Copyright Exception for ‘Non-Enjoyment’ Purposes – Recent Amendment in Japan and Its Implication’ (2021) 70 GRUR Int 145

Unterthiner T and others, ‘Predicting Neural Network Accuracy From Weights’ (2020) < https://arxiv.org/pdf/2002.11448> accessed 21 May 2026

Van Den Hoogen J and others, ‘Hyperparameter Analysis of Wide-Kernel CNN Architectures in Industrial Fault Detection: An Exploratory Study’ (2024) 18 International Journal of Data Science and Analytics 423

Vasconcellos Grubow J, ‘OK Computer: The Devolution of Human Creativity and Granting Musical Copyrights to Artificially Intelligent Joint Authors’ (2018) 40 Cardozo L Rev 387

Vedder A and Naudts L, ‘Accountability for the Use of Algorithms in a Big Data Environment’ (2017) 31 International Review of Law, Computers & Technology 206

Verikas A, Gelzinis A and Bacauskiene M, ‘Mining Data with Random Forests: A Survey and Results of New Tests’ (2011) 44 Pattern Recognition 330

Wachter S, Mittelstadt B and Floridi L, ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’ (2017) 7 IDPL 76

Walker RK, ‘Breaking With Convention: The Conceptual Failings of Scènes à Faire’ (2020) 38 Cardozo Arts & Entertainment Law Journal 435

Wan Y and Lu H, ‘Copyright protection for AI-generated outputs: The experience from China.’ (2021) 42 CLS Rev 105581

Weerts HJ, Mueller AC and Vanschoren J, ‘Importance of Tuning Hyperparameters of Machine Learning Algorithms’ (2020) < https://arxiv.org/pdf/2007.07588> accessed 21 May 2026

Weinreb LL, ‘Copyright for Functional Expression’ (1998) 5 HLR 1149

Wenskay DL, ‘Intellectual Property Protection for Neural Networks’ (1990) 3 Neural Networks 229

Wenskay DL, ‘Neural Networks: A Prescription for Effective Protection’ (1991) 8 Computer Law 12

Wiley JS, ‘Copyright at the School of Patent’ (1991) 58 UCLR 119

Wu AJ, ‘From Video Games to Artificial Intelligence: Assigning Copyright Ownership to Works Generated by Increasingly Sophisticated Computer Programs’ (1997) 25 American Intellectual Property Law Association 131

Wu T, ‘Copyright's Communications Policy’ (2004) 103 Mich L Rev 278

Wymeersch P, ‘EU Copyright Exceptions and Limitations and the Three-step Test: One Step Forward, Two Steps Back’ (2023) 72 GRUR Int 631

Yamashita R and others, ‘Convolutional Neural Networks: an Overview and Application in Radiology’ (2018) 9 Insights into Imaging 611

Yanisky-Ravid S and Liu XJ, ‘When Artificial Intelligence Systems Produce Inventions: The 3A Era and an Alternative Model for Patent Law’ (2017) 39 Cardozo L Rev 2215

Yanisky-Ravid S and Velez-Hernandez LA, ‘Copyrightability of Artworks Produced by Creative Robots and Originality: The Formality-Objective Model’ (2018) 19 Minnesota Journal of Law, Science & Technology 1

Zhuk A, ‘Navigating the legal landscape of AI copyright: a comparative analysis of EU, US, and Chinese approaches’ (2024) AI Ethics 4

Legislation

European Union

Council Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs [1991] OJ L122/42

Council Directive 93/98/EEC of 29 October 1993 harmonizing the term of protection of copyright and certain related rights [1993] OJ L290/ 9

Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data [1995] OJ L281/31

Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases [1996] OJ L77/20

Directive 98/71/EC of the European Parliament and of the Council of 13 October 1998 on the legal protection of designs [1998] OJ L289/28

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society [2001] OJ L 167/10

Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights [2006] OJ L87/417

Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (codified version of 91/250/EEC) [2009] OJ L 111/16

Directive 2014/65/EU of the European Parliament and of the Council of 15 May 2014 on markets in financial instruments and amending Directive 2002/92/EC and Directive 2011/61/EU [2014] OJ L173/349

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1

Directive (EU) 2018/1808 of the European Parliament and of the Council of 14 November 2018 amending Directive 2010/13/EU on the coordination of certain provisions laid down by law, regulation or administrative action in Member States concerning the provision of audiovisual media services (Audiovisual Media Services Directive) in view of changing market realities [2018] OJ L 303/69

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC [2019] OJ L130/92

Directive (EU) 2019/2161 of the European Parliament and of the Council of 27 November 2019 amending Council Directive 93/13/EEC and Directives 98/6/EC, 2005/29/EC and 2011/83/EU of the European Parliament and of the Council as regards the better enforcement and modernisation of Union consumer protection rules [2019] OJ L328/7

Regulation (EU) 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services [2019] OJ L186/ 57

Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 [2022] OJ L152/1

Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) [2022] OJ L265/1

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277/1
Regulation 2023/2854 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) OJ L2023/2854

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2023] OJ L 2024/1689 12.7.2024

International Treaties and Agreements:

Agreement on Trade-Related Aspects of Intellectual Property Rights (Annex 1C) (15 April 1994) ("TRIPS")

Berne Convention for the Protection of Literary and Artistic Works (Paris Act of 1971) (28 September 1979)

International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organizations (26 October 1961)

Marrakesh Agreement Establishing the World Trade Organization (15 April 1994)

WIPO Copyright Treaty (20 December 1996) ("WCT")

WIPO Performances and Phonograms Treaty (20 December 1996) ("WPPT")

Japan:

Copyright Act No 48 of 1970 ("JCA")

United Kingdom:

Copyright (Computer Software) Amendment Act 1985

Copyright, Designs and Patents Act 1988 (“CDPA”)

The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014

United States:

The Computer Software Copyright Act Pub. L. 96-517 (1980)

Digital Millenium Copyright Act Pub. L. 105-304 (1998) (“DMCA”)

United States Code Title 17 – Copyrights Pub. L. 94–553 (1976) (“US Copyright Act”)

Monographs

Alpaydin E, *Machine Learning* (MIT Press 2021)

Aplin T, *Research Handbook on Intellectual Property and Digital Technologies* (Edward Elgar Publishing 2020)

Aplin T and Bently L, *Global Mandatory Fair Use: the Nature and Scope of the Right to Quote Copyright Works* (CUP 2020)

Bainbridge DI, *Information Technology and Intellectual Property Law* (6th edn, Bloomsbury Professional 2014)

Balganesh S, Loon N-LW and Sun H, *The Cambridge Handbook of Copyright Limitations and Exceptions* (Cambridge University Press 2021)

Band J and Katoh M, *Interfaces on Trial 2.0* (MIT Press 2011)

Bentham J, *The Works of Jeremy Bentham III* (William Tait 1843)

Beunen A, *Protection for Databases : The European Database Directive and its Effects in the Netherlands, France and the United Kingdom* (Wolf Legal Publishers 2007)

Borghi M and Karapapa S, *Copyright and Mass Digitization* (OUP 2013)

Boyle J and Jenkins J, *Intellectual Property: Law & the Information Society* (Duke Law School 2014)

Brown MS, *Data Mining for Dummies* (John Wiley & Sons 2014)

Burri M and Cottier T, *Trade Governance in the Digital Age: World Trade Forum* (CUP 2012)

Corrales M, Fenwick M and Forgo N (eds), *Robotics, AI and the Future of Law* (Springer 2018)

Derclaye E, *The Legal Protection of Databases: A Comparative Analysis* (Edward Elgar Publishing 2008)

Drahos P and Braithwaite J, *Information Feudalism* (Earthscan 2002)

Dreier T and Hugenholtz PB, *Concise European Copyright Law* (Kluwer Law International BV 2016)

