

# MASTER THESIS | MASTER'S THESIS

Titel | Title

Intellectual Property Law implications in the EU and US when training  
Generative AI models

verfasst von | submitted by

Grace Egan

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of  
Master of Laws (LL.M.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt |  
Degree programme code as it appears on the  
student record sheet:

UA 999 082

Universitätslehrgang lt. Studienblatt |  
Postgraduate programme as it appears on  
the student record sheet:

Europäisches und Internationales Wirtschaftsrecht  
[Vollzeit Englisch]

Betreut von | Supervisor:

ao. Univ.-Prof. Mag. Dr. Siegfried Fina

### **Abstract English**

This paper examines the role of data scraping when training generative AI models primarily in the EU and the US. With the continued development of generative AI, there has been a notable increase in the demand for data to train AI models. This has sparked concerns about data collection methods and their subsequent impact on the protection of intellectual property rights. The training of large language models relies heavily on accessing vast data sets, which involves the automated extraction of information from sources such as databases and third-party platforms. This directly affects creators and owners of IP protected works, particularly when it is conducted without their payment or consent. This has raised the question of whether data scraping activities infringe intellectual property rights, particularly in relation to copyright. The legal landscape in this area is complex, with many existing IP laws predating modern AI systems and varying across jurisdictions, complicating their application. This paper discusses current legal frameworks and their applicability, focusing on ongoing case law in the field and on whether exceptions such as text and data mining or fair use provisions apply.

## **Abstract Deutsch**

Diese Arbeit untersucht die Rolle des Data Scraping beim Training generativer KI-Modelle vor allem in der EU und den USA. Mit der kontinuierlichen Weiterentwicklung generativer KI ist die Nachfrage nach Daten zum Training von KI-Modellen deutlich gestiegen. Dies hat zu Bedenken hinsichtlich der Methoden der Datenerhebung und deren Auswirkungen auf den Schutz geistiger Eigentumsrechte geführt. Das Training großer Sprachmodelle ist in hohem Maße auf den Zugriff auf umfangreiche Datensätze angewiesen, was die automatisierte Extraktion von Informationen aus Quellen wie Datenbanken und Plattformen von Drittanbietern beinhaltet. Dies hat direkte Auswirkungen auf die Urheber und Eigentümer von urheberrechtlich geschützten Werken, insbesondere wenn dies ohne deren Bezahlung oder Zustimmung erfolgt. Dies hat die Frage aufgeworfen, ob Data Scraping-Aktivitäten gegen Rechte des geistigen Eigentums verstoßen, insbesondere in Bezug auf das Urheberrecht. Die Rechtslage in diesem Bereich ist komplex, da viele bestehende Gesetze zum Schutz geistigen Eigentums aus der Zeit vor modernen KI-Systemen stammen und sich je nach Rechtsordnung unterscheiden, was ihre Anwendung erschwert. In dieser Arbeit werden die aktuellen rechtlichen Rahmenbedingungen und ihre Anwendbarkeit erörtert, wobei der Schwerpunkt auf der aktuellen Rechtsprechung in diesem Bereich und der Frage liegt, ob Ausnahmen wie Text- und Data-Mining oder Bestimmungen zur fairen Nutzung gelten.

## Table of Contents

|                                                                 |    |
|-----------------------------------------------------------------|----|
| <i>List of abbreviations</i> .....                              | 6  |
| 1. <i>Introduction</i> .....                                    | 7  |
| 2. <i>Generative Artificial Intelligence</i> .....              | 9  |
| 2.1 Defining Generative AI and Outlining Its Applications ..... | 9  |
| 2.2 What is data scraping? .....                                | 13 |
| 3. <i>Legal Landscape</i> .....                                 | 15 |
| 3.1 European Union .....                                        | 17 |
| 3.1.1 Objectives and Scope of the AI Act .....                  | 19 |
| 3.1.2 Obligations for general-purpose AI models .....           | 22 |
| 3.1.3 Code of Practice .....                                    | 23 |
| 3.1.4 EU Copyright Law .....                                    | 24 |
| 3.1.5 EU Case Law .....                                         | 27 |
| Decision 24-D-03 of 15 March 2024 .....                         | 28 |
| <i>LAION v Robert Kneschke</i> .....                            | 30 |
| <i>Like Company v Google Ireland</i> .....                      | 31 |
| 3.2 United States .....                                         | 33 |
| 3.2.1 AI Governance at the Federal Level .....                  | 34 |
| 3.2.2 AI Governance at the State Level .....                    | 36 |
| 3.2.3 US Copyright Law .....                                    | 38 |
| 3.2.4 Fair Use Doctrine .....                                   | 41 |
| First Factor .....                                              | 42 |
| Second Factor.....                                              | 43 |
| Third Factor .....                                              | 43 |
| Fourth Factor.....                                              | 44 |
| 3.2.5 Trademark Law .....                                       | 45 |
| 3.2.6 US Case Law.....                                          | 46 |
| <i>Getty Images (US) Inc &amp; Ors v Stability AI Ltd</i> ..... | 47 |
| <i>The New York Times v Microsoft and OpenAI</i> .....          | 49 |
| <i>Andersen v Stability AI</i> .....                            | 51 |
| <i>Bartz v Anthropic PBC</i> .....                              | 52 |
| 3.3 International Guidelines on Artificial Intelligence .....   | 54 |
| 3.3.1 The Council of Europe .....                               | 55 |
| 3.3.2 Additional Guidelines: OECD, UN & G7 .....                | 56 |
| 3.3.3 International Copyright Law .....                         | 58 |
| 4. <i>Policy Approaches</i> .....                               | 59 |

|       |                                                                               |    |
|-------|-------------------------------------------------------------------------------|----|
| 4.1   | Comparative analysis of jurisdictions with and without a legal framework..... | 59 |
| 4.1.1 | China .....                                                                   | 60 |
| 4.1.2 | United Kingdom.....                                                           | 62 |
| 4.2   | Policy recommendations based on legal gaps.....                               | 65 |
| 5.    | <i>Conclusion</i> .....                                                       | 68 |
|       | <i>Bibliography</i> .....                                                     | 71 |

## **List of abbreviations**

**AI** – Artificial Intelligence

**Autorité de la concurrence** – French Competition Authority

**CAC** – Cyberspace Administration of China

**CDSM** – Directive on Copyright in the Digital Single Market

**CJEU** – Court of Justice of the European Union

**CMI** – Copyright Management Information

**DMCA** – Digital Millennium Copyright Act

**EO** – Executive Order

**EU** – European Union

**FLOP** – Floating point operations

**GenAI** – Generative Artificial Intelligence

**GPAI** – General-Purpose AI model

**GPT** – Generative Pre-Trained Transformer

**IIC** - International Review of Intellectual and Competition Law

**InfoSoc Directive** – Information Society Directive

**Interim Measures for Management of Generative Artificial Intelligence Services** – AI Measures

**IP** - Intellectual Property

**LLMs** – Large Language Models

**ML** – Machine Learning

**MPs** – Members of Parliament

**NY** – New York

**NYT** – New York Times

**RAG** – Retrieval-Augmented Generation

**RMI** – Rights Management Information

**RNN** – Recurrent Neural Network

**TDM** – Text and Data Mining

**TMs** – Technological Measures

**TRIPS** – Trade-Related Aspects of Intellectual Property Rights

**UK** –United Kingdom

**UN** – United Nations

**Urheberrechtsgesetz** - Act on Copyright and Related Rights

**US** – United States

**WCT** – World Intellectual Property Organisation Copyright Treaty

**WIPO** – World Intellectual Property Organisation

**WTO** – World Trade Organisation

# Intellectual Property Law implications in the EU and US when training Generative AI models

## 1. Introduction

As technology continues to advance, recent developments in artificial intelligence (“AI”), particularly generative AI, have raised concerns regarding the intellectual property (“IP”) landscape. The increasing demand for data to train AI models has sparked concerns about certain data collection methods and their subsequent impact on the protection of IP rights. Data scraping involves the automated extraction of information from sources such as websites or databases.<sup>1</sup> It is a frequently used method of data collection to train generative AI models, and the focus of many IP rightsholders’ concerns.

While the technology surrounding generative AI has been gradually developing since the 1960s, commonly recognised models, such as ChatGPT or Gemini, are a much newer concept. OpenAI introduced ChatGPT in November 2022. Since then, it and its similar models have accomplished a new level of AI.<sup>2</sup> This has been achieved through developers’ continuous effort to improve their platforms, subsequently leading to an increasing demand for data to train the models. The efficacy of large language models (“LLMs”) relies heavily on accessing vast data sets to incorporate information and learn from their patterns of words. “The latest models of GPT are trained on trillions of words”<sup>3</sup>. This puts into perspective the sheer amount of data LLMs require to create new content. This is “a data set so big that it would be ‘the equivalent of a Microsoft Word document that is over 3.7 billion pages long’.”<sup>4</sup>

Numerous data collection methods are adopted to train and inform AI models. Data scraping, also known as web scraping, is one of the most common methods used. This comprises information from various sources, including databases, third-party websites and social media platforms, through automated extraction. “While data scraping for AI systems and similar practices are not new, its use has significantly increased with the rapid growth of generative

---

<sup>1</sup>OECD, Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping (2025) OECD Artificial Intelligence Papers, No. 33 p9

<sup>2</sup>Keith Foote, ‘A Brief History of Generative AI’ (Dataversity, 5 March 2024) <<https://www.dataversity.net/a-brief-history-of-generative-ai/>> accessed 30 June 2025

<sup>3</sup>Audrey Pope, ‘NYT v. OpenAI: The Time’s About -Face’ (Harvard Law Review, 10 April 2024) <<https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/>> accessed 26 May 2025

<sup>4</sup>Pope (n3)

AI”.<sup>5</sup> When “undertaken responsibly, AI data scraping can provide access to diverse and vast amounts of data which is essential to advance AI research and innovation”.<sup>6</sup> However, creators and owners of IP rights can be negatively affected by this when it is conducted with payment or consent of rightsholders.<sup>7</sup> It should, therefore, be viewed as a careful balance between advancing technology and safeguarding intellectual property rights.

There is much regulatory uncertainty surrounding the area, which has led to litigation arising to determine whether web scraping amounts to an IP infringement. There are several cases on this topic, including the New York Times’<sup>8</sup> (“NYT”) dispute with OpenAI and Microsoft. The central concerns of the NYT case are the alleged unauthorised and unlicensed use and reproduction of their work during the training of OpenAI’s model, ChatGPT. This is exacerbated by the fact that occasionally models memorise work and will repeat it almost verbatim. LLMs may also produce ‘synthetic’ search results, which essentially allow users to circumvent the NYT paywall.

In this instance and others, judges are beginning to consider whether the US copyright law’s fair use doctrine applies, providing tech companies with immunity from the main allegations.<sup>9</sup> While many of these cases are currently ongoing, it will be necessary to discuss them further in this thesis as their judgments “could have a significant impact on the relationship between generative AI and copyright law”<sup>10</sup>.

Furthermore, this is an issue that has not gone unnoticed. In 2023, G7 leaders identified IP rights infringements as one of the major risks deriving from generative AI.<sup>11</sup> The OECD also reported that “generative AI poses several new policy challenges to current IP frameworks.

---

<sup>5</sup>Cade Metz, Cecelia Kang, Sheera Frenkel, Stuart Thompson, and Nico Grant, ‘How Tech Giants Cut Corners to Harvest Data for A.I.’ (New York Times, 8 April 2024)

<<https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html>> accessed 2 July 2025

<sup>6</sup>OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping’ (2025) OECD Artificial Intelligence Papers No. 33 p13

<sup>7</sup>OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping’ (2025) OECD Artificial Intelligence Papers No. 33 p9

<sup>8</sup>*The New York Times Co v Microsoft et al* 1:23-cv-11195 (S.D.N.Y.)

<sup>9</sup>Blake Brittain, ‘Judge explains order for New York Times in OpenAI copyright case’ (Reuters, 4 April 2025) <<https://www.reuters.com/legal/litigation/judge-explains-order-new-york-times-openai-copyright-case-2025-04-04/>> accessed 3 July 2025

<sup>10</sup>Audrey Pope, ‘NYT v. OpenAI: The Time’s About -Face’ (Harvard Law Review, 10 April 2024) <<https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-times-about-face/>> accessed 26 May 2025

<sup>11</sup>OECD, ‘G7 Hiroshima Process on Generative Artificial Intelligence (AI)’ (2023) Japanese G7 Presidency and the G7 Digital and Tech Working Group p3

Governments are taking note globally and beginning to reflect and act on how such challenges should be addressed”.<sup>12</sup>

An issue with IP data scraping is that the legal landscape surrounding it is complex and rapidly evolving. “Existing IP laws, many predating modern AI practices, differ across jurisdictions, complicating their application.”<sup>13</sup> In many jurisdictions, this has raised the question as to whether it infringes IP rights, most notably through breach of copyright allegations. However, the issue lacks clear parameters, raising the argument of whether exceptions such as text and data mining (“TDM”) and fair use provisions are applicable.

As aforementioned, the issue lacks legislative guidance surrounding this topic, which has led to litigation emerging from various jurisdictions, including the United States (“US”) and the European Union (“EU”). Technology often develops quicker than legislation can be implemented. Jurisdictions currently without a legal framework rely on the judiciary to maintain a balance between IP protection and scientific progression until regulations are enacted. This thesis will examine the role of data scraping when training generative AI, primarily in the EU and the US. This will be accomplished by assessing the current legal framework, or lack thereof and by analysing the various case law.

## **2. Generative Artificial Intelligence**

This chapter will provide insight into generative AI through its definition and by outlining its applications. The chapter will then continue to the topic of web scraping to provide an understanding of how generative AI operates, why web scraping is an integral part of its training and how this affects authors of IP protected works.

### **2.1 Defining Generative AI and Outlining Its Applications**

With the emergence of new technology, it is no easy feat to define something that is continually evolving. Therefore, it is notable that an initial definition of generative AI may not sufficiently cover its later iterations.<sup>14</sup> In 2020, IBM described generative AI as “a class of machine learning

---

<sup>12</sup>OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping’ (2025) OECD Artificial Intelligence Papers No. 33 p12

<sup>13</sup>OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping’ (2025) OECD Artificial Intelligence Papers No. 33 p9

<sup>14</sup>Tom Taulli, *Generative AI: How ChatGPT and Other AI Tools Will Revolutionise Business* (1<sup>st</sup> edn, Apress L. P. 2023) p3

technology that learns to generate new data from training data.”<sup>15</sup> Similarly, Sequoia Capital stated that “this new category is called ‘Generative AI’, meaning the machine is generating something new rather than analysing something that already exists.”<sup>16</sup>

Furthermore, the EU and the US have outlined the concept of generative AI through their respective legislations, the AI Act<sup>17</sup> and the National Artificial Intelligence Initiative<sup>18</sup>. The EU AI Act classifies generative AI as general-purpose AI models (“GPAI”). The Act defines such as “an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market.”<sup>19</sup> Whereas, the National Artificial Intelligence Initiative defines generative AI as, “the class of AI models that emulate the structure and characteristics of input data in order to generate derived synthetic content. This can include images, videos, audio, text, and other digital content.”<sup>20</sup>

A distinguishing factor of generative AI is its ability to create work. This is a unique characteristic for AI models; up until this point, their strengths lay in analysing data sets and identifying patterns for various applications, such as detecting spam or fraud. This has been referred to as traditional AI or ‘Analytical AI’.<sup>21</sup> Examples of traditional AI models include virtual assistants such as Alexa, chatbots and streaming platforms’ recommendation systems. “Up until recently, machines had no chance of competing with humans at creative work – they were relegated to analysis and rote cognitive labour.”<sup>22</sup>

---

<sup>15</sup>Stephanie Houde, Vera Liao, Jacquelyn Martino, Michael Muller, David Piorkowski, John Richards, Justin Weisz and Yunfeng Zhang, ‘Business (mis)Use Cases of Generative AI’ (2020) 2848 HAI-GEN 1

<sup>16</sup>Sonya Huang, Pat Grady and GPT-3, ‘Generative AI: A Creative New World’ (Sequoia, 19 September 2022) <<https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/>> accessed 02 May 2025

<sup>17</sup>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) [2024] OJ L, 2024/1689 (“AI Act”)

<sup>18</sup>National Artificial Intelligence Initiative of 2020, 15 USC 9401

<sup>19</sup>AI Act, art 3(63)

<sup>20</sup>National Artificial Intelligence Initiative Act, sec 3(p)

<sup>21</sup>Huang *et al* (n16)

<sup>22</sup>Huang *et al* (n16)

With these new advances, generative AI can produce a wide range of content including text, images, videos and audio. The technologies underpinning generative AI, enabling it to provide these improved outputs, are rather new. It has taken decades of research to create the technologies which established generative AI. A gradual development and improvement in technologies can be seen during the 1980s to 1990s with recurrent neural networks (“RNN”). RNN is a form of Machine Learning (“ML”) that detects sequential data and was modelled after the human brain. While it can generate text, it was more limited in its ability than current generative AI models and was not easily scaled up.<sup>23</sup>

Significant improvements in generative AI functions occurred with the introduction of the transformer architecture by Google researchers in 2017 and in around 2019 with developments in generative pre-trained transformer (“GPT”) models. With the introduction of the transformer model sequence, whole sentences could be processed rather than analysing on a word-by-word basis. Using such methods as well as a mathematical technique called attention or self-attention, GPT models have become highly scalable, faster to train and more efficient in understanding context.<sup>24</sup>

The most recognisable form of generative AI is ChatGPT, which is freely available online. Users provide it with a few short lines of instructions, referred to as a prompt. An answer will then be generated in response. Systems such as ChatGPT are trained on a large data set of examples. Through this training, the model learns patterns and structures in the data. Once the system has been trained, it can create new content based on the examples from its data set, allowing the model to generate human-like text and generate realistic images.<sup>25</sup> This data set is often referred to as ‘training data’, which is defined by the EU AI Act as “data used for training an AI system through fitting its learnable parameters”.<sup>26</sup>

Information for data sets is extracted from sources such as databases, social media platforms and third-party websites. “The progress in this area has been swift, with models becoming increasingly large and complex. For instance, Large Language Models such as GPT-4 from

---

<sup>23</sup>Laurie Harris, ‘Generative Artificial Intelligence: Overview, Issues and Considerations for Congress’ (2025) IF12426 Library of Congress 1

<sup>24</sup>Harris (n 23) p1

<sup>25</sup>Tom Taulli, *Generative AI: How ChatGPT and Other AI Tools Will Revolutionise Business* (1<sup>st</sup> edn, Apress L. P. 2023) p8

<sup>26</sup>AI Act, art 3(29)

OpenAI have been trained on vast amounts of text data and can perform a wide range of tasks without task-specific training. This demonstrates the flexibility and potential of GenAI to impact various fields, from creative writing to software development.”<sup>27</sup>

The availability of large amounts of data and the capacity of their language models have played a critical role in advancing generative AI technologies. The language of LLMs is often modelled off of training data, which consists of millions or billions of parameters that determine how inputs are transformed into outputs through numbers in the model. Generative AI models then seek to mimic the appearance and style underlying training data.<sup>28</sup> This has been the source of contention for creators of IP protected works. Using such IP works as part of the parameters without the creators’ permission, as well as generating outputs that compete with or displace their work, has led to litigation arising from IP infringement allegations.

Furthermore, the use of IP-protected works is an issue that has the potential to extend beyond a dispute between developers and creators. Incorporating AI tools into various businesses has become increasingly popular as generative AI is “becoming not just faster and cheaper, but better in some cases than what humans create by hand”<sup>29</sup>. Therefore, the World Intellectual Property Organisation (“WIPO”) noted, “The risk is not limited to AI developers but potentially extends to users of generative AI tools. In many countries, liability for vicarious forms of IP infringement, such as making a copy of a copyright work, does not depend on the intention or knowledge of the alleged infringer.”<sup>30</sup>

Current pending litigation has been limited to creators and developers. However, WIPO’s report seeks to provide businesses with guidance for using AI tools, as it highlights that while AI presents substantial opportunities for various fields, it is also important to understand IP risks.<sup>31</sup> This uncertainty as to the extent of IP rightsholders’ work and the degree to which developers can use protected works as training data for AI models is why litigation has arisen, and discussions have emerged regarding the need for a legal framework to provide guidance in this area.

---

<sup>27</sup>Ryan W. White and Chirag Shah, *Information Access in the Earlier Era of Generative AI* (1<sup>st</sup> edn, Springer 2024) p3

<sup>28</sup>Harris (n23) p1

<sup>29</sup>Huang *et al* (n16)

<sup>30</sup>IP and Frontier Technologies Division, ‘Generative AI Navigating Intellectual Property’ (2024) WIPO Conversation, 8<sup>th</sup> Session p7

<sup>31</sup>IP and Frontier Technologies Division (n30) p3

## 2.2 What is data scraping?

“The collection of data from websites at great scale – so-called data scraping – is the foundation for ChatGPT and most other Generative AI (“GenAI”) tools.”<sup>32</sup> Data scraping is the essence of generative AI models, as “the data needed for AI training is in most cases obtained by this means.”<sup>33</sup> Without access to massive amounts of information through data scraping, generative AI would not be able to generate outcomes, as it currently does.

There are two different ways in which pre-trained generative AI models are associated with data scraping. Firstly, in their pre-training process, where substantial amounts of data are used for deep learning, the only way this can be obtained is through data scraping. Secondly, they are proficient in data scraping, allowing the systems to operate and draft data scraping software.<sup>34</sup> The use of such technology has the potential to lead to both significant productivity gains, but also runs the risk of IP infringements and data breaches.<sup>35</sup>

Generative pre-trained transformers, with the ‘P’ referring to the pre-training aspect of the process, involve access to vast amounts of data known as the ‘training data set’ collected from the internet. This data is used to inform the deep learning of the system. Essentially, it learns patterns from the information it gathers, without which the transformer system could not generate outcomes. This deep learning is rooted in the reading of data sets, which means the purpose of the data scraping is to data mine, not broadcast or store content.<sup>36</sup> The issue with the process is that, often, copyrighted material is scraped along with open-sourced data, raising IP infringement concerns.

There is a lack of transparency surrounding the training data used by developers for LLM models. Even those termed open source, implying information about the model is freely available for public use, are not as upfront about the data as they portray. Material deriving

---

<sup>32</sup>Qian Li and Konrad Kollnig, ‘Data Scraping for the Training of Generative AI: Lessons from Chinese Case Law and Regulation’ (2024) 25(2) Computer Law Review International 33

<sup>33</sup>Javier Torre de Silva and López de Letona, ‘The Right to Scrape Data on the Internet: From the US Case *hiQLabs, Inc. v. LinkedIn Corp.* to the ChatGPT Scraping Cases: Differences Between US and EU Law’ (2024) 5 Global Privacy Law Review 5

<sup>34</sup>de Silva *et al* (n33) p5

<sup>35</sup>Li *et al* (n32) p33

<sup>36</sup>de Silva *et al* (n33) p5

from open-access journals such as Frontiers and PLOS is recognised as prominently featuring in the C4 data set, according to a study by the Allen Institute for Artificial Intelligence and the Washington Post. This C4 data set has been used in several LLMs, including Meta’s generative AI model Llama.<sup>37</sup> This C4 data set is significant as it was one of the first of its kind, designed specifically for large language model pre-training.<sup>38</sup>

However, “It is also widely suspected that, just as copyrighted books have been used to train LLMs, so have non-open-access research papers.”<sup>39</sup> The use of such materials has increasingly raised concerns, evident from the upsurge in cases about generative AI training data sets and whether it is permissible under current laws. WIPO has stated that “it is unclear whether collecting data or using them to create LLM outputs is considered copyright infringement, or whether these activities fall under one of several exemptions, which differ by jurisdiction.”<sup>40</sup>

Due to this uncertainty, several AI developers have found themselves embroiled in lawsuits focusing on different angles of data scraping<sup>41</sup>. A landmark case in this area is *Andersen v Stability AI*<sup>42</sup>. Andersen, along with several other artists, filed a copyright infringement claim against a number of AI companies, including Stability AI. The focus of the case is the artists’ work, which was scraped from the internet and compiled in the LAION data set to assist companies in developing their AI image generator. “While there had been prior lawsuits by copyright owners against AI developers, this case was among the first where creatives united to challenge the use of their works for AI training, with many similar cases following shortly after”.<sup>43</sup>

Much of the litigation in this area is still ongoing; thus, the legality of data scraping remains unresolved. In light of this, the WIPO has provided guidelines to mitigate risks, with

---

<sup>37</sup>The International Journal of Science, ‘AI must play fair when they use academic data in training’ (2024) 632 Nature 953

<sup>38</sup>Joe El Khoury, ‘How Have Pre-Training Data sets for Large Language Models Evolved?’ (Medium, 13 July 2024) <<https://medium.com/@jelkhoury880/how-have-pre-training-data-sets-for-large-language-models-evolved-13d74c01f8e8>> accessed 14 July 2024

<sup>39</sup>The International Journal of Science (n37) p953

<sup>40</sup>The International Journal of Science (n37) p953

<sup>41</sup>Li *et al* (n32) p33

<sup>42</sup>*Andersen v Stability AI* 3:23-cv-00201 (N.D. Cal)

<sup>43</sup>Zack Schor, ‘Andersen v Stability AI: The Landmark Case Unpacking the Copyright Risks of AI Image Generators’ (NYU Journal of Intellectual Property & Environmental Law, 2 December 2024) <<https://jipel.law.nyu.edu/andersen-v-stability-ai-the-landmark-case-unpacking-the-copyright-risks-of-ai-image-generators/>> accessed 05 July 2025

suggestions including using licensed, public domain or the user's own training data as well as implementing code that informs tools as to what data can and cannot be scraped.<sup>44</sup>

It is, however, evident from the various cases that there is much uncertainty regarding data scraping, particularly as to whether it is permissible under EU text and data mining exemptions and US fair use criteria. Applicable legal framework and exemptions to such will be assessed in the next section of this thesis, primarily looking at the legal framework in the EU and the US, followed by a discussion of the case law that has arisen in this area.