Drexl J, *Research Handbook on Intellectual Property and Competition law* (Edward Elgar Publishing 2010)

Du Sautoy M, *The Creativity Code: How AI is Learning to Write, Paint and Think* (Fourth Estate 2019)

Ducci F, *Natural Monopolies in Digital Platform Markets* (Cambridge University Press 2020)

Dutt S, Subramanian C, Das, AK, *Machine Learning* (Pearson India 2018)

Ebers, M, Cantero Gamito, M (eds), *Algorithmic Governance and Governance of Algorithms* (Springer 2021)

Efroni Z, Access-*Right: The Future of Digital Copyright Law* (Oxford University Press 2011)

Evans DS, Fels A and Tucker C (eds), *The Evolution of Antitrust in the Digital Era: Essays on Competition policy.* Volume 1 and 2 (Competition Policy International 2020)

Fitsilis F, *Imposing Regulation on Advanced Algorithms* (Springer 2019)

Foster D, *Generative Deep Learning* (O'Reilly 2019)

Ganea P and others, *Japanese Copyright Law: Writings in Honour of Gerhard Schricker* (Kluwer Law International 2005)

Géron A, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems* (2nd edn, O'Reilly 2019)

Glanert S (ed), *Comparative Law - Engaging Translation* (Routledge 2014)

Goldstein P, *Goldstein on Copyright* (Wolters Kluwer 2007)

Goodfellow I, Bengio Y and Courville A, *Deep Learning* (MIT Press 2016)

Han J and Kamber M, *Data Mining: Concepts and Techniques* (3rd edn, Elsevier 2012)

Haskell J and Westlake S, *Capital without Capital* (Princeton University Press 2018)

Haskell J and Westlake S, *Restarting the Future: How to Fix the Intangible Economy* (Princeton University Press 2022)

Haykin S, *Neural Networks and Learning Machines* (Pearson 2009)

Heaton J, *Introduction to Neural Networks with Java* (Heaton Research 2008)

Kanai S, Takahashi J and Miyagawa T, ＤＸ時代におけるデジタルコンテンツ著作権法「ぎょうせい２０２１」| *Digital Contents and Copyright Law in an Era of Digital Transformation* (Gyōsei 2021)

Kato M, 著作権法逐条講義 『六訂新版』「著作権情報センター」2013 | *An Article by Article Commentary on Japanese Copyright Law 6th Edition* (Copyright Research Information Centre 2013)

Kato M, 著作権法逐条講義 『七訂新版』「著作権情報センター2021」 | *An Article by Article Commentary on Japanese Copyright Law 7th Edition* (Copyright Research Information Centre 2021)

Kelleher JD, *Deep Learning* (MIT Press 2019)

Kelleher JD, MacNamee B and D'Arcy A, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies* (MIT Press 2020)

Kelleher JD and Tierney B, *Data Science* (MIT Press 2018)

Knuth DE, *The Art of Computer Programming: Fundamental Algorithms, Volume 1* (Addison-Wesley Professional 1997)

Koepsell DR, *The Ontology of Cyberspace: Philosophy, Law, and the Future of Intellectual Property* (Open Court Publishing 2003)

Koo J, *The Right of Communication to the Public in EU Copyright Law* (Hart 2019)

Lantz B, *Machine Learning with R* (Packt Publishing 2019)

Leaffer MA, *Understanding Copyright Law* (LexisNexis 2010)

Lee J-A, Hilty R and Liu K-C, *Artificial Intelligence and Intellectual Property* (OUP 2021)

Lemley M, Samuelson P and others, *Software and Internet Law* (Aspen Publishing 2011)

Liu B, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data.* Vol 1 (Springer 2011)

Marr D, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* (Henry Holt & Co 1982)

Martín JG and others, *AI Knowledge Transfer from the University to Society* (CRC 2022)

Marušić B, *The Autonomous Legal Concept of Communication to the Public* (Elgar 2023)

Maskus KE, *Intellectual Property Rights in the Global Economy* (Peterson Institute 2000)

Maskus KE, *Encouraging International Technology Transfer.* Vol 7 (Citeseer 2004)

Meier GM and Stiglitz JE, *Frontiers of Development Economics: The Future in Perspective* (World Bank Publications 2001)

Menell PS, Merges RP, Samuelson P, Lemley M, *Software and Internet Law 4th Edition* (Aspen Law & Business 2011)

Meyer B, *Object-Oriented Software Construction* (Prentice-Hall 1988)

Morimoto D and Hamano T, AI・データ関連契約の実務 —AI 技術「中央経済社 2020」| *Practical Aspects of AI and Data-Related Contracts – AI Technology* (Chūōsha 2020)

Müller AC and Guido S, *Introduction to Machine Learning with Python* (O'Reilly 2016)

Nakayama N. これでいいのか！２０１８年著作権法改正－ほど遠い「日本版フェアユース」確立への道 「インプレス２０１９」 | *Is This Good Enough? The 2018 Copyright Reform and the Long Road to a Japanese-Style Fair Use System* (Impress 2019)

Nakayama N, ソフトウエアの法的保護「有斐閣 １９９０」 | *Legal Protection of Software* (Yūhikaku 1990)

Nakayama N, 著作権法第３版「有斐閣 ２０２０」| *Copyright Law 3rd Edition* (Yūhikaku 1990)

Nilsson NJ, *The Quest for Artificial Intelligence* (Cambridge University Press 2009)

Nimmer D, *Nimmer on Copyright* (Mathew Bender 2019)

Ogura H and Kanai S, 著作権法コメンタールⅠ　改訂版「第一法規 ２０２０」 | *A Commentary on Copyright Law I Revised Edition* (Daiichihōki 2020)

Øverby H and Audestad JA, *The Digital Economy - Foundations, Business Models and Case Studies* (Springer 2021)

Pagallo U, *The Laws of Robots: Crimes, Contracts, and Torts* (Springer 2013)

Patry WF, *How to Fix Copyright* (OUP 2012)

Patry WF, *Patry on Fair Use* (Thomson Reuters 2015)

Perzanowski A, *The Right to Repair* (CUP 2022)

Petit N, *Big Tech and the Digital Economy: The Moligopoly Scenario* (OUP 2020)

Pointer I, *Programming PyTorch for Deep Learning - Creating and Deploying Deep Learning Applications* (O'Reilly 2019)

Ricketson S, *The Berne Convention for the Protection of Literary and Artistic Works: 1886-1986* (Kluwer 1987)

Rikap C, *Capitalism, Power and Innovation: Intellectual Monopoly Capitalism Uncovered* (Routledge 2021)

Russell S and Norvig P, *Artificial Intelligence – A Modern Approach* (Pearson Education 2010)

Senftleben M, *Copyright, Limitations and the Three-Step Test* (Kluwer Law International 2004)

Shemtov N, *Beyond the Code: Protection of Non-Textual Features of Software* (OUP 2017)

Shemtov N and Walden I, *Free and Open Source Software: Policy, Law and Practice* (OUP 2014)

Sherman B and Wiseman L, *Copyright and the Challenge of the New* (Aspen Publishers 2012)

Siems M and Yap PJ (eds), *The Cambridge Handbook of Comparative Law* (CUP 2024)

Stucke M and Grunes A, *Big Data and Competition Policy* (OUP 2016)

Theobald O, *Data Analytics for Absolute Beginners* (Scatter Plot Press 2017)

Turner JD, *Intellectual Property and EU Competition Law* (OUP 2010)

Twining W, *Academic Law and Legal Development* (University of Lagos 1976)

Walter M and von Lewinski S, *European Copyright Law: A Commentary* (OUP 2010)

Warwick K, *Artificial intelligence: The Basics* (Routledge 2012)

Watkins D and Burton M (eds), *Research Methods in Law* (Routledge 2026)