### 3. Legal Landscape

AI has been reshaping industries worldwide, leading to an increasing need for effective AI governance and necessitating a comprehensive understanding of the global frameworks applying to AI technology. However, under the current fast-pasted yet fragmented landscape, this can prove difficult.<sup>45</sup> In most jurisdictions, the legislators have sought to strike a balance between implementing rules and protections against potential harms while encouraging AI innovation and investment. However, jurisdictions have taken substantially different approaches, leading to an inconsistent AI regulatory environment.<sup>46</sup>

Given the transformative nature of AI technology, a pattern of implementation has emerged and been adopted at various rates across jurisdictions. This often involves beginning with an ethics policy or national strategy instead of legislating from the outset.<sup>47</sup> The development of such policies and subsequent legislation will be discussed throughout this chapter. The scope of such frameworks and the extent to which this applies to data scraping when training generative AI will be the primary focus.

With the rapid nature with which AI develops, governments and regulatory bodies have had to act quickly to ensure their frameworks do not become obsolete. Furthermore, a number of

---

<sup>44</sup>IP and Frontier Technologies Division (n30) pp6-16

<sup>45</sup>Bird & Bird, 'AI Regulatory Horizon Tracker' (Bird & Bird, 30 April 2025) <<https://www.twobirds.com/en/capabilities/artificial-intelligence/ai-legal-services/ai-regulatory-horizon-tracker>> accessed 06 July 2025

<sup>46</sup>White & Case, 'AI Watch: Global Regulatory Tracker' (White & Case, 2025) <<https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker#home>> accessed 06 July 2025

<sup>47</sup>iapp25, 'Global AI and Policy Tracker' (iapp 25, May 2025) <<https://iapp.org/resources/article/global-ai-legislation-tracker/>> accessed 06 July 2025

international organisations, including the Council of Europe, the OECD, the UN and the G7, have reacted to this technological shift by publishing their own AI frameworks. There has, however, been a scramble to keep pace with technological advancement, and emerging signs indicate that it may prove difficult to do this in line with the rate of development.<sup>48</sup>

Another significant challenge is that emerging regulations take various forms and adopt different conceptual approaches. As a result, there is no uniform legal structure for the overall framework. For instance, the EU AI Act is a regulation that applies directly to all EU member states without the need for national implementation, whereas compliance obligations greatly differ in the US, where there is a combination of White House Executive Orders, federal and state initiatives, as well as actions by regulatory agencies. This highlights the lack of consistency when it comes to the conceptual approach: where certain frameworks are legally binding, some are enforced by regulators; certain regulations are sector-specific, others apply across all sectors, while others are merely guidelines or recommendations.<sup>49</sup> Therefore, the approach to IP implications arising from data scraping when training generative AI will differ from jurisdiction to jurisdiction.

As previously outlined, data scraping for use in training generative AI has become a subject of growing concern amongst IP rightsholders. Due to the legal uncertainty surrounding whether web scraping constitutes an IP infringement, increasing legal disputes have emerged querying the judiciary on the scope of rights. Litigation is primarily focused on copyright, with some cases also including trademark elements. However, WIPO has noted that there is potential for disputes to extend beyond these areas to include patents, industrial designs and database rights.<sup>50</sup> The main focus of this chapter will be AI and IP legislation. The IP regulations will primarily focus on copyright, with some consideration also placed on trademark law, followed by a review of current cases in this area.

---

<sup>48</sup>White & Case (n46)

<sup>49</sup>White & Case (n46)

<sup>50</sup>IP and Frontier Technologies Division, 'Generative AI Navigating Intellectual Property' (2024) WIPO Conversation, 8<sup>th</sup> Session p6

### 3.1 European Union

In April 2021, the first EU Artificial Intelligence law was proposed by the European Commission.<sup>51</sup> Since then, the European Union has indisputably introduced a landmark legislation governing Artificial Intelligence. The introduction of the EU AI Act has been considered a pioneering endeavour with the key aim of becoming the global hub for trustworthy, human-centric AI.<sup>52</sup> “The AI Act, officially published in July 2024 and entered into force in August 2024, is the first comprehensive framework on AI worldwide.”<sup>53</sup>

An AI Liability Directive was initially proposed to complement the AI Act, aiming to address the inconsistencies in liability regimes across the EU. However, the Commission withdrew the draft in February 2025 after citing a lack of consensus on core issues.<sup>54</sup>

The AI Act creates a harmonised legal framework in which the rules are outlined in the regulation itself. This leaves little scope for national legislation, therefore, implementing an approach that ensures all Member States uniformly protect EU values.<sup>55</sup> The AI Act’s ambitious objective is to foster trustworthy AI within the EU. It seeks to achieve this by supporting innovation while ensuring the deployment and development of AI systems respect safety, fundamental rights, democracy and the rule of law.<sup>56</sup>

The Act exemplifies “a significant, albeit complex, effort to bridge the regulatory gap. It proposes a regulatory framework designed to govern AI while addressing the unique challenges posed by this rapidly evolving technology.”<sup>57</sup> The goal of introducing the Act was to ensure a well-functioning internal market for AI systems. The Commission recognised that AI systems have the potential to bring innovation, societal benefits, economic growth and global competitiveness. However, they also acknowledged that certain aspects of AI systems could

---

<sup>51</sup>European Parliament, ‘EU AI Act: first regulation on Artificial Intelligence’ (European Parliament, 08 June 2023) <[<sup>52</sup>White & Case \(n46\)](https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence#:~:text=In%20April%202021%2C%20the%20European,risk%20they%20pose%20to%20users.>intelligence#:~:text=In%20April%202021%2C%20the%20European,risk%20they%20pose%20to%20users.>” accessed 09 July 2025</a></p></div><div data-bbox=)

<sup>53</sup>Francesca Palmiotto, ‘The AI Act Roller Coaster: The Evolution of Fundamental Rights Protection in the Legislative Process and the Future of the Regulation’ (2025) 16 European Journal of Risk Regulation 771

<sup>54</sup>White & Case (n46)

<sup>55</sup>Marco Almada and Nicolas Petit, ‘The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights’ (2025) 62 Common Market Law Review 90

<sup>56</sup>Palmiotto (n52) p2

<sup>57</sup>Stanley Greenstein and Mauro Zamboni, ‘Navigating the legislative dilemma: evaluating the EU AI Act’s approach to regulating emerging technologies’ (2025) The Theory and Practice of Legislation 2

pose risks, leading to uncertainty and potentially a slower uptake in their use. Thus, the Commission responded with legislative action, through the AI Act, addressing both risks and benefits adequately.<sup>58</sup>

The Act was designed as a product safety instrument, which was approached with risk-based regulation. This essentially assumes that classes of AI practices and systems can be classified into levels of unacceptable or high-risk.<sup>59</sup> General-purpose AI was not explicitly referenced in the first draft of the Act. It was also not excluded per se. If GPAI entailed a high-risk purpose, it then had to comply with the high-risk requirements.<sup>60</sup>

However, as technology continues to rapidly advance, it is increasingly outpacing the ability of traditional legislative frameworks to regulate its use effectively. This is made evident by the often unpredictable and dynamic nature of technological innovation, creating a regulatory gap.<sup>61</sup> The introduction of ChatGPT by OpenAI exemplifies this fast-paced landscape.<sup>62</sup> The lack of mention surrounding general-purpose AI in the April 2021 draft of the AI Act highlighted the challenges in regulating emerging technologies. The dramatic public attention received by generative AI with the launch of ChatGPT made this clear.<sup>63</sup> This resulted in extensive negotiations on the approach to regulating GPAI.<sup>64</sup> In line with concerns, the draft text was changed to include specific regulations for generative AI, which includes certain restrictions on general-purpose AI models and imposes transparency obligations.<sup>65</sup>

This chapter provides insight into the AI Act and outlines the provisions most relevant to generative AI, with a focus on Chapter V of the AI Act, which specifically regulates general-purpose AI. Furthermore, it will aim to examine the value of safeguards put in place for IP

---

<sup>58</sup>European Commission, 'Artificial Intelligence - Questions and Answers' (2024) European Commission QANDA/21/1683 p1

<sup>59</sup>Almada *et al* (n55) pp3-4

<sup>60</sup>Nidia Andrade Moreira, Pedro Miguel Freitas, and Paulo Novais, 'The AI Meets General Purpose AI: The Good, The Bad and The Uncertain' (2023) 14116 Lecture Notes in Computer Science 157

<sup>61</sup>Greenstein *et al* (n57) p2

<sup>62</sup>Palmiotto (n53) p12

<sup>63</sup>Greenstein *et al* (n57) p2

<sup>64</sup>Oskar Gstrein, Noman Haleem and Andrej Zwitter, 'General Purpose AI regulation and the European Union Act' (2024) 13(3) Internet Policy Review 4

<sup>65</sup>Desara Dushi, 'How is ChatGPT regulated by the EU AI Act: Reflections on Higher Education' (Global Campus of Human Rights, 05 September 2024) <<https://www.gchumanrights.org/preparedness/how-is-chatgpt-regulated-by-the-eu-ai-act-reflections-on-higher-education/>> accessed 10 July 2025

rightsholders regarding IP protected works used in training data through reference to relevant legislation and case law.

### 3.1.1 Objectives and Scope of the AI Act

As aforementioned, the AI Act generally takes a risk-based perspective. This implements an approach where different requirements are applied depending on the level of risk, rather than focusing on specific sectors such as finance or insurance.<sup>66</sup> “Risk regulation (“the privileged methodological tool for regulating risks in Europe”) offers a guide which then requires enforcement for understanding the actual risk and the related planning and type of intervention.”<sup>67</sup>

The application of the Act is split into four different classifications: unacceptable risk, high risk, limited risk and minimal or no risk. AI systems considered an unacceptable risk are banned for posing a clear threat to livelihoods, safety and people’s rights.<sup>68</sup> Article 5 of the Act<sup>69</sup> outlines prohibited AI practices, namely AI that is harmful based on manipulation and deception, exploitation of vulnerabilities, social scoring, etc. High-risk AI systems are extensively outlined in Article 6 of the Act<sup>70</sup> and include any system that can pose risks to safety, health or fundamental rights, such as AI safety components used in critical infrastructure, which could place citizens’ lives and health at risk. This category is subject to strict obligations before they are placed on the market, which include appropriate human oversight measures and adequate risk assessment.

Limited risk refers to risks with exigencies for transparency and specific disclosure obligations around their use. For instance, informing someone that they are interacting with a chatbot. The vast majority of systems in the EU currently fall into the minimal or no risk category. These include spam filters and AI-enabled video games for which no rules have been introduced.<sup>71</sup>

---

<sup>66</sup>Martin Braun, Anne Valley, Itsiq Benizri and Nancy Stephen, ‘Navigating Generative AI Under the European Union’s Artificial Intelligence Act’ (WilmerHale, 02 October 2024)

<<https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20241002-navigating-generative-ai-under-the-european-unions-artificial-intelligence-act>> accessed 10 July 2025

<sup>67</sup>Nicoletta Rangone and Luca Megale, ‘Risks Without Rights? The EU AI Act’s Approach to AI in Law and Rulemaking’ (2025) European Journal of Risk Regulation 10

<sup>68</sup>European Commission, ‘AI Act’ (European Commission, 03 June 2025) <<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>> accessed 10 July 2025

<sup>69</sup>AI Act, art 5(1)

<sup>70</sup>AI Act, art 6

<sup>71</sup>European Commission (n68)

Generative AI does not easily fall into one of the four categories since, by definition, it is not limited to a specific purpose, preventing it from slotting into a specific risk category.<sup>72</sup>

Of particular importance for this thesis is Chapter V of the Act<sup>73</sup>, which governs general-purpose AI. GPAI, also known as a foundational model, are defined by Article 3(63)<sup>74</sup> as including models trained on large-scale data using self-supervision, that show broad general capabilities, can handle diverse tasks effectively, regardless of how they are marketed, and can be integrated into various downstream systems or applications.<sup>75</sup> Recitals 98 and 99 of the Act<sup>76</sup> state that GPAI is also determinable by the number of parameters, including models with at least a billion parameters and trained with large amounts of data. Large generative AI models are a typical example of this. This may include models with a single modality of texts, audio, images or video, if it is flexible enough.<sup>77</sup>

GPAIs are further classified in the Act into AI models or AI systems. A GPAI system is an AI system that is based on a general-purpose AI model and can provide a variety of functions, both for direct use and integration into other AI systems.<sup>78</sup> For example, Generative AI, which can create, generate, and develop images, code texts, etc., is a subset of GPAI systems. Furthermore, GPAI systems can be used or integrated as components into high-risk systems, requiring providers to comply with such obligations.<sup>79</sup>

General-purpose AI risk classification is divided into two groups: GPAI with systematic risk and GPAI without systematic risk. Systematic risks are exposure to large-scale harm from the most advanced models or from models which have the equivalent impact.<sup>80</sup> This is expanded in the recital<sup>81</sup> to include, but is not limited to, the lowering of barriers for biological or chemical weapon development or the unintended issues of control on autonomous GPAI

---

<sup>72</sup>Gstrein *et al* (n64) p7

<sup>73</sup>AI Act, Chapter V

<sup>74</sup>AI Act, art 3(63)

<sup>75</sup>Dushi *et al* (n65)

<sup>76</sup>AI Act, recs 98 and 99

<sup>77</sup>European Commission, ‘General Purpose AI Models in the AI Act – Questions and Answers’ (European Commission, 12 June 2025) <<https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers>> accessed 10 July 2025

<sup>78</sup>AI Act, art 3(66)

<sup>79</sup>Dushi *et al* (n65)

<sup>80</sup>AI Act, art 3(65)

<sup>81</sup>AI Act, rec 110

models. It should be noted that models that fall below this threshold may be considered to pose a systematic risk through scalability, reach or scaffolding.<sup>82</sup>

A GPAI is considered a systematic risk if it is one of the most advanced models at the time or has an equivalent impact.<sup>83</sup> The definition is subject to change over time in line with evolving technology and societal adaptation to increasingly advanced models. There are currently only a handful of companies whose models fall into this classification.<sup>84</sup> A threshold of  $10^{25}$  floating-point operations (“FLOP”) in training for the models is used to determine the most advanced models.<sup>85</sup> This threshold will be monitored and may be updated by the AI Office.<sup>86</sup>

The scope of the Act applies to providers who place such models on the market in the EU, irrespective of whether they are located or established in the Union or within a third country.<sup>87</sup> Furthermore, this includes a natural or legal person, a public authority, an agency or other authority that develops a GPAI model or has developed such and placed it on the market, regardless of whether or not there is a charge.<sup>88</sup>

A clear exception is made for models used purely for prototyping activities, research or development.<sup>89</sup> A model is considered to be placed on the market from the first making available of it on the Union market.<sup>90</sup> This means its first distribution or use in the course of commercial activity.<sup>91</sup> For GPAI models, integration by the provider of a model into its own AI system constitutes being placed on the market unless it is purely for internal purposes, not providing a service or product to a third party, as well as in instances where a natural person’s rights are not affected, and it is not a GPAI model with systematic risk.<sup>92</sup> The extent to which general-purpose AIs were defined and set apart from the concept of AI systems highlights the need for legal certainty in the area, considering their potentially significant negative effects.<sup>93</sup>

---

<sup>82</sup>AI Act, art 51(1)(b) and Annex XIII, see also European Commission (n76)

<sup>83</sup>AI Act, art 51(1)

<sup>84</sup>European Commission (n77)

<sup>85</sup>AI Act, arts 51(1)(a) and (2)

<sup>86</sup>AI Act, art 51(3)

<sup>87</sup>AI Act, art 2(1)(a)

<sup>88</sup>AI Act, art 3(3)

<sup>89</sup>Gstrein *et al* (n64) p9

<sup>90</sup>AI Act, art 3(9)

<sup>91</sup>AI Act, art 3(10)

<sup>92</sup>AI Act, rec 97

<sup>93</sup>AI Act, rec 97

### 3.1.2 Obligations for general-purpose AI models

The AI Act came into force in August 2024, with the first obligations taking effect from February 2025.<sup>94</sup> GPAI obligations for models will apply from August 2025, whereas transparency requirements regarding AI-generated content, deep-fakes and AI-generated content come into effect from August 2026.<sup>95</sup>

Obligations placed on providers of general-purpose AI models comprise of documenting technical information about their models to be able to provide this upon request to the AI Office and competent national authorities.<sup>96</sup> This information is also required to be made available to downstream providers.<sup>97</sup> However, this does not apply to models released as open-source and free license models and their parameters, provided certain information is made publicly available, such as information on model usage and architecture.<sup>98</sup>

Furthermore, the issue of IP rights, specifically copyright, is addressed under GPAI obligations. It is outlined that providers must implement a policy to comply with EU law on copyright and related rights.<sup>99</sup> In addition to this, a sufficiently detailed summary of the content used as training data must be prepared and made publicly available.<sup>100</sup> The accountability requirements surrounding the law of copyright and related rights, and the transparency obligation to publicise content used in training data, display the proactive approach the EU is taking to issues arising from generative AI models. Discussions have been raised on the introduction of GPAI regulations and the EU's shift from proactive regulation to a more reactive approach. However, given that there has generally been a global change towards legislating AI, this decision appears to have been a clear choice.<sup>101</sup>

The introduction of these obligations in the Act has provided more clarity concerning IP infringements arising out of data used to train generative AI. While these obligations are not yet applicable, it appears the EU has made a concerted effort to put provisions in place to protect IP rights. The extent to which these obligations will be strictly enforced will become

---

<sup>94</sup>AI Act, art 113

<sup>95</sup>Braun *et al* (n66)

<sup>96</sup>AI Act, art 53(1)(a)

<sup>97</sup>AI Act, art 53(1)(b)

<sup>98</sup>AI Act, art 53(2)

<sup>99</sup>AI Act, art 53(1)(c)

<sup>100</sup>AI Act, art 53(1)(d)

<sup>101</sup>Gstrein *et al* (n64) pp16-19

clearer in the coming months. Failure to comply with the general-purpose AI model requirements will entail fines not exceeding 3% of the provider's annual total worldwide turnover or €15,000,000, whichever is higher.<sup>102</sup>

### 3.1.3 Code of Practice

The AI Act has also introduced a General-Purpose AI Code of Practice based on Article 56 of the Act. The Code of Practice was finalised on 10 July 2025 and is applicable from August 2025. It was designed to help providers comply with the general-purpose AI rules outlined in the Act. The code consists of three chapters: the first and second are Transparency and Copyright, which apply to all GPAI; the third, Safety and Security, applies only to the most advanced models.<sup>103</sup>

The Code of Practice allows organisations to exhibit compliance with the Act on a voluntary basis. Those who agree to comply will see a reduced administrative burden and more legal certainty.<sup>104</sup> The Transparency chapter provides a user-friendly Model Documentation Form to allow easy documentation of the necessary information in one place. The Copyright segment outlines practical suggestions to enact a policy complying with EU Copyright Law.<sup>105</sup> This includes measures such as drawing up and maintaining a copyright policy, extracting and reproducing only accessible content, complying with rights reservations when web crawling, mitigating copyright infringement risks and appointing a point of contact to receive complaints.<sup>106</sup>

As with the obligations in the Act, the extent to which the code will be adopted will become clearer once it comes into effect. It should also be noted that the code specifies that the Copyright chapter does not affect the application and enforcement of EU law on copyright and related works. Interpretation of such is for the courts of Member States and ultimately the Court of Justice of the EU to interpret.<sup>107</sup> The application and interpretation of the EU intellectual

---

<sup>102</sup>AI Act, art 101

<sup>103</sup>European Commission, 'General Purpose AI Code of Practice now Available' (European Commission, 10 July 2025) <[https://ec.europa.eu/commission/presscorner/detail/en/ip\\_25\\_1787](https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1787)> accessed 11 July 2025

<sup>104</sup>Caitlin Andrews, 'European Commission receives final version of General-Purpose AI Code of Practice' (iapp, 10 July 2025) <<https://iapp.org/news/a/european-commission-receives-final-version-of-general-purpose-ai-code-of-practice>> accessed 11 July 2025

<sup>105</sup>European Commission (n103)

<sup>106</sup>Code of Practice for General-Purpose AI Models (2025) Copyright Chapter

<sup>107</sup>Code of Practice for General-Purpose AI Models (2025) Copyright Chapter rec (b)

property law in the context of IP infringements when training generative AI models will be discussed in the following section.

### 3.1.4 EU Copyright Law

Copyright under EU law is governed by an array of directives and regulations. This section will focus on two key regulations: the Information Society Directive<sup>108</sup> (“InfoSoc Directive”) and the Copyright in the Digital Single Market Directive<sup>109</sup> (“CDSM Directive”). An analysis of rights conferred by the InfoSoc Directive will be discussed first, followed by an assessment of exemptions introduced by the CDSM Directive to determine whether data scraping is permitted under EU copyright laws.

The InfoSoc Directive was adopted in 2001 to harmonise copyright and related rights in light of the information society and technological developments. A key aspect of the Directive is providing protection for intellectual property law.<sup>110</sup> The Directive protects the economic interests of copyright holders, ensuring that they retain control over their work and receive any revenue derived from it. InfoSoc is therefore focused on the authorisation or prohibition of making and distributing copies, along with how they are communicated to the public.<sup>111</sup> Thus, the Directive essentially incentivises creative works, protecting both their originality and creativity.<sup>112</sup>

Article 2 of the Directive<sup>113</sup> protects creators’ reproduction rights, which must be granted exclusively to protected content, the right to permit or prevent reproduction, whether direct or indirect, temporary or permanent. This applies to a wide range of creators, from authors and film producers to broadcasting organisations.<sup>114</sup> The Directive provides further protections

---

<sup>108</sup>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 (Information Society Directive) [2001] *OJ L 167/10* (“InfoSoc Directive”)

<sup>109</sup>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 (Directive on Copyright in the Digital Single Market) *OJ L 130/1* (“CDSM Directive”)

<sup>110</sup>European Union, ‘Copyright and related rights in the information society’ (European Union, 23 September 2021) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l26053>> accessed 17 July 2025

<sup>111</sup>European Union, ‘EU copyright’ (European Union) <[https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legisum:eu\\_copyright](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legisum:eu_copyright)> accessed 17 July 2025

<sup>112</sup>Anett Pogácsás, ‘New Ways of Plagiarism? At the Meeting Point of the Specific Form of Co-Authorship and the Transparency Criteria of the EU AI Regulation’ (2025) IIC 3 <<https://link.springer.com/article/10.1007/s40319-025-01608-2>> accessed 05 July 2025

<sup>113</sup>InfoSoc Directive, art 2

<sup>114</sup>InfoSoc Directive, art 2 (a)-(e)

over the right to publicly communicate works and their availability to the public,<sup>115</sup> as well as distribution rights.<sup>116</sup> Similarly, these protections vest exclusively in the rightsholders, thereby deterring unauthorised use of their works.