Witten I, Frank E, Hall, MA, Pal, C, *Data Mining: Practical Machine Learning Tools and Techniques* (4th edn, Elsevier 2016)

Monograph Chapters

Adede A, 'Origins and History of the Trips Negotiations' in Bellman C and others (eds), *Trading in Knowledge: Development Perspectives on Trips, Trade and Sustainability* (Routledge 2003)

Aizawa H, 著作権法のパラダイムへの小論 in E Aizawa and others (eds), 知的財産法の理論と現代的課題 中山信弘先生還暦記念論文集「弘文堂２００５」| 'An Essay on the Paradigm of Copyright Law' in E Aizawa and others (eds), *Theory of Intellectual Property Law and Contemporary Issues: Commemorative Essays Honouring Professor Nobuhiro Nakayama on his 60th Birthday* (Kōbundō, 2005)

Baker J, 'The Role of Antitrust Law in Protecting and Fostering Online Platform Competition' in Evans DS and others (eds), *The Evolution of Antitrust in the Digital Era Volume Two* (Competition Policy International 2021)

Barwise P and Watkins L, 'The Evolution of Digital Dominance - How and Why We Got to GAFA' in Moore M and Tambini D (eds), *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (OUP 2018)

Borghi M, 'Exceptions and Users' Rights?' in Rosati E (ed), *The Routledge Handbook of EU Copyright Law* (Routledge 2021)

Borghi M, 'Owning Form, Sharing Content: Natural-right Copyright and Digital Environment' in Macmillan F (ed), *New Direction in Copyright Law* Vol 5 (Edward Elgar 2007)

Borghi M and White B, 'Data Extractivism and Public Access to Algorithms' in Borghi M and Brownsword R (eds), *Law, Regulation and Governance in the Information Society: Informational Rights and Informational Wrongs* (Routledge 2022)

Brownsword R, 'Transformative Technologies and Responsive Legal Scholarship' in Mendis D and others (eds), *3D Printing and Beyond* (Edward Elgar Publishing 2019)

Calboli I, 'Comparative Legal Analysis and Intellectual Property Law: A Guide for Research' in Calboli I and Montagnani M L (eds), *Handbook of Intellectual Property Research* (OUP 2021)

Calboli I, 'The role of comparative legal analysis in intellectual property law: From good to great?' in Dinwoodie G B (ed), *Methods and Perspectives in Intellectual Property* (Edward Elgar 2013)

Chen H Y, 'Copyright Protection for Software 2.0?: Rethinking the Justification of Software Protection Under Copyright Law' in Hilty R and others (eds), *Artificial Intelligence and Intellectual Property* (OUP 2021)

Cheng HWJ and Parra-Lancourt M, 'The Fourth Industrial Revolution, Development and Intellectual Property – The World Economic and Social Survey 2018 and Beyond' in Heath C, Sanders AK & Moerland A (eds), *Intellectual Property Law and the Fourth Industrial Revolution* (Wolters Kluwer 2020)

Colombo G F and Dragoni M, 'The Legal Protection of Software in Japan – an Original Model?' in Bellantuono G and others (eds), *Law, Development and Innovation* (Springer 2016)

Cosman F and Cohen I, 'Risks of Semi-Supervised Learning' in Chapelle O and others (eds), *Semi-Supervised Learning* (MIT Press 2006)

Deltorn J M, 'The Elusive Intellectual Property Protection of Trained Machine Learning Models: a European Perspective' in Abbott R (ed), *Research Handbook on Intellectual Property and Artificial Intelligence* (Edward Elgar 2022)

den Hertog J, 'General Theories of Regulation' in Bouckaert B and De Geest G (eds), *Encyclopedia of Law and Economics. Volume III. The Regulation of Contracts* (Edward Elgar 2000)

Determann L and Nimmer D, 'Software Copyright in the Cloud' in Yoo CS and Blanchette J F (eds), *Regulating the Cloud: Policy for Computing Infrastructure* (MIT Press 2015)

Drexl J, 'Is There a 'more Economic Approach' to Intellectual Property and Competition Law?' in Drexl J (ed), *Research Handbook on Intellectual Property and Competition* Law (Edward Elgar 2008)

Fisher W, ‘Theories of Intellectual Property’ in Munzer S (ed), *New Essays in the Legal and Political Theory of Property* (CUP 2001)

Fukui K and Ishiyama H, 知財実務の動き－AI ネットワークの近未来予測と知的財産権 in Takahashi R and others (eds), 年報知的財産法２０１６－２０１７「日本評論社２０１６」| ‘Developments in Intellectual Property Practice: Prospects for the Near Future of AI Networks and Intellectual Property Rights’ in Takahashi R and others (eds), *Yearbook of Intellectual Property Law 2016–2017* (Nihon Hyōronsha 2016)

Geiger C, ‘Copyright as an Access Right, Securing Cultural Participation Through the Protection of Creators’ Interests’ in Giblin R and Weatherall K (eds), *What if We Could Reimagine Copyright?* (ANU Press 2017)

Geiger C, ‘The Social Function of Intellectual Property Rights, or How Ethics Can Influence the Shape and Use of IP Law’ in Dinwoodie GB (ed), *Methods and Perspectives in Intellectual Property* (Edward Elgar Publishing 2013)

Geiger C, Schönherr F and Jütte BJ, ‘Limitations to Copyright in the Digital Age, Safeguards for User’s Rights, Creativity and Authors Remuneration Interests’ in Savin A and Trzaskowski J (eds), *Research Handbook on EU Internet Law* (Edward Elgar Publishing 2023)

Goldfarb A and Trefler D, ‘Artificial Intelligence and International Trade’ in Agrawal A and others (eds), *The Economics of Artificial Intelligence* (University of Chicago Press 2019)

Graef I, ‘When Data Evolves Into Market Power – Data Concentration and Data Abuse Under Competition Law’ in Moore M and Tambini D (eds), *Digital Dominance: the Power of Google, Amazon, Facebook, and Apple* (OUP 2018)

Handke C and others, ‘Is Europe Falling Behind in Data Mining? Copyright’s Impact on Data Mining in Academic Research’ in Schmidt B and Dobreva M (eds), *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust* (IOS Press 2015)

He T, ‘Copyright Exceptions Reform and AI Data Analysis in China: A Modest Proposal’ in Hilty R, Lee JA, Liu KC (eds), *Artificial Intelligence & Intellectual Property* (OUP 2021)

Harding A and Lee YS, ‘Law and Development’ in Siems M and Yap PJ (eds) ,*The Cambridge Handbook of Comparative Law* (CUP 2024)

Hilty R and others, ‘Intellectual Property Justification for Artificial Intelligence’ in Hilty R, Lee JA, Liu KC (eds), *Artificial Intelligence & Intellectual Property* (Oxford University Press 2021)

Hussa J, ‘Traditional Methods’ in Siems M and Yap PJ (eds), *The Cambridge Handbook of Comparative Law* (CUP 2024)

Hutchinson T, ‘Doctrinal Research’ in Dawn Watkins and Mandy Burton (eds), *Research Methods in Law* (Routledge 2026)

Jongsma D, ‘The Nature and Content of the Three-Step Test in EU Copyright Law: A Reappraisal’ in Rosati E (ed), *The Routledge Handbook of EU Copyright Law* (Routledge 2021)

Karapapa S, ‘The Quotation Exception Under EU Copyright Law’ in Rosati E (ed), *The Routledge Handbook of EU Copyright Law* (Routledge 2021)

Kelsey J, ‘The Risks for ASEAN of New Mega-Agreements That Promote the Wrong Model of E-Commerce’ in Chen L and Kimura F (eds), *Developing the Digital Economy in ASEAN* (Routledge 2019)