The Directive provides a number of exceptions and limitations to these protections under Article 5.<sup>117</sup> This allows temporary acts of reproduction of protected works as outlined in Article 2.<sup>118</sup> It has been argued that data scraping could be classified as such and thus this exception could be applied. However, the temporary and transient nature of the data is integral for this exemption. Since long-term storage is necessary for high-quality data analysis and AI training methods, data scraping would not be permissible under Article 5(1).<sup>119</sup> Long-term storage is only admissible under certain exceptions outlined in the CDSM Directive.

A further exception is also granted for scientific research if the use is non-commercial.<sup>120</sup> Some generative AI companies focused purely on innovation may qualify for this exception, but commercial developers like OpenAI, which earn revenue from subscriptions and licensing, would be ineligible.

With the introduction of the CDSM Directive in 2019, certain exceptions were adopted to adapt to the evolving digital and global environment. The Directive was introduced to provide greater legal certainty for copyright exemptions, including text and data mining. TDM “means 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”.<sup>121</sup>

Under the Directive, text and data mining by research organisations and cultural heritage institutions is permitted for the purpose of scientific research.<sup>122</sup> The exemptions scope appears quite limited and is further constrained by the interpretation of a ‘scientific organisation’ in

---

<sup>115</sup>InfoSoc Directive, art 3

<sup>116</sup>InfoSoc Directive, art 4

<sup>117</sup>InfoSoc Directive, art 5

<sup>118</sup>InfoSoc Directive, art 5(1)

<sup>119</sup>Katharina de la Durantaye, ‘Control and Compensation. A Comparative Analysis of Copyright Exceptions for Training Generative AI’ (2025) 56(4) IIC 742, see also InfoSoc Directive, art 5(1)

<sup>120</sup>InfoSoc Directive, art 5(3)(a)

<sup>121</sup>CDSM, art 2(2)

<sup>122</sup>CDSM, art 3

Recital 12.<sup>123</sup> For the application of this Directive, scientific research covers both the natural and human sciences, while research organisations are considered to be a widely applicable term and therefore seen as generally sharing a common public interest goal or a non-profit basis.<sup>124</sup> Consequently, commercially driven AI developers would not be able to rely on this exception. Furthermore, institutions permitted to carry out TDM must have lawful access to the works or other subject matter in question.<sup>125</sup>

Beyond scientific research, TDM is permitted under Article 4 of the Directive,<sup>126</sup> provided it meets certain conditions. Reproductions and extractions carried out per Article 4<sup>127</sup> may only store the data as long as required for text and data mining. In addition, this exception or limitation will only apply if the works are not expressly reserved in a suitable way by their rightsholders. The Directive suggests a machine-readable means for publicly available content or contractual agreements as an example of an appropriate manner.<sup>128</sup> Recital 18 of the Directive expands on these TDM measures. The right of copyright holders to derive economic benefit from their work is underscored by the provision allowing them to request a licensing fee prior to exercising their opt-out.<sup>129</sup>

There was initially uncertainty surrounding whether TDM was intended for AI training uses.<sup>130</sup> However, it has been made clear through the AI Act that it is permissible for training purposes. Article 53(1)(c)<sup>131</sup> provides that not only should generative AI framework be compliant with copyright laws, but also in accordance with Article 4(3) of the CDSM. Furthermore, Recital 105 of the AI Act allows extensive use of TDM provided opt-out rights have been observed. The inclusion of TDM in the AI Act has been considered a step backwards by cultural organisations. Axel Voss, a German European Parliament member involved in drafting the CDSM, stated that TDM was not intended to be used to deal with generative AI models and has created a legal loophole. He views this as the EU prioritising support for big tech over the

---

<sup>123</sup>Artha Dermawan, 'Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese "nonenjoyment purposes"' (2024) 27 *The Journal of World Intellectual Property* 52, see also CDSM, rec 12

<sup>124</sup>CDSM, rec 12

<sup>125</sup>CDSM, art 3(1)

<sup>126</sup>CDSM, art 4

<sup>127</sup>CDSM, art 4(2)

<sup>128</sup>CDSM, art 4(3) and rec 18

<sup>129</sup>Salitorn Thongmeensuk, 'Rethinking copyright exceptions in the era of generative AI: Balancing innovation and intellectual property protection' (2024) 47(2) *The Journal of World Intellectual Property* 282

<sup>130</sup>de la Durantaye (n119) p743

<sup>131</sup>AI Act, art 53(1)(c)

protection of creative endeavours.<sup>132</sup> Furthermore, despite the introduction of relevant policies, legal ambiguity remains, as evidenced by the pending landmark case in the EU, *Like Company v Google Ireland*.<sup>133</sup>

The practicality and effectiveness of reserving rights have also been the subject of discussion. Key issues that have remained unclear include the lack of a technical standard for declaring reservations, ownership of the right, whether it can be transferred, and how rightsholders can assert their rights over online content that was shared before the existence of generative AI or without their permission and metadata.<sup>134</sup>

### 3.1.5 EU Case Law

So far, litigation in the EU has not been as prevalent as in the US. There are a number of reasons for this variance in development. One of these rationales is that the legislation in each jurisdiction reflects differing priorities. US regulations are predominantly industry-oriented, whereas the EU emphasises a rights-oriented framework, which is evident throughout the AI Act.<sup>135</sup> Additionally, copyright protection issues in the US, regarding the use of protected content when training generative AI, are currently being addressed through case law. This contrasts with the EU's approach, relying on provisions in the AI Act and CDSM.<sup>136</sup>

Another influential factor is the US's more robust AI ecosystem. The US has established itself as the global leader in AI through well-recognised machine learning models, significant private investment and considerable AI merger and acquisition activity.<sup>137</sup> This is facilitated by regulations, which emphasise AI innovation, and which will be discussed further in the following section. In contrast, AI development is not as dominant in the EU. In 2023, France was ranked in sixth place in the Global AI Leaders index, the highest of EU Member States.<sup>138</sup>

---

<sup>132</sup>Jennifer Rankin, 'EU accused of leaving 'devastating' copyright loophole in AI Act' (The Guardian, 19 February 2025) <<https://www.theguardian.com/technology/2025/feb/19/eu-accused-of-leaving-devastating-copyright-loophole-in-ai-act>> accessed 10 August 2025

<sup>133</sup>C-250/25 *Like Company v Google Ireland* [2025]

<sup>134</sup>de la Durantaye (n119) p743

<sup>135</sup>Kaigeng Li, Hong Wu and Yupeng Dong, 'Copyright Protection during the training stage of generative AI: Industry-oriented U.S. law, rights-oriented EU law, and fair remuneration rights for generative AI training under the UN's international governance regime for AI' (2024) 55 Computer Law and Security Review 2

<sup>136</sup>Li *et al* (n135) pp4-5

<sup>137</sup>Stanford HAI staff, 'Global AI Power Rankings' (Stanford University Human-Centred Artificial Intelligence, 21 November 2024) <<https://hai.stanford.edu/news/global-ai-power-rankings-stanford-hai-tool-ranks-36-countries-in-ai>> accessed 27 July 2025

<sup>138</sup>Stanford HAI staff (n137)

Therefore, the EU's less dominant stance in the AI market could be a contributing factor to the lower level of litigation.

Up until this point, Member States in the EU have dealt with IP issues at a local level in line with EU laws. The rise in litigation is a relatively recent development, suggesting that existing legislation may not adequately address all issues. Two of the first Member States to issue decisions in this area are France and Germany. In March 2024, the French Competition Authority ("Autorité de la concurrence") fined Google €250 million for breaches of EU Intellectual Property Law.<sup>139</sup> Similarly, the District Court in Hamburg published a decision on whether data sets for training generative AI constitute a legitimate form of TDM or if they breach copyright.<sup>140</sup>

### **Decision 24-D-03 of 15 March 2024**

The decision by the French Competition Authority pertains to Google and its artificial intelligence service, Gemini, formerly Bard. It is one of four decisions which have been issued on this case. The dispute stems from the transposition of the CDSM into French law, as the Creation of Neighbouring Rights for the Benefit of Press Agencies and Publishers.<sup>141</sup> The legislative framework was introduced to create more balanced conditions for negotiations between publishers, press agencies and digital platforms. The legislation recognised the decline in print press and growth in digital audiences, negatively affecting the press sector.<sup>142</sup>

The case began with interim orders against Google for failing to comply with legislation, resulting in a fine.<sup>143</sup> Subsequently, an order to comply with the initial injunctions was issued under penalty.<sup>144</sup> Following this, Google proposed several commitments to alleviate

---

<sup>139</sup>Decision 24-D-03 of 15 March 2024 regarding compliance with the commitments in Decision 22-D-13 of 21 June 2022 of the Autorité de la concurrence regarding practices implemented by Google in the press sector

<sup>140</sup>*LAION v Robert Kneschke* Hamburg District Court, 310 O.22723

<sup>141</sup>Law No 2019-775 of 24 July 2019, on the Creation of Neighbouring Rights for the Benefit of Press Agencies and Publishers

<sup>142</sup>Autorité de la concurrence, 'Related Rights: the Autorité fines Google €250 million for non-compliance with some of its commitments made in June 2022 (Autorité de la concurrence, 20 March 2024)

<<https://www.autoritedelaconcurrence.fr/en/press-release/related-rights-autorite-fines-google-eu250-million-non-compliance-some-its>> accessed 28 July 2025

<sup>143</sup>Decision 20-MC-01 of 9 April 2020 on requests for interim measures by the Syndicat des éditeurs de la presse magazine, the Alliance de la presse d'information générale and others and Agence France-Presse

<sup>144</sup>Decision 21-D-17 of 12 July 2021 on compliance with the injunctions issued against Google in Decision 20-MC-01 of 9 April 2020

competition concerns, which were accepted by the French competition authority.<sup>145</sup> However, Google was recently found to be in breach of its promise to cooperate with the monitoring of trustees, as well as four of its seven commitments. This includes a guarantee to negotiate using transparent, objective and non-discriminatory tactics,<sup>146</sup> provide sufficient information to allow press agencies and publishers to determine remuneration for related rights,<sup>147</sup> and take measures to ensure that economic relationships between Google and the parties are not affected by negotiations.<sup>148</sup>

This is connected to the training of generative AI models, as Bard incorporated press agencies and publishers' content into its foundational model without their permission. This protected content was included without permitting the concerned parties to opt out, which is required per Article 4 and recital 18 of the CDSM, which was transposed into French law. This hindered their right to negotiate remuneration for the use of their works.<sup>149</sup> Press publishers and news agencies' domain content was present in not only the foundational model but also the grounding, where the AI refers a query to Google search to answer a prompt, and display, showing the user this answer. This was conducted without rightsholders' consent.<sup>150</sup>

While it has not yet been determined whether protections under the related rights legislation extend to the use of such content in AI, the Autorité believes that at a minimum, Google breached Commitment 1, which guarantees transparency. Google failed to provide a technical solution to allow press publishers and news agencies to object without affecting the displays of their works on Google until the introduction of Google Extended. This only occurred seven months after the launch of Bard. Before this, those seeking to opt out had to insert a program prohibiting Google from crawling their content across all its services. Furthermore, Google did not contest the allegation when fined.<sup>151</sup> Although the decision does not offer a definitive stance on the legality of data scraping for generative AI training in the EU, it does provide insight into how at least one Member State views this particular intellectual property concern. "In the

---

<sup>145</sup>Cases No 19/0074F, 19/0078F and 19/0080F – Related Rights Commitments Proposal of Alphabet Inc, Google LLC, Google Ireland Limited and Google France, see also Decision 22-D-13 of 21 June 2022 regarding practices implemented in the press sector

<sup>146</sup>Cases No 19/0074F (n144), commitments 1 and 4

<sup>147</sup>Cases No 19/0074F (n144), commitment 2

<sup>148</sup>Cases No 19/0074F (n144), commitment 6

<sup>149</sup>Autorité de la concurrence (n141)

<sup>150</sup>Autorité de la concurrence (n141)

<sup>151</sup>Autorité de la concurrence (n141)

future, the *Autorité* will be particularly attentive as regards the effectiveness of opt-out systems implemented by Google.”<sup>152</sup>

### ***LAION v Robert Kneschke***

In *LAION v. Robert Kneschke*,<sup>153</sup> the Hamburg District Court ruled on whether web scraping publicly available images to create a data set for AI training constitutes a copyright infringement. The plaintiff was Robert Kneschke, a photographer who claimed the defendant, a non-profit called LAION, infringed copyright by reproducing one of his images in their data set built to train AI models.<sup>154</sup> This data set consisted of aggregated data from third parties, image URLs and textual descriptions, which were run through software to check whether the image and text corresponded. Those that did not match were filtered out, and the remaining metadata was extracted to create the data set, which was publicly available free of charge and consisted of almost 6 billion images with corresponding texts.<sup>155</sup>