Kohashi K, 技術的保護手段と著作物の自由利用 in Koizumi N and Tamura Y (eds), 羽ばたきー２１世紀の知的財産法 「弘文堂 ２０１５」| ‘Technological Protection Measures and the Free Use of Copyrighted Works’ in Koizumi N and Tamura Y (eds), *Taking Flight: Intellectual Property Law in the 21st Century* (Kōbundō 2015)

Leistner M, ‘Big Data and the EU Database Directive 96/9/EC: Current Law and Potential for Reform’ in Lohsse S, Schulze R and Staudenmayer D (eds), *Trading Data in the Digital Economy: Legal Concepts and Tools: Münster Colloquia on EU Law and the Digital Economy III* (Nomos 2017)

Leistner M, ‘The Existing European IP Rights System and the Data Economy – An Overview with Particular Focus on Data Access and Portability’ in Drexl J (ed), *Data Access, Consumer Protection and Public Welfare* (Nomos 2021)

Margoni T, ‘The Harmonisation of EU Copyright Law: The Originality Standard’ in Perry M (ed), *Global Governance of Intellectual Property in the 21st Century* (Springer 2016)

Maskus KE, ‘Policy Space in Intellectual Property Rights and Technology Transfer: a New Economic Research Agenda’ in Correa C and Seuba X (eds), *Intellectual Property and Development: Understanding the Interfaces* (Springer 2019)

Mezei P, ‘You AIn't Seen Nothing Yet’: Arguments Against the Protectability of AI-Generated Outputs by Copyright Law’ in Borghi M and Brownsword R (eds), *Law, Regulation and Governance in the Information Society: Informational Rights and Informational Wrongs* (Routledge 2022)

Nimmer D, ‘Security Measures and the Future of Fair Use’ in Balganesh S and others (eds), *The Cambridge Handbook of Copyright Limitations and Exceptions* (Cambridge University Press 2021)

Oh S, Schiele B and Fritz M, ‘Towards Reverse-engineering Black-box Neural Networks’ in Samek W (ed), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer 2019)

Okediji RL, ‘Reframing International Copyright Limitations and Exceptions as Development Policy’ (2017) in Okediji RL (ed), *Copyright Law in an Age of Limitations and Exceptions* (CUP 2017)

Palmiotto F, ‘The Black Box on Trial: The Impact of Algorithmic Opacity on Fair Trial Rights in Criminal Proceedings’ in Ebers M and Gamito MC (eds), *Algorithmic Governance and Governance of Algorithms* (Springer 2021)

Reichman JH, ‘The Limits of “Limitations and Exceptions” in Copyright Law’ in Okediji RL (ed), *Copyright Law in an Age of Limitations and Exceptions* (CUP 2017)

Samuelson P, ‘A Square Peg in a Round Hole? Copyright Protection for Computer Programs’ in Sherman B and Wiseman L (eds), *Copyright and the Challenge of the New* (Wolters Kluwer 2012)

Samuelson P, ‘Justifications for Limitations and Exceptions’ in Okediji RL (ed), *Copyright Law in an Age of Limitations and Exceptions* (CUP 2017)

Sganga C, ‘The Right of Reproduction’ in Rosati E (ed), *Routledge Handbook of EU Copyright Law* (Routledge 2021)

Siems J, ‘Protecting Deep Learning: Could the New EU-Trade Secrets Directive Be an Option for the Legal Protection of Artificial Neural Networks?’ in Ebers M and Gamito MC (eds), *Algorithmic Governance and Governance of Algorithms* (Springer 2021)

Slowinski P, ‘Rethinking Software Protection’ in Hilty R and others (eds), *Artificial Intelligence and Intellectual Property* (OUP) 2021)

Sobel B, ‘A Taxonomy of Training Data: Disentangling the Mismatched Rights, Remedies, and Rationales for Restricting Machine Learning’ in (2020) in Hilty R, Lee JA, Liu KC (eds), *Artificial Intelligence & Intellectual Property* (OUP 2021)

Stiglitz J, ‘Towards a Broader View of Competition Policy’ (2017) 270 in Bonakele T, Fox E and Mncube L (eds), *Competition Policy for the New Era: Insights From the BRICS Countries* (OUP 2018)

Sucker M, ‘The Software Directive - Between the Combat Against Piracy and the Preservation of Undistorted Competition’ (1993) in Lehmann M and & Tapper CF (eds), *A Handbook of European Software Law* (Clarendon Press 1993)

Sugiyama K, フェアユースと教育利用 in A Moriizumi (ed), 著作権法と民法の現代的課題: 半田正夫先生古稀記念論集 「法学書院２００３」 | ‘Fair Use and Use for Educational Purposes’ in A Moriizumi (ed), *Contemporary Issues in Copyright Law and Civil Law: Essays in Honour of Professor Masao Handa on His Seventieth Birthday* (Hōgakushoin 2003)

Surblytė G, ‘Enhancing Trips: Trade Secrets and Reverse Engineering’ in Ullrich H and others (eds), *Trips Plus 20: From Trade Rules to Market Principles* (Springer 2016)

Turing AM, ‘Computing Machinery and Intelligence’ in Epstein R and others (eds), *Parsing the Turing Test* (Springer 2009)

Ueno T, 権利制限の一般規定 in Nakayama N and Kaneko T (eds), しなやかな著作権制度に向けて「信山社２０１７」 | ‘General Provisions in Limitations and Exceptions’ in Nakayama N and Kaneko T (eds), *Towards a Flexible Copyright System* (Shinzansha 2017)

Ueno T, 創作性 in Takahashi R and others (eds) 知的財産法の理論的探究 現代知的財産法講座１「評論社２０１２」| ‘Creativity’ in Takahashi R and others (eds), *Theoretical Explorations in Intellectual Property Law: Contemporary Intellectual Property Law Series 1* (Hyōronsha 2012)

Unni V, ‘Software Protection Under Copyright Law’ in Sinha MJ and Mahalwar V (eds), *Copyright Law in the Digital World* (Springer 2017)

van Pottelsberghe de la Potterie B, ‘Innovation Economics Approach to Intellectual Property Law Scholarship’ in Calboli I and Montagnani ML (eds), *Handbook of Intellectual Property Research* (OUP 2021)

Von Lewinski K, ‘BDSG § 6a Automatisierte Einzelentscheidung’ in Wolff HA and Brink S (eds), *Beck’scher Online-Kommentar Datenschutzrecht* (Beck-online 2016)

Wang RL, ‘Ancillary Orders of Compulsory Licensing and Their Compatibility with the TRIPS Agreement’ in Hilty R and Liu KC (eds), *Compulsory Licensing: Practical Experiences and Ways Forward* (Springer 2014)

Wataru S, プログラム著作物の創作性について in Koizumi N and Tamura Y (eds), 羽ばたき－２１世紀の知的財産法－中山信弘先生古稀記念論文集「弘文堂２０１５」| ‘Creativity in Software Works’ in Koizumi N and Tamura Y (eds), *Taking Flight: Intellectual Property Law in the 21st Century* (Kōbundō 2015)

Newspaper Articles and Online Publications

Autocrypt.io, ‘Top 4 Car Features That Rely on Vehicle-to-cloud Connectivity’*Autocrypt.io* (Seoul, 16th February 2022) < https://www.autocrypt.io/4-car-features-rely-on-vehicle-to-cloud-connectivity/> accessed 20 May 2026

Alcantara C and others, ‘How Big Tech Got So Big: Hundreds of Acquisitions’ *The Washington Post* (Washington, 21 April 2021) < www.washingtonpost.com/technology/interactive/2021/amazon-apple-facebook-google-acquisitions/> accessed 16 December 2025

Chuvpilo G, ‘AI Research Rankings 2020: Can the United States Stay Ahead of China?’ *Medium* (San Francisco, 20 December 2020) < https://chuvpilo.medium.com/ai-research-rankings-2020-can-the-united-states-stay-ahead-of-china-61cf14b1216> accessed 15 December 2025