One of Kneschke’s copyright protected photos was downloaded, analysed and incorporated into the data set. The watermarked image had been made available via the website of a photo agency. The website’s terms of use contained a usage reservation in English prohibiting the scraping or downloading of content by automated programs. Subsequently, the plaintiff sought to compel the defendant to refrain from using his photos in future data sets. This was not upheld by the Hamburg Regional Court.<sup>156</sup>

The crux of the dispute was the permissibility of the downloaded image, its use in the image-text comparison, and the inclusion of the photo in the data set. The Court ruled that in this context, the conduct constituted text and data mining, therefore exempting the defendant from copyright infringement. Since the TDM was conducted by a non-commercial research organisation for scientific research purposes, it was not only exempt from copyright

---

<sup>152</sup>Autorité de la concurrence (n141)

<sup>153</sup>*LAION v Robert Kneschke* Hamburg District Court, 310 O.22723

<sup>154</sup>EUIPO, ‘Germany - Hamburg District Court, 310 O.22723, *LAION v Robert Kneschke*, [27 September 24] (European Union Intellectual Property Office) <<https://www.euipo.europa.eu/en/law/recent-case-law/germany-hamburg-district-court-310-o-22723-laion-v-robert-kneschke>> 31 July 2025

<sup>155</sup>WIPO, ‘Germany DE048-j 2024 WIPO IP Judges Forum Informal Case Summary – Hamburg Regional Court, Germany [2024]: *Robert Kneschke v LAION e.V.* Case No. 310 227/23’ (WIPO) <<https://www.wipo.int/wipolex/en/text/592042>> accessed 31 July 2025

<sup>156</sup>WIPO (n155)

infringement but also did not require an opt-out option.<sup>157</sup> This was decided under Article 3 of the CDSM implemented by §60d of the Act on Copyright and Related Rights (“Urheberrechtsgesetz”)<sup>158</sup> in Germany. The court elected to apply §60d rather than §44b<sup>159</sup> for non-scientific purposes due to the non-profit nature of the company, the non-commercial purpose of the data set and the fact that it was published free of charge. It was argued that web scraping has commercial purposes as it competes with the original works, but this was dismissed by the court.<sup>160</sup>

In an *obiter dictum* statement by the court, it was held that the natural language used by the photo agency in their terms of use to reserve rights would likely be considered machine-readable. This provides some clarity to rightsholders about what constitutes ‘machine-readable’ in instances where the opt-out is applicable. Furthermore, the Chamber did not address whether the use of the defendant’s data set by a subsequent party would infringe on the plaintiff’s copyright. This case helps to provide a little insight into the implementation of the CDSM by a Member State.

### ***Like Company v Google Ireland***

Thus far, only one case has been referred to the Court of Justice of the European Union (“CJEU”), which occurred in April of this year. *Like Company v Google Ireland*<sup>161</sup> is being hailed as a landmark case on generative AI. The decision has the potential to redefine AI providers’ responsibilities in the EU and the scope of press publishers’ rights in the digital landscape.<sup>162</sup>

Like Company is a news publisher from Hungary that alleges segments of their protected articles were reproduced and made available without their consent by Gemini, Google’s generative AI. In their claim to the Hungarian court, Like Company provided an example of an

---

<sup>157</sup>WIPO (n155)

<sup>158</sup>Gesetz über Urheberrecht und verwandte Schutzrechte (“Urheberrechtsgesetz”), Act on Copyright and Related Rights version applicable since 7 June 2021

<sup>159</sup>Gesetz über Urheberrecht und verwandte Schutzrechte (“Urheberrechtsgesetz”), Act on Copyright and Related Rights version applicable since 7 June 2021

<sup>160</sup>EUIPO (n154)

<sup>161</sup>C-250/25 *Like Company v Google Ireland* [2025]

<sup>162</sup>Barry Scannell and Leo Moore, ‘CJEU to Rule on AI Chatbots and Copyright in Landmark Case Against Google’ (William Fry, May 2025) <<https://www.williamfry.com/knowledge/cjeu-to-rule-on-ai-chatbots-and-copyright-in-landmark-case-against-google/>> accessed 27 July 2025

article they published about a Hungarian singer’s plan to introduce dolphins to Lake Balaton. When Gemini was prompted by users, it would generate a detailed summary of the article, which allegedly includes significant elements of the plaintiff’s original content. The court in Budapest has referred the issue to the CJEU to interpret provisions in both the InfoSoc Directive and CDSM. The central question posed is whether the training and operation of models and their responses amount to reproduction or communication to the public under EU law.<sup>163</sup>

Four questions have been submitted to the CJEU for a preliminary ruling:

1. Whether including content in chatbot replies that reflects press articles and goes beyond a “very short extract” qualifies as communication to the public under Article 15(1) CDSM and Article 3(2) InfoSoc Directives.
2. Whether training LLMs using tokenisation and language patterns derived from protected works amounts to reproduction under Article 2 of the InfoSoc Directive.
3. If it is considered a mode of reproduction, does it qualify for the TDM exception under Article 4 of the CDSM.
4. Whether a chatbot’s response containing protected content constitutes reproduction by the AI provider requiring rightsholder authorisation.

Like Company argues that Gemini’s training and output amount to unlawful use of copyrighted content. They also claim that the chatbot’s responses go beyond what is permitted under Article 15 of the CDSM, while the systematic reproduction of their protected works falls outside the scope of TDM, as it is not for scientific research purposes and causes economic harm. Like Company has further submitted that Gemini’s responses interfere with visits to its website.<sup>164</sup>

Google responded, stating that its AI model does not store or retrieve copies of articles. It uses a method called tokenisation, which generates new text based on probabilistic modelling. Google claims any similarity between content is merely incidental or the result of a hallucination by the AI model. They are purporting to rely on the exception of temporary acts of reproduction under EU law and TDM, asserting that no new public has been reached through Gemini.<sup>165</sup>

---

<sup>163</sup>Scannell *et al* (n162)

<sup>164</sup>Scannell *et al* (n162)

<sup>165</sup>Scannell *et al* (n162)

The decision in this case will have the opportunity to provide crucial clarity on several points of law in this area. These include whether AI-generated works that mirror protected content constitute communication to the public under EU law, whether the meaning of reproduction under Article 2 of the InfoSoc Directive includes the tokenisation of works for pattern analysis, whether training carried out on publicly available information falls into research-related or commercial TDM, and whether prompts which that generate content with identifiable elements of protected works is considered infringement.<sup>166</sup>

The outcome of this case will likely have far-reaching consequences for copyright holders and AI developers. The decision is expected to clarify the limits and applicability of TDM under Article 4 CDSM. Many AI developers rely on this exception to justify their large-scale data collection. This could therefore negatively affect their web scraping activities going forward. Furthermore, if the CJEU rules that training and outputs are not covered by the exceptions, AI developers may have to obtain licenses to access content for training. On the other hand, if generative AI is considered transformative or the outputs are too indirect, developers may be able to continue to rely on this exemption, encouraging the development of LLMs in the EU.<sup>167</sup>

### 3.2 United States

The unified and comprehensive legal framework adopted by the EU greatly contrasts with the fragmented approach taken by the US to AI governance, which relies on a combination of federal initiatives, state-level legislation and voluntary industry standards.<sup>168</sup> Initiatives at the federal level have primarily been led by sector-specific guidelines and non-binding interventions that fall within the executive powers. This absence of a legally binding framework could be viewed as a policy direction or the impact of a polarised and partisan make-up of the US Congress.<sup>169</sup>

---

<sup>166</sup>Scannell *et al* (n162)

<sup>167</sup>Scannell *et al* (n162)

<sup>168</sup>Nathalie Moreno and Amanda McAllister, 'Key insights into AI regulations in the EU and the US: navigating the evolving landscape' (Kennedys, 21 January 2025) <<https://kennedyslaw.com/en/thought-leadership/article/2025/key-insights-into-ai-regulations-in-the-eu-and-the-us-navigating-the-evolving-landscape/>> accessed 12 July 2025

<sup>169</sup>Thomas Reuters Foundation, 'AI governance in the United States' (Thomas Reuters Foundation, 07 May 2025) <[https://www.trust.org/toolkit/ai-governance-for-africa-part-1/7-ai-governance-in-united-states/#\\_ftn1](https://www.trust.org/toolkit/ai-governance-for-africa-part-1/7-ai-governance-in-united-states/#_ftn1)> accessed 12 July 2025

Discussions surrounding the development of AI in recent years have centred around two components in the AI geopolitical competition: a technical and a regulatory race. With the introduction of the AI Act, the EU has positioned itself as a leading regulatory power, whereas the US has typically placed a greater emphasis on fostering innovation over implementing a legal framework.<sup>170</sup> The reluctance to govern AI highlights the US's more reactive approach to legislation, while also being influenced by the greater level of investment in AI companies within its jurisdiction. The number of AI companies funded in the US in 2022 was reported to be 1.9 times greater than the total investment in the EU and UK combined.<sup>171</sup>

### 3.2.1 AI Governance at the Federal Level

Under the Biden administration, an attempt was made to provide guidance and progress towards an enforceable legal framework. In October 2023, an Executive Order (“EO”) on Safe, Secure and Trustworthy Development and Use of Artificial Intelligence<sup>172</sup> was released. This order introduced a government-wide effort to guide the responsible development and deployment of AI through industry regulation, federal agency leadership and engagement with international parties.<sup>173</sup> However, this EO was rescinded in January 2025 under the current administration with the Executive Order Removing Barriers to American Leadership in Artificial Intelligence.<sup>174</sup> The justification by the administration for the removal of Biden’s EO is to maintain the US’s long-standing reputation as the country at the forefront of AI, which is reasoned must be achieved free from ideological bias or engineered social agendas.<sup>175</sup>

The White House Blueprint for an AI Bill of Rights<sup>176</sup> was also issued under the Biden Administration prior to the EO. The AI Bill of Rights is a non-binding document that establishes five principles to mitigate potential risks.<sup>177</sup> The principles are algorithmic

---

<sup>170</sup>Aifang Ma, ‘Regulation in pursuit of artificial intelligence (AI) sovereignty: China’s mix of restrictive and facilitative modalities’ (2024) 34 AJIC pp3-4

<sup>171</sup>Nir Kshetri, ‘Navigating EU Regulations: Challenges for US Technology Firms and the Rise of Europe’s Generative AI Ecosystem’ (2024) 57(10) IEE Computer Society 113

<sup>172</sup>Executive Order 14110 of October 30, 2023. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

<sup>173</sup>US Congress, ‘Highlights of the 2023 Executive Order on Artificial Intelligence for Congress’ (2024) US Congress Reports R47843 p2

<sup>174</sup>Executive Order 14179 of January 23, 2024. Removing Barriers to American Leadership in Artificial Intelligence

<sup>175</sup>EO 14179, sec 1

<sup>176</sup>Executive Office of the President “Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People” (October 2022)

<sup>177</sup>Tom Krantz, ‘What is the AI Bill of Rights’ (IBM, 27 September 2024)  
<<https://www.ibm.com/think/topics/ai-bill-of-rights>> accessed 12 July 2024

discrimination protections, data privacy, safe and effective systems, notice and explanation, and human alternatives. In addition to this, the administration secured voluntary commitments from 16 US AI developers such as OpenAI, Google and Meta to conform to certain standards for testing and transparency by mid-2024.<sup>178</sup>

While the Removing Barriers EO did not specifically revoke the AI Bill of Rights, under the current administration, such commitments could be seen to be rolled back. Similarly, further development of these principles and in this regulatory area seems unlikely given the perception that they are not in line with enhancing US global AI dominance.<sup>179</sup> This deregulatory stance could extend beyond federal regulations. States have been implementing legislation in response to emerging AI issues. Under the new AI policy, resisting AI industry pressure may become more difficult. An absence of federal policy direction may also result in states struggling to implement an AI governance ‘patchwork’.<sup>180</sup>

The COPIED Act<sup>181</sup> is a recent bipartisan proposal which recommends new federal transparency guidelines requiring authenticating, marking and detecting AI-generated content to hold violators accountable for their infringements. One of the goals of the Act is to promote transparency, making it easier to identify the origin of the content as well as whether it has been generated and manipulated by AI. It would give content owners the right to take legal action against those who use their content without their permission.<sup>182</sup> Despite reintroducing the Act to Congress earlier this year, it remains pending.<sup>183</sup> Establishing such an act at the federal level would have a significant impact on the rights of content creators concerning the use of their