Ciuriak D, ‘Industrial-era Investment Strategies Won’t Work in a Data-driven Economy’ *Centre for International Governance Innovation* (Waterloo, 15th November 2018) < www.cigionline.org/ articles/industrial-era-investment-strategies-wont-work-data-driven-economy/ > accessed 15 December 2025

Fingas J, ‘EU Copyright Laws Face Uncertain Fate After 11 Countries Reject Proposal’ *engadget* (San Francisco, 19 January 2019) < www.engadget.com/2019-01-19-eu-copyright-directive-rejected-at-council.html> accessed 15 December 2025

Gakuto T and Nakajima Y, ‘Japan and UK to Protect Encryption Keys in New Trade Deal’ *Nikkei Asia* (Tokyo, 25 July 2020) <https://asia.nikkei.com/politics/international-relations/japan-and-uk-to-https://www.google.com/search?client=safari&rls=en&q=Hill+R.+%E2%80%98Second+Contribution+to+the+June-September+2017+Open+Consultation+of+the+ITU+CWG-Internet%E2%80%99+(Association+for+Proper+Internet+Governance+2017)&ie=UTF-8&oe=UTF-8&channel=46protect-encryption-keys-in-new-trade-deal> accessed 20 May 2026

Hill R, ‘Second Contribution to the June-September 2017 Open Consultation of the ITU CWG-Internet’ *Association for Proper Internet Governance* (Geneva, 5 April 2018) < https://committees. parliament. uk/writtenevidence/88879/html/> accessed 20 May 2026

Hill R, ‘Why Should Data Flow Freely?’ *Association for Proper Internet Governance* (Geneva, March 2018) < www.apig.ch/Forum%202018%20Policy%20statement.pdf> accessed 20 May 2026

Hugenholtz PB, ‘The “Data Producer’s Right”: Unwelcome Guest in the House of IP’ *Kluwer Copyright Blog* (Amsterdam, 25 August 2017) < https://legalblogs.wolterskluwer.com/copyright-blog/the-data-producers-right-unwelcome-guest-in-the-house-of-ip/> accessed 10 December 2025

Japan Wire, ‘Japan OKs highly autonomous “level-4” driving from April 2023’ *Japan Wire* (Tokyo, 20 December 2022) <https://english.kyodonews.net/articles/-/38229> accessed 12 January 2023

Jolly J and Wearden G, ‘Landmark Moment as Uber Unveils First Annual Profit as Limited Company’ *The Guardian* (London, 7 February 2024) < www.theguardian.com/technology/2024/ feb/07/landmark-moment-as-uber-unveils-first-annual-profit-as-limited-company> accessed 14 August 2024

Karpathy A, ‘Software 2.0’ *Medium* (San Francisco, 11 November 2017) < https://karpathy.medium.com/ software-2-0-a64152b37c35> accessed 14 August 2024

Kaye K, ‘How the Tech Industry Coordinated to Squelch Algorithm Transparency in the New NAFTA Deal’ *Redtail* (Portland, 8 November 2018) < https://redtailmedia.org/2018/11/08/how-the-tech-industry-prevented-algorithm-transparency-in-nafta-2-0/> accessed 9 November 2021

Kondō K, ソフトバンク孫代表、株主総会で「人工知能の進化」について熱弁をふるう | ‘SoftBank CEO Son Speaks Passionately About “The Evolution of Artificial Intelligence” at the Annual Shareholders’ Meeting’ *My Navi News* (Tokyo, 23 June 2016) < https://news.mynavi.jp/article/20160623-singularity/ > accessed 29 July 2022

Love J, ‘TPP Has Provision Banning Requirements to Transfer of or Access to Source Code of Software’ *Knowledge Ecology International* (Geneva, 6 November 2015) <www.keionline.org/ 22935> accessed 17 January 2023

Luscome R, ‘Google Engineer Put on Leave After Saying AI Chatbot Has Become Sentient’ *The Guardian* (London, 12 June 2022) <www.theguardian.com/technology/2022/jun/12/google-engineer-ai-bot-sentient-blake-lemoine> accessed 20 May 2026

Malone M, ‘Silicon Insider: When Did Amazon Get Profitable?’ *ABC News* (New York, 29 January 2002) < https://abcnews.com/Business/story?id=87393&page=1> accessed 20 May 2026

Margoni T and Kretschmer M, ‘The Text and Data Mining Exception in the Proposal for a Directive on Copyright in the Digital Single Market: Why It is Not What EU Copyright Law Needs’ *CREATe* (Glasgow, 25 April 2018) <www.create.ac.uk/blog/2018/04/25/why-tdm-exception-copyright-directive-digital-single-market-not-what-eu-copyright-needs/> accessed 20 May 2022

Markoff J, ‘Business Technology. What’s the Best Answer? It’s Survival of the Fittest.’ *New York Times* (New York, 29 August 1990) < www.nytimes.com/1990/08/29/business/business-technology-what-s-the-best-answer-it-s-survival-of-the-fittest.html> accessed 17 January 2023

McAllister N, ‘Study: Most Projects on GitHub Not Open Source Licensed.’ *The Register* (London, 18 April 2013) < www.theregister.com/software/2013/04/18/study-most-projects-on-github-not-open-source-licensed/847877#:~:text=But%20new%20research%20shows%20that,as%20open%20source%20software%20uncertain.> accessed 20 May 2026

Meinert R, 'Optimizing Hyperparameters in Random Forest Classification' *Medium* (San Francisco, 5 June 2019) < https://medium.com/data-science/optimizing-hyperparameters-in-random-forest-classification-ec7741f9d3f6> accessed 20 May 2026

Murray-Rust P, 'British Library Document on Copyright' *Petermr's blog* (Cambridge, 9 May 2009) <https://blogs.ch.cam.ac.uk/pmr/> accessed 17 January 2022

Naughton J, 'AI is Making Literary Leaps – Now We Need the Rules to Catch Up' *The Guardian* (London, 2 November 2019) < www.theguardian.com/commentisfree/2019/nov/02/ai-artificial-intelligence-language-openai-cpt2-release> accessed 1 December 2025

Nelson J, 'You Might Be Resizing Your Images Incorrectly' *Roboflow* (Des Moines, 31 January 2021) < https://blog.roboflow.com/you-might-be-resizing-your-images-incorrectly/> 19 November 2022

New York Times, 'Computer Program Copyrighted for First Time; Columbia Law Student Gets Approval for Plans – Sees Wide Industry Impact' *New York Times* (New York, 8 May 1964) 43

Oliver WD, 'Quantum Computing Takes Flight' *Nature* (London, 2 October 2019) 487

Ollero DJ, 'El Algoritmo Secreto del Gobierno que Decide Si Te Llevas una Subvención para la Factura de la Luz' *El Mundo* (Madrid, 3 July 2019) < www.elmundo.es/tecnologia/2019/07/03/5d1b89fbfc6c83a2358b46ca.html> accessed 20 May 2026

O'Flaherty K, 'All the Data WhatsApp and Instagram Send to Facebook' *Wired* (San Francisco, 24 July 2021) < www.wired.com/story/whatsapp-instagram-facebook-data/> accessed 20 January 2025

Pasquinelli KM, 'Adapt Your IP Strategy for Artificial Intelligence' *Robins Kaplan LLP* (2019) < www.robinskaplan.com/assets/htmldocuments/uploads/pdfs/publications/b0fe3a1f87594916817ca748b8897596__adapt-your-ip-strategy-for-artificial-intelligence.pdf> accessed 20 May 2023

Ravi Kanth D, 'Pluri-Accord on e-Com Could Reinforce Lawlessness in Digital Trade' *Third World Network Berhard* (13 August 2019) < https://twn.my/title2/wto.info/2019/ti190803.htm> accessed 20 May 2026

Reed B, 'Microsoft Has Lost $11 Billion Trying to Compete with Google' *Yahoo News* (July 2013) < www.yahoo.com/news/microsoft-lost-11-billion-trying-compete-google-185008151.html> accessed 20 May 2026

Rendyk, 'Tuning the Hyperparameters and Layers of Neural Network Deep Learning' *Analytics Vidhya* (Gurugram, 19 November 2021) < www.analyticsvidhya.com/blog/2021/05/tuning-the-hyperparameters-and-layers-of-neural-network-deep-learning/> accessed 20 May 2026

Rogerson B, 'Almost 60% of Musicians Are Already Using AI in Their Productions, Study Suggests' *Musicradar* (21 April 2023) < https://www.musicradar.com/news/musicians-already-using-ai> accessed 20 May 2026

Sag M, Samuelson P and Sprigman CJ, 'Comments in Response to the Copyright Office's Notice of Inquiry' (6 December 2023) <https://downloads.regulations.gov/COLC-2023-0006-10299/attachment_1.pdf> accessed 20 May 2026

Sagar R, 'Under the Hood of TESLA's AI Chip That Takes the Driverless Battle to Nvidia's Home Turf' *Analytics India Magazine* (Bangalore, 26 April 2019) < https://analyticsindiamag.com/deep-tech/under-the-hood-of-teslas-ai-chip-that-takes-the-driverless-battle-to-nvidias-home-turf> accessed 20 May 2026

Sayak P, 'Learning to Resize in Computer Vision' *Keras.io* (Carquefou, 30 April 2021) < https://keras.io/ examples/vision/learnable_resizer/> accessed 20 May 2026

Schick P, 'Tesla Business Model: Innovation, Scale and Network Effects' (2022) < https:// peterschick.eu/tesla-business-model-innovation-scale-and-network-effects/> last accessed 21 January 2024

Thompson B, 'Tensorflow and Monetizing Intellectual Property' *Stratechery* (12 November 2015) < https://stratechery.com/2015/tensorflow-and-monetizing-intellectual-property/> accessed 20 May 2025

Weston S, 'Google Cloud Lost £4.1 Billion in 2020' *IT Pro* (London, 3 February 2021) <www.itpro.com/ cloud/358506/google-cloud-lost-56-billion-in-2020> accessed 20 May 2026

Whittaker C, ‘7 Innovative Uses of Clustering Algorithms in the Real World’ *Datafloq* (The Hague, 4 April 2019) < https://datafloq.com/7-innovative-uses-of-clustering-algorithms/> accessed 20 May 2026

Workman D, ‘Top Industrial Robots Exporters’ *World’s Top Exports* (2025) <www.worldstopexports.com/top-industrial-robots-exporters/> accessed 20 May 2026

Zavrel J, ‘Must Read: the 100 Most Cited Papers in 2022’ *Zeta Alpha* (Amsterdam, 2 March 2023) < https://www.zeta-alpha.com/post/must-read-the-100-most-cited-ai-papers-in-2022> accessed 20 May 2026

Official Publications

European Union

European Commission, ‘Artificial Intelligence – A European Approach to Excellence and Trust’ (White Paper) COM 2020 65 final

European Commission, ‘Commission Staff Working Document on the Free Flow of Data and Emerging Issues of the European Data Economy’ (Staff Working Document) SWD 2017 2 final

European Commission, ‘Communications on Artificial Intelligence in Europe’ (Staff Working Document) COM 2018 237 final

European Commission, ‘Coordinated Plan on Artificial Intelligence’ (Communication) COM 2018 795 final

European Commission, ‘European Strategy for Data’ (Communication) COM 2020 66 final

European Parliamentary Research Service, ‘Polarisation and the Use of Technology in Political Campaigns and Communication’ (European Parliamentary Research Service 2019)

European Commission, ‘Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics’ (Report) COM 2020 64 final

Publications Office of the European Union, ‘The Draghi Report: Part B In-depth Analysis and Recommendations’ (Publications Office of the European Union 2025)

United Nations and United Nations Conference on Trade and Development (UNCTAD)

Banga R, 'Joint Statement Initiative on e-Commerce (JSI): Economic and Fiscal Implications for the South' (UNCTAD Research Paper 58. 2021)

UN, 'Eighth United Nations Conference to Review All Aspects of the Set of Multilaterally Agreed Equitable Principles and Rules for the Control of Restrictive Business Practices in 2020' (United Nations 2019)

UN, 'The Age of Digital Interdependence. Report of the UN Secretary-General's High-level Panel on Digital Cooperation' (United Nations 2019)

UNCTAD, 'Industry 4.0 for Inclusive Development' (UNCTAD 2022)

UNCTAD, 'Internationalisation Production Beyond the Pandemic' (UNCTAD World Investment Report 2020)

UNCTAD, 'Investment and the Digital Economy' (UNCTAD World Investment Report 2017)

UNCTAD, 'Technology and Innovation Report' (UNCTAD 2021)

UNCTAD, 'Trade and Development Report 2018: Power, Platforms and the Free Trade Delusion' (UNCTAD 2018)

UNCTAD, 'Value Creation and Capture: Implications for Developing Countries' (UNCTAD 2019)

UNCTAD, 'What is at Stake for Developing Countries in Trade Negotiations on E-Commerce? The Case of the Joint Statement Initiative' (UNCTAD 2020)

United States

Committee on the Judiciary, 'House Report on the Copyright Act of 1976' HR Rep 94-1476 (US Congress 3 September)

Library of Congress, 'Final Report of the National Commission on New Technological Uses of Copyrighted Works' (CONTU) (Library of Congress, 1978)

WTO

‘Work Programme on Electronic Commerce’ (World Trade Organisation 25 September 1998)

‘WTO Electronic Commerce Negotiations – Consolidated Negotiating Text’ INF/ECOM/62 (World Trade Organisation 7 December 2020)

Reports

Acumen Research and Consulting, ‘Generative AI Market and Region Forecast 2022 – 2030’ (Acumen Research and Consulting 2023)

Ademuyiwa I and Adeniran AP, ‘Assessing Digitalization and Data Governance Issues in Africa’ (Centre for International Governance Innovation. CIGI Paper 244 2020)

Aaronson SA and Struett T, ‘Data is Divisive: A History of Public Communications on e-Commerce 1998-2020’ (CIGI Papers No 247 2021)

Agrawal A, Gans J and Goldfarb A, ‘How to Win with Machine Learning’ (Harvard Business Review 2020)

Anaconda, ‘2020 State of Data Science: Moving from Hype Toward Maturity’ (Anaconda 2020)

Andrenelli A, Gourdon J and Moïsé E, ‘International Technology Transfer Policies’ (OECD 2019)

Association Littéraire et Artistique Internationale, ‘Report and Opinion on the Making Available and Communication to the Public in the Internet Environment: Focus on Linking Techniques on the Internet’ (Association Littéraire et Artistique Internationale 2013)

Ausloos J, and others, ‘Operationalizing Research Access in Platform Governance What to Learn from Other Industries?’ (Algorithm Watch 2020)

Backer KD and Flaig D, ‘The Future of Global Value Chains’ (OECD Science, Technology and Industry Policy Papers No 41. 2017)

Benaich N and Hogarth I, ‘State of AI Report’ (State of AI 2019)

Bessen JE and others, ‘The Business of AI Startups’ (Boston University School of Law, Law and Economics Research Paper 2018)

BSA, ‘Eighth Annual BSA Global Software Piracy Study – Trends in Software Piracy 1994 – 2002’ (BSA 2002)

Brynjolfsson E, Hitt LM and Yang S, ‘Intangible Assets: Computers and Organizational Capital’ (Brookings Papers on Economic Activity 137 2002)

Burri M, ‘Current and Emerging Trends in Disruptive Technologies: Implications for the Present and Future of EU's Trade Policy’ (Directorate General for External Policies 2017)