---

<sup>178</sup>Thomas Reuters Foundation (n169)

<sup>179</sup>White & Case (n46)

<sup>180</sup>Katherine Grillaert, Matt Kennedy and Chinasa T. Okolo, ‘Risks of State-Led AI Governance in a Federal Policy Vacuum’ (Tech Policy Press, 06 February 2025) <<https://www.techpolicy.press/risks-of-state-led-ai-governance-in-a-federal-policy-vacuum/>> accessed 13 July 2025

<sup>181</sup>Content Origin Protection and Integrity from Edited and Deepfaked Media Act of 2025, S 1396

<sup>182</sup>US Senate Committee on Commerce, Science and Transportation, ‘Cantwell, Blackburn, Heinrich Introduce Legislation to Increase Transparency, Combat AI Deepfakes & Put Journalists, Artists & Songwriters Back in Control of Their Content’ (US Senate Committee on Commerce, Science and Transportation, 11 July 2024) <<https://www.commerce.senate.gov/2024/7/cantwell-blackburn-heinrich-introduce-legislation-to-combat-ai-deepfakes-put-journalists-artists-songwriters-back-in-control-of-their-content>> accessed 13 July 2025

<sup>183</sup>US Senate Committee on Commerce, Science and Transportation, ‘Cantwell, Blackburn, Heinrich Reintroduce Bipartisan Bill to Increase Transparency, Combat AI Deepfakes & Put Journalists, Artists & Songwriters Back in Control of Their Content’ (US Senate Committee on Commerce, Science and Transportation, 10 April 2025) <<https://www.commerce.senate.gov/2025/4/cantwell-blackburn-heinrich-reintroduce-bipartisan-bill-to-increase-transparency-combat-ai-deepfakes-put-journalists-artists-songwriters-back-in-control-of-their-content>> accessed 13 July 2025

work in generative AI data sets. However, considering the deregulatory stance of the current administration, it remains to be seen whether the Act passes.

### 3.2.2 AI Governance at the State Level

Therefore, given the current political climate, governing AI in the US primarily relies on State intervention. The recently enacted One Big Beautiful Bill<sup>184</sup> sought to move away from this with a 10-year federal moratorium on state regulation of AI. However, this proved a source of contention and was struck down despite major AI firms expressing interest in moving away from states' panoply of differing policies.<sup>185</sup> Thus, the US legal framework remains rather fragmented, with several states passing legislation in lieu of federal action.<sup>186</sup>

While many states have enacted various forms of regulations regarding AI output, decision-making, data processing and protection, this section will focus more specifically on policies concerning the use of IP protected works in generative AI training data. In September 2024, California lawmakers passed several AI-related bills. Given the state's status as home to many of the world's leading AI companies, the enactment of these Bills has the potential to be impactful on technology companies.<sup>187</sup> Of these Bills, the most significant for this thesis is the Generative artificial intelligence: training data transparency.<sup>188</sup>

The law is broad-reaching, encompassing developers who produce, design, code or modify substantially an AI system or service for use by the public.<sup>189</sup> Interestingly, the framework applies retrospectively to any generative AI system or service or those substantially modified, released on or after January 2022.<sup>190</sup> Beginning in January 2026, transparency requirements, which include but are not limited to publishing a high-level summary of data sets used in

---

<sup>184</sup>One Big Beautiful Bill Act of 2025, HR 492, 119<sup>th</sup> Congress

<sup>185</sup>David Morgan and David Shepardson, 'US Senate Strikes AI regulation ban from Trump megabill' (Reuters, 01 July 2025) <<https://www.reuters.com/legal/government/us-senate-strikes-ai-regulation-ban-trump-megabill-2025-07-01/>> accessed 13 July 2025

<sup>186</sup>Grant Gross, 'What IT leaders want from US AI regulation' (Cio, 04 September 2024) <<https://www.cio.com/article/3498566/what-it-leaders-want-from-us-ai-regulation.html>> accessed 13 July 2024

<sup>187</sup>Ben Gould, 'California's New AI Laws Focus on Training Data, Content Transparency' (Cooley, 16 October 2024) <<https://www.cooley.com/news/insight/2024/2024-10-16-californias-new-ai-laws-focus-on-training-data-content-transparency>> accessed 13 July 2024

<sup>188</sup>Generative artificial intelligence: training data transparency (2024) AB2013

<sup>189</sup>AB2013, s3110(b)

<sup>190</sup>AB2013, s3111

development, outlining owners or sources of data sets, whether data protected by IP laws are within the domain and whether the data sets were purchased or licensed, will take effect.<sup>191</sup>

New York (“NY”) has proposed an Act on AI Training Data Transparency<sup>192</sup>, which seeks to introduce similar transparency requirements to California. Given that cases on data scraping and their subsequent IP infringement allegations are mainly focused on NY and California, it is unsurprising that a need for a form of framework has been recognised in these states. The definitions and obligations proposed in the NY Act are akin to the Californian framework. A developer is defined as “a person, partnership, state or local government agency, or corporation that designs, codes, produces or substantially modifies an artificial intelligence model or service for use by members of the public”.<sup>193</sup> Developers are similarly required to provide high-level summaries of their data sets, which include but are not limited to twelve obligations, including declaring whether data is protected by copyright, trademark or patent, as well as stating if the data set was purchased or licensed.<sup>194</sup>

The introduction of transparency obligations provides IP rightsholders with greater insight into the data used to train generative AI models. Although these requirements do not resolve the question of whether the use of IP protected works in training data constitutes infringement, they promote openness and may inform judicial reasoning in pending cases.

An Act<sup>195</sup> has also been introduced in Arkansas, providing a brief insight into the data set and AI-generated content ownership. It notes that the provider of the data is the owner of the resulting trained model unless the data is unlawfully acquired.<sup>196</sup> Washington has a pending Act<sup>197</sup> that largely follows the definitions and requirements outlined in the NY and California frameworks. The legal landscape in the U.S. remains largely undeveloped in this area, with only a few states having implemented or proposed frameworks. This lack of comprehensive legislation has contributed to a surge in legal cases aimed at clarifying whether the use of IP-

---

<sup>191</sup> AB2013, s3111(a)

<sup>192</sup> AI Training Data Transparency Act (2025) S 6955

<sup>193</sup> S 6955, s 1421(2)

<sup>194</sup> S 6955, s1422(e) and (f)

<sup>195</sup> An Act regarding the ownership of model training and content generated by a generative AI tool (2025) HB1876

<sup>196</sup> HB1876, s1(2)(a)

<sup>197</sup> Transparency in AI (2025) H 1168

protected works in generative AI training data sets constitutes infringement, underscoring the fragmented and uncertain nature of the current regulatory environment.

### 3.2.3 US Copyright Law

In the US, copyright legislation is codified at a federal level under the US Copyright Act.<sup>198</sup> The Act encompasses all “original works of authorship fixed in any tangible medium of expression”.<sup>199</sup> Similar to EU legislation, US authors are granted exclusive rights to the distribution and reproduction of their works, subject to limitations.<sup>200</sup> This includes adaptations and transformations of works, as well as reproductions and derivatives.<sup>201</sup>

In August 2023, the US Copyright Office published a Notice of Inquiry querying the public on which parts of generative AI training amount to copyright infringements and what conditions would have to be satisfied to allow fair use of protected works.<sup>202</sup> In May 2025, the pre-publication version was issued. This Notice of Inquiry acknowledges the increasing amount of litigation developing in the area, and in response to this, issued a publication with consolidated opinions from its Office and the public answering technical questions about generative AI training and copyrighted works.<sup>203</sup> The Notice focuses on a range of concerns stemming from generative AI training, such as *prima facie* infringements and fair use.

*Prima facie* infringements are considered under a number of headings, including data collection and curation, retrieval-augmented generation (“RAG”), outputs, and training. It is established that *prima facie* cases only need to meet two elements to constitute an infringement. The first is valid ownership of a copyright, the second that the copied contents are original work.<sup>204</sup> In terms of data collection and curation, the Copyright Office and many of the respondents agreed that reproduction rights are infringed through the establishment of data sets comprising copyrighted works. The consensus was that the steps involved in the process, including works being copied through downloads, transferring of such works, converting their

---

<sup>198</sup>Copyright Act of 1976, 17 USC ss101-810

<sup>199</sup>Copyright Act, s102(a)

<sup>200</sup>Copyright Act, s106

<sup>201</sup>Celeste Shen, ‘Fair Use, Licensing, and Authors’ Rights in the Age of Generative AI’ (2024) 22 *Northwestern Journal of Technology and Intellectual Property* 164

<sup>202</sup>de la Durantaye (n119) p746

<sup>203</sup>United States Copyright Office, ‘Copyright and Artificial Intelligence Part 3: Generative AI Training’ (2025) A Report of the Register of Copyrights Pre-Publication Version p1

<sup>204</sup>US Copyright Office (n202) p26

formats, and using them in subsets or developing modified versions, implicated reproduction rights.<sup>205</sup> RAG and outputs are outside the scope of generative AI model training. However, both were acknowledged as having the potential to infringe on copyrights.

The Notice also addresses potential IP breaches resulting from the training process. There are several ways in which the training process could breach copyrights. This can occur in the first step, where the data has to be downloaded and copied to a high-performance storage before training to enable its speed and scale. Then, in the next phase, works or pieces thereof are briefly reproduced and shown to the model. Whether this breaches copyright law depends on the amount of time the developer retains the works, which depends on the model, hardware and software used.<sup>206</sup>

It is not quite as clear whether the third aspect of the training process infringes copyrights. This part of the training development involves supplying training data, testing the performance on expected outputs and continuously updating the weights to optimise results. Depending on how much information the model weights retain, this could lead to the training data containing copies of works.<sup>207</sup> Model weights allow AI systems to ascertain which characteristics of the data are important. It uses numerical parameters to learn the data set's important features. These parameters are regularly adjusted to emphasise or deemphasise certain inputs which significantly affect the outputs of the model.<sup>208</sup>

The extent to which training inputs are memorised and whether this constitutes a breach of copyright has been debated. According to the US Copyright Notice, this largely depends on two factors: whether the work is fixed and if it is substantially similar to the original. Model weights consist of a list of numbers which do not change and generate and display memorised works, therefore satisfying the fixed and reproduction requirements.<sup>209</sup> Whether these criteria amount to an infringement must then be considered.

---

<sup>205</sup>US Copyright Office (n203) p26

<sup>206</sup>US Copyright Office (n203) p27

<sup>207</sup>US Copyright Office (n203) p28

<sup>208</sup>Min Jun Jung, 'AI Essentials: What are model weights?' (Medium, 10 October 2024)

<<https://engineadvocacyfoundation.medium.com/ai-essentials-what-are-model-weights-2e5b47ec77a1>>

accessed 24 July 2025

<sup>209</sup>US Copyright Office (n203) p29

In *Kadrey v Meta Platforms*<sup>210</sup>, the allegation of infringing derivative works was dismissed as the plaintiffs failed to argue that the models could produce substantially similar or copied works. Despite this ruling, Chhabria J stated that it did not mean the use of copyrighted works in AI training is lawful; only that in this instance, the grounds for the arguments were incorrect and failed to develop the correct one. Furthermore, the decision was noted by the Judge to have limited consequences, stating that it would be difficult to see how using copyrighted works in AI training could qualify as fair use in cases with more persuasive arguments.<sup>211</sup>

This line of reasoning can be seen in *Andersen v Stability AI*<sup>212</sup>. In this case, a third party's motion for dismissal was denied despite claims that they merely downloaded and used the trained AI model. They had not contributed to the training process but had downloaded a trained model, which had sufficient allegations that elements of protected works were incorporated, in some form, within the model. This not only exemplifies the scope of implication but also emphasises how central the retention of part or all of the works in model weights will be in determining infringements in subsequent cases.

The Digital Millennium Copyright Act<sup>213</sup> ("DMCA"), a federal statute enacted in 1998, provides additional copyright protections in the US. It does so in the form of digital rights management, which is used to stop infringements of rights and privileges protecting Internet Service Providers through safe harbours. It was introduced to adopt the WIPO Copyright Treaty<sup>214</sup> ("WCT") into US legislation. The Act brought in many new changes to US copyright law, including augmented protections against new copyright infringement methods, particularly through the internet. The framework includes Copyright Protection and Management Systems.<sup>215</sup> This prohibits circumvention of copyright protection systems<sup>216</sup> and maintains the integrity of copyright management information.<sup>217</sup> This is significant because several of the cases discussed in this section have relied on these provisions in their claims.