Castro D and McLaughlin M, ‘Who is Winning the AI Race: China, the EU, or the United States?’ (Information Technology & Innovation Foundation 2021)

CB Insights, ‘AI 100: The Most Promising AI Startups of 2022’ (CB Insights 2022)
Centre for Economic Policy Research. ‘Addressing Impediments to Digital Trade’ (Centre for Economic Policy Research 2021)

Ciuriak D, ‘Economic Rents and the Contours of Conflict in the Data-Driven Economy’ (Centre for International Governance Innovation Policy Brief 245 June 2020)

Ciuriak D,‘The Data-Driven Economy: Implications for Canada’s Economic Strategy’ (Centre for International Governance Innovation. Policy Brief No. 151 2019)

Computer & Communications Industry Association, ‘Request for Comments on Negotiating Objectives Regarding Modernization of the North American Free Trade Agreement with Canada and Mexico’ (CCIA 2017)

Cory N, ‘Cross-Border Data Flows: Where Are the Barriers, and What Do They Cost?’ (Information Technology & Innovation Foundation 2017)

Crémer J, De Montjoye Y-A and Schweitzer H, ‘Competition Policy for the Digital Era’ (Directorate-General for Competition 2019)

Crouzet N & Eberly JC, ‘Understanding Weak Capital Investment: The Role of Market Concentration and Intangibles’ (National Bureau of Economic Research 2019)

Darsinouei A E, ‘Understanding E-Commerce Issues in Trade Agreements: A Development Perspective Towards MC11 and Beyond’ (CUTS International 2007)

Ducci F, ‘Gatekeepers and Platform Regulation – Is the EU Moving in the Right Direction?’ (Sciences Po 2021)

Ernst D, ‘Competing in Artificial Intelligence Chips: China’s Challenge Amid Technology War’ (Centre for International Governance Innovation Special Report 2020)

Ferracane M and van der Marel E, ‘Do Data Policy Restrictions Inhibit Trade in Services?’ (Robert Schuman Centre for Advanced Studies Research Paper No 2019/29 2019)

Ficsor M, ‘Guide to Copyright and Related Rights Treaties Administered by WIPO and Glossary of Copyright and Related Rights Terms’ (WIPO 2003)

Garcia G, ‘Artificial Intelligence in Japan – Industrial Cooperation and Business Opportunities for European Companies’ (EU-Japan Centre for Industrial Cooperation 2016)

Guglya L and Maciel M, ‘Addressing the Digital Divide in the Joint Statement Initiative on e-Commerce’ (International Institute for Sustainable Development and CUTS International 2020)

Gurumurthy A and Chami N, ‘The Wicked Problem of AI Governance’ (Artificial Intelligence in India Vol 2 IT for Change 2019)

Gurumurthy A and Nandini C, ‘Digital Public Goods: A Precondition for Realising SDGs’ (IT for Change 2019)

Gurumurthy A and others, ‘Platform Planet: Development in the Intelligence Economy’ (IT For Change 2019)

Hufbauer GC and Lu Z, ‘Global E-Commerce Talks Stumble on Data Issues, Privacy, and More’ (Peterson Institute for International Economics 2019)

Irion K, ‘AI Regulation in the European Union and Trade Law: How Can Accountability of AI and a High Level of Consumer Protection Prevail Over a Trade Discipline on Source Code?’ (Verbraucherzentrale Bundesverband 2021)

Irion K and Williams J, ‘Prospective Policy Study on Artificial Intelligence and EU Trade Policy’ (The Institute for Information Law Amsterdam University 2019)

Ismail Y, ‘E-Commerce in the World Trade Organization’ (International Institute for Sustainable Development 2020)

James D, ‘Digital Trade Rules: A Disastrous New Constitution for the Global Economy Written by and for Big Tech’ (Center for Economic and Policy Research 2020)

Johnson L, ‘Space for Local Content Policies and Strategies’ (Deutsche Zusammenarbeit 2016)

Joseph KJ, ‘Information Technology Agreement of WTO: Call for a Revisit’ (Centre for WTO Studies WP/200/9 2013)

Kaggle, ‘State of Data Science and Machine Learning’ (Kaggle 2021)

Kaggle. ‘State of Data Science and Machine Learning’ (Kaggle 2022)

Kallummal M, ‘Process of Trade Liberalisation Under the Information Technology Agreement (ITA): The Indian Experience’ (Centre for WTO Studies WP/200/3 2012)

Kim L, ‘Technology Transfer & Intellectual Property Rights’ (UNCTAD - ICTSD Project on Intellectual Property Rights and Sustainable Development 2003)

Korinek A and Stiglitz JE, ‘Artificial Intelligence, Globalization, and Strategies for Economic Development’ (National Bureau of Economic Research 2021)

Körner K, ‘(How) Will the EU Become an AI Superstar?’ (Deutsche Bank Research 2020)

Kuczyńska A, ‘Analysis of Opportunities for EU SMEs in Japan’s Data Economy and Artificial Intelligence in Connection with Robotics’ (EU-Japan Centre for Industrial Cooperation 2019)

Lambrecht A and Tucker CE, ‘Can Big Data Protect a Firm from Competition?’ (Competition Policy International 2015)

Lee-Makiyama H, ‘AI & Trade Policy’ (Tallinn Digital Summit 2018)

Lianos I and Ivanov A, ‘Digital Era Competition BRICS Report’ (BRICS Competition and Law Centre 2019)

Manyika J and others, ‘Digital Globalization: The New Era of Global Flows’ (McKinsey Digital 2016)

Margoni T, ‘Artificial Intelligence, Machine Learning and EU Copyright Law: Who Owns AI?’ (CREATe Working Paper 2018/12 2018)

Mathys & Squire Consulting Ltd, ‘Economic Impact of IP on Investment in AI’ (Mathys & Squire Consulting Ltd 2022)

McKinsey Global Institute, ‘Notes from the AI Frontier. Modelling the impact of AI on the World Economy’ (McKinsey Global Institute 2018)

McKinsey Global Institute, ‘Notes From the AI Frontier: Tackling Europe’s Gap in Digital and AI’ (McKinsey Global Institute 2019)

McKinsey Global Institute, ‘Twenty-Five Years of Digitization: Ten Insights into How to Play It Right’ (McKinsey Global Institute 2019)

Mendis D, White B and Hong Dukki.,‘Copyright and Open Norms in Seven Jurisdictions: Benefits, Challenges & Policy Recommendations’ (CIPPM Research Paper 2024)

Neeraj RS, ‘Trade Rules on Source Code-deepening the Digital Inequities by Locking Up the Software Fortress’ (Centre for WTO Studies 2017)

Netherland Enterprise Agency, ‘Artificial Intelligence in Japan’ (Netherlands Enterprise Agency 2020)

NTT Data Institute of Management Consulting, 令和３年度産業経済研究委託事業・海外におけるデザイン・ブランド保護等新たな知財制度上の課題に関する実態調査「別紙２」「株式会社エヌ・ティ・ティ・データ経営研究所２０２１」| ‘2021 Industrial and Economic Research Commissioned Project: A Survey of New Intellectual Property System Issues Relating to Design and Brand Protection Overseas – Appendix 2’ (NTT Data Institute of Management Consulting Inc 2021)

Ocean Tomo, ‘Intangible Asset Market Value Study’ (Ocean Tomo 2020)

OECD, ‘Algorithms and Collusion’ (OECD 2017)

OECD, ‘Competition Economics of Digital Ecosystems’ (OECD 2021)

OECD, ‘Who Develops AI-Related Innovations, Goods and Services?’ (OECD Science, Technology and Industry Policy Papers 2021)

PwC, ‘Global Top 100 Companies by Market Capitalisation’ (PwC May 2021)

PwC, ‘Sizing the Prize. What’s the Real Value of AI for Your Business and How Can You Capitalise?’ (PwC 2017)

Rashmi B, ‘Implications of Signing Information Technology Agreement (ITA-1) and Expansion of ITA (ITA-2)’ (Centre for WTO Studies WP/200/57 2020)

Singh PJ, ‘Developing Countries in the Emerging Global Digital Order - A Critical Geopolitical Challenge to Which the Global South Must Respond’ (IT For Change 2017)

Smith SR, ‘Some Preliminary Implications of WTO Source Code Proposal’ (Third World Network Briefings 2017)

Stanford University, ‘Artificial Intelligence Index Report’ (Stanford University 2021)

Stanford University, ‘Artificial Intelligence Index Report’ (Stanford University 2023)

Stanford University, ‘Artificial Intelligence Index Report’ (Stanford University 2024)

Statista, ‘AI-Related Publications Worldwide 2016-2020 by Country’ (Statista 2023)

Statista, ‘Economic Impact on IP on Investment in AI’ (Statista 2022)

Statista, ‘Net Loss of TESLA’ (Statista 2022)

Statista, ‘Research and Development Expenses of TESLA’ (Statista 2022)

Stephenson D, ‘Fostering Growth in Digital Trade Through Bilateral Cooperation in the Development of Trade Rules’ (Centre for International Governance Innovation 2020)

Thormundsson B, ‘Artificial Intelligence (AI) Worldwide – Statistics & Facts’ (Statista 2024)

UK Intellectual Property Office, ‘Economic Impact of IP on Investment in AI’ (Intellectual Property Office Research Paper 2022)

Walch K, ‘Data Engineering, Preparation and Labelling for AI 2020’ (Cognilytica 2020)

Walport M, ‘The Internet of Things: Making the Most of the Second Digital Revolution – A Report by the UK Government Chief Scientific Adviser.’ (Government Office of Science 2014)

WIPO, ‘Artificial Intelligence’ (WIPO Technology Trends 2019)

Wu M, ‘Digital Trade-Related Provisions in Regional Trade Agreements: Existing Models and Lessons for the Multilateral Trade System.’ (International Centre for Trade and Sustainable Development (ICTSD) and the Inter-American Development Bank 2017)

Annex 1: A Guide for the Courts: Establishing the Extent of Copyright Protection for Trained AI Models Created with a ML Toolkit

Feature	Questions	Comments
Model Parameters (e.g. weights, biases, co-efficients, support vectors etc)	N/A	As individual datapoints automatically generated from, or isolated from data (support vectors), they lack a human author and are not protected by copyright law.
AI Toolkits	• Which AI toolkit or toolkits were used? • Which precise parts of the code not sourced from the toolkit are new and have been programmed by the plaintiff or defendant?	Where an AI toolkit / software library has been employed to create a trained model, establishing copyright protection beyond that which inheres in the toolkit will be significantly more onerous than where new programming has been undertaken from scratch. In such instances, it will be important to distinguish which specific elements of the code are original and have been authored by the plaintiff or defendant.
Model Architectures	• Is the model architecture entirely novel? • Are sub-elements of the architecture novel? • To what extent is the architecture derived from the toolkit, an open repository or from the literature?	When a party asserts that the architecture or part of the architecture is novel, independent verification from an expert witness will be required.
Model Architectures and Source Code	• Which elements are determined by the data? • Which elements are determined by the nature of the task?	The relevance of doctrines such as scènes à faire, merger and national equivalents, banal expression etc should be evaluated.

	• Which elements are determined by the software, hardware and model algorithm selected? • Which elements are determined by well-known heuristics and industry practice? • Which elements are determined by the business need and the local technical environment where the model is deployed?	
Model Architectures and Source Code	• Which elements are fixed by the optimisation and regularisation software or any other hyperparameter-related software?	Where features of the model are selected by the programmer based on optimal performance calculated by the software, their structure and arrangement constitute a non-protected idea.
Model Architectures and Source Code	• Which elements are only present in the training phase? • Which elements are present in the deployed model?	Given that copyright fixation requirements vary by jurisdiction establishing which elements of the program reach the threshold for protection across the development and deployment lifecycle is required.
Source Code	• Which elements of the code are determined by the toolkit and the programming environment? e.g. other software, hardware etc. • Which new elements of the code have been written not determined by the toolkit and programming environment?	Having distilled elements of the trained model which are not subject to copyright protection or are openly licensed and whose copyright is held by third parties, an analysis of the remaining authored software code created for deployment of the trained model will be required to determine whether, or to what extent, copyright subsists.

Glossary

Agglomerative clustering	A type of hierarchical clustering algorithm.
ANN	Artificial Neural Network. A type of model containing at least one hidden layer of calculations known as nodes or neurons.
Assembler	A program that converts assembly code into machine code.
Assembly code	A low-level language that is translatable into machine code.
Autoencoder	A system that captures important features from data (encoder) in a lower dimensional form which are then for the output mapped back into a higher format modified form. (decoder)
Compiler	A program that converts a high-level language into machine code.
CNN	Convolution Neural Network. A neural network with at least one convolutional layer which acts as a filter on inputs, creating a feature map summarising the presence of a certain feature.
DBSCAN	A density-based non-parametric algorithm for clustering.
Decision tree	A supervised learning algorithm commonly used for regression, classification, prediction etc.
Deep learning	Any type of machine learning which uses neural networks.
Ensemble model	A model which comprises of multiple individual models.
Foundation model	Models trained on a broad set of data that can be fine-tuned to perform different tasks.
Function	A discrete sub-program that performs a specific task.
GAN	Generative Adversarial Network. A model which develops by learning how to distinguish valid data from invalid data.
Generative AI	A probabilistic deep learning model which can produce original content that looks like it has been created using the same rules and relationships as the original training data.
Generative pre-trained transformers (GPT)	A family of a predictive transformer-based LLM developed by OpenAI.
General purpose model	See foundation model.
Hyperparameter	A topological or algorithmic model value that can be set by a data scientist.
Hyperplane	A linear "decision boundary" on high dimensional space informed by the position of the support vectors.

K-means clustering	An algorithm which categorises data points into clusters by using a mathematical distance measure.
K-nearest neighbours	A non-parametric supervised learning model commonly used for predictions, classifications etc.
Linear regression	A model that estimates the linear relationship between two or more variables.
LLM	Large language models. A language model containing many parameters.
Machine code	Binary code which directly runs the computer CPU.
Memorisation	See overfitting.
Nearest neighbours	Supervised and unsupervised algorithms which find the nearest data to the query point.
Neural network	See ANN.
Object code	Computer-readable statements and commands produced by an assembler or compiler from the source code.
Optimisation	Mechanisms which improve the accuracy of the prediction by reducing the model error.
Overfitting	Creating a model that matches too closely the training data, so it is unable to make accurate predictions on unseen data.
Parameters	Model variables which are created from learning from the data and the settings of the model hyperparameters. e.g. weights, coefficients, support vectors etc.
Random forests	A model which combines multiple decision trees.
Regularisation	Any mechanism or process which reduces overfitting.
Reinforcement learning	Models which learn and update through trial and error.
Source code	Plain "human readable" text statements and commands written in a high-level programming language.
Support vectors	Data points which inform the position of a hyperplane in an SVM model.
Supervised learning	Models which during the training phase use labelled data containing the correct target information.
Semi-supervised learning	Models which utilise a small amount of labelled data alongside a majority of unlabelled data.
SVM	Support vector model.
Transfer learning	Reusing elements from a prior trained model in a new task.

Transformer	A deep learning architecture developed by researchers at Google that analyses relationships in data and can make predictions.
Unsupervised learning	Models which are tasked “blind” to detect patterns in data.
VAE	Variational Autoencoder. A type of autoencoder useful for generative AI.
Weights	A type of parameter associated with neural networks, linear and logistic regression models, random forests etc.