---

<sup>210</sup>*Kadrey v Meta Platforms* No. 3:23-cv-03417 (N.D. CAL)

<sup>211</sup>*Kadrey et al v Meta Platforms, Inc.*, 3:2023cv03417 Document 598 (N.D. Cal 2025) pp5 and 39

<sup>212</sup>*Andersen et al v Stability AI* 3:23-cv-00201 Document 166 (N.D. Cal 2024)

<sup>213</sup>Digital Millennium Copyright Act of 1998, HR 2281, 105<sup>th</sup> Congress

<sup>214</sup>World Intellectual Property Organisation Copyright Treaty of 1996

<sup>215</sup>DMCA, 17 USC Chap 12

<sup>216</sup>DMCA, s1201

<sup>217</sup>DMCA, s1202

### 3.2.4 Fair Use Doctrine

These *prima facie* infringements may, however, be permitted under the fair use doctrine. The fair use doctrine can be used as a copyright infringement defence in instances of limited reproductions, including for criticism or research.<sup>218</sup> Fair use is applied by the courts in a multi-factor test, which balances the particular purpose against the authors' rights. The doctrine establishes an all-or-nothing hard line where either use is granted without consent and payment, or it's prohibited and can be negotiated for a fee.<sup>219</sup>

The criteria for fair use comprise four considerations, outlined in the Copyright Act. These are: the purpose and character of use, taking into consideration whether or not it is commercially driven; the nature of the copyrighted works: the measure of copyrighted work used, and its effect on the potential market or value of the work.<sup>220</sup> These criteria are evaluated collectively, and it is not necessary to satisfy all four components.<sup>221</sup> The factors are applied through judicial balance, taking into account all relevant circumstances.<sup>222</sup> TDM is also recognised under US copyright law and can be exempt under fair use, provided it meets the outlined criteria.<sup>223</sup>

In the context of data scraping, the application of the 'fair use' exception under section 107<sup>224</sup> is highly disputed. Certain contributors to the US Copyright Notice argued that fair use has a detrimental effect on artists' livelihoods. In contrast, developers asserted that licensing copyright works would stifle development and doubted whether AI would be possible without fair use. Furthermore, several developers continue to claim that the aggregated data is only used to identify language patterns, not to reproduce protectable expressions.<sup>225</sup> As with the earlier discussion on training processes, the decision on this matter will be left to the judiciary and will depend on the strength of the parties' claims to sufficiently prove the presence or absence of copyrighted work.

---

<sup>218</sup>Shen (n201) p160

<sup>219</sup>de la Durantaye (n119) p745

<sup>220</sup>Copyright Act, s107

<sup>221</sup>Thongmeensuk (n129) p285

<sup>222</sup>*Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 19 (2021)

<sup>223</sup>de la Durantaye (n119) p285

<sup>224</sup>Copyright Act, s107

<sup>225</sup>US Copyright Office (n203) pp32-35

## First Factor

The court places significant consideration on the first factor of the fair use test. It is often referred to as the ‘transformative use factor’. A use meets this transformative requirement if it adds something new or the purpose differs from the original material, irrespective of whether the work remains largely the same.<sup>226</sup> It has been suggested that since the use of copyright works varies at different stages of generative AI development, it should be assessed separately. The US Copyright Office acknowledged that there should be separate examinations considered in the context of the overall use.<sup>227</sup> Similar to TDM in the EU, a distinction is placed on a use that promotes innovation while continuing to support creative activities, which will be more likely to be permitted. In *Andy Warhol Foundation v. Goldsmith*, these justifications were seen as a matter of degrees.<sup>228</sup> In *Google Books*, establishing a searchable database by scanning books was considered to have a “highly transformative purpose”<sup>229</sup>.

There have been various opinions surrounding whether copyrighted work in AI development is sufficiently transformative. Those who believe it is assert that the statistical analysis in machine learning differs significantly from the original works, that the use is non-expressive and have likened AI training to human learning. Those opposed, however, opine that the work is shown in a way that simply compresses the expressive elements, that specific work is not necessarily required for development and instead could be licensed and disagree with the non-expressive defence. The Copyright Office considers the transformative use to lie somewhere between the two perspectives, depending on the functionality and deployment of the model. Furthermore, they disagree with the non-expressive and human-like learning argument, as LLMs are trained on large tokens of information, where the linguistic expression is absorbed, making non-expressive use unlikely. Human-like learning also appears to be inaccurate, as humans retain imperfect impressions, whereas AI can produce perfect copies.<sup>230</sup>

The commerciality of the use is also examined under this first factor. This is similar to the EU’s distinction between non-profit and for-profit purposes. Unlike in the EU, where non-profit and research purposes are subject to different exemptions than for-profit purposes, the US acknowledges the difference and decides based on whether the developer stands to profit from

---

<sup>226</sup>Thongmeensuk (n129) p285

<sup>227</sup>US Copyright Office (n203) pp36-37

<sup>228</sup>US Copyright Office (n203) pp38-39, see also *Andy Warhol Foundation v. Goldsmith* 598 U.S. 508 (2023) p2

<sup>229</sup>*Authors Guild, Inc. v. Google Inc.*, No. 13-4829-cv (2d Cir. Oct. 16, 2015) p21

<sup>230</sup>US Copyright Office (n203) pp45-48

exploiting the copyrighted materials.<sup>231</sup> This has become a topic of controversy with the rise of ‘data laundering’. Stability AI are an example of this, focusing on financial backing for researchers and distancing itself from the direct development of software.<sup>232</sup> However, the Copyright Office has noted that the designation of the company is not the deciding factor when considering commerciality. It has also outlined that while fair use will not be determined on the grounds of unlawful access to works, it will certainly be influential in a ruling<sup>233</sup>

## **Second Factor**

The second factor looks at the nature of the copied works, assessing whether it is factual or creative. While this factor is generally of lesser significance, it is still important to note that factual works are afforded less protection than creative works in order to promote research and development.<sup>234</sup>

## **Third Factor**

The third criterion examines the portion of copyrighted content used and its substantiality in the context of the copyrighted work as a whole. This is considered under three headings. The first is the amount of work used. This considers the amount of work used and how indispensable the content is to its function. A claim involving only a small portion of work may still weigh against fair use, depending on how central it is to the use. The second aspect looks at the reasonableness of the purpose. Copying of entire works may be justified if it is carried out on reasonable grounds. For example, in *Google Books*<sup>235</sup>, copying whole texts to make a full-text searchable database was reasonable. Since the complete works were not made publicly available and only a portion of them was used to inform the database, fair use was allowed. This differs from generative AI training, which is not as clearly justified, and certain models necessitate the use of whole portions of works. The third discussion is about the amount of content made available to the public. The court considers the degree to which the publicly available works serve as a competing substitute for the original. In *Google Books*<sup>236</sup>, only snippets of works, not capable of substitution, were made available. Generative AI does not

---

<sup>231</sup>US Copyright Office (n203) p48

<sup>232</sup>Thongmeensuk (n129) p285

<sup>233</sup>US Copyright Office (n203) pp50-52

<sup>234</sup>Thongmeensuk (n129) p285

<sup>235</sup>*Authors Guild, Inc. v. Google Inc.*, No. 13-4829-cv (2d Cir. Oct. 16, 2015)

<sup>236</sup>*Authors Guild, Inc. v. Google Inc.*, No. 13-4829-cv (2d Cir. Oct. 16, 2015)

tend to make all of its content available to the public, and its outputs often do not contain protected expressions. It generally employs limits and guardrails to prevent potentially infringing outputs. Therefore, courts will likely place greater consideration on the guardrails developers have implemented to prevent infringing outputs than on publicly available information.<sup>237</sup>

#### **Fourth Factor**

The fourth factor plays a significant role in the court's fair use analysis and, along with the first factor, is given substantial weight in the final decision. The factor examines the market impact of the reproduction. Copyright intends to protect an author's right to derive profit from their creation. Therefore, if the reproduction constitutes a competing alternative or preferable substitute to the original work, it will not be considered fair use.<sup>238</sup>

The first element of this is lost sales incurred by the original artist. This looks at the effect the copied work has on the original artist's revenue. Courts take into consideration not only the impact of a certain use but also, if it were to become widespread and unrestricted, it queries what the substantial adverse impact would be. Some argue that generated works are becoming substitutive, with the example of news stories from local news stations being generated word for word, while opposing opinions believe the impact does not sufficiently meet the fourth requirement. Pirated works and their negative impact on sales are also a recognised consideration.<sup>239</sup>

Secondly, it has been argued that this market effect should be assessed more broadly, taking into consideration competition in a market for a specific type of work, even if the copied work is not substantially similar. Some regard this as too broad of a scope, potentially stifling innovation and development. However, the US Copyright Office disagrees, noting that this factor should not be interpreted too narrowly. This is based on the premise that an influx of AI-generated content could dilute the market, competing with human-authored works.<sup>240</sup>

---

<sup>237</sup>US Copyright Office (n203) pp57-60

<sup>238</sup>Thongmeensuk (n129) p286

<sup>239</sup>US Copyright Office (n203) pp62-64

<sup>240</sup>US Copyright Office (n203) pp64-66

Licensing opportunities also influence the fair use analysis under the fourth factor. In theory, actual or potential licensing agreements could be implemented between artists and developers. Therefore, the courts should consider the likelihood of this and the loss of revenue sustained by failing to come to such an agreement. A willingness to license content has been stated with individual and collective licensing agreements either in existence or under development. This is evident in the number of publicised deals between companies such as OpenAI and Shutterstock, and Getty Images. Subsequent discussions about collective agreements in various areas, such as news publishing, are also in the works. However, as this area progresses, difficulties may still arise, including sectors lacking such infrastructure, and transactional costs becoming a barrier exceeding the value of the works. This aspect may take the viewpoint that, in circumstances where licensing work is an option, this should be favourable, whereas in instances where licensing creates an insurmountable barrier then fair use may be preferred.<sup>241</sup>

An evaluation of the public benefits has also been recommended. Those seeking to rely on fair use assert public benefits such as fostering creativity and innovation. While those opposed insist that it will impede artistic growth, public interest is not sufficiently established and licensing the content could resolve these disputes. To further this argument, several examples of AI models trained on licensed or public domain works have been highlighted, for example, Adobe's Firefly.<sup>242</sup> These are all considerations the courts will have to balance in pending litigation when deciding if the use of copyrighted work in generative AI training is an infringement. The Notice concludes by acknowledging that the application of fair use will be circumstantial. Many purposes for the copyrighted works meet the transformative use requirement. The courts' decisions will likely depend on the purpose for which the model is deployed. Whether licensing copyrighted content will assuage concerns will be a key consideration.

### **3.2.5 Trademark Law**

Although trademark disputes have been less prevalent than copyright issues, they remain a relevant consideration, particularly in light of Getty Images'<sup>243</sup> lawsuit against Stability AI for alleged trademark infringement. In the US, trademarks are regulated under the Lanham Act,

---

<sup>241</sup>US Copyright Office (n203) pp66-71

<sup>242</sup>US Copyright Office (n203) pp71-73

<sup>243</sup>*Getty Images (US), Inc. v Stability AI, Inc.*, 1:23-cv-00135-JLH

also known as the Trademark Act.<sup>244</sup> “A trademark is any word, name, symbol, or design, or any combination thereof, used in commerce to identify and distinguish the goods of one manufacturer or seller from those of another and to indicate the source of the goods”.<sup>245</sup>

The Act’s objective is to prevent the use of similar marks likely to cause confusion to consumers or to avoid dilution of well-known marks. Trademark infringements are established under two categories in the Lanham Act: either as a registered mark<sup>246</sup> or an unregistered mark.<sup>247</sup> To substantiate a trademark infringement claim, a plaintiff must meet three criteria. Firstly, a valid or legally protectable mark must be proven. Secondly, ownership of the mark has to be demonstrated. Thirdly, the likelihood of confusion between the parties’ goods or services must be substantiated.<sup>248</sup>

Although trademark infringements are not the primary focus of litigation surrounding generative AI training, they remain an important topic for consideration. Given the significant amount of data that generative AI scrapes, there is potential for trademarked content to be collected and generated similarly or identically in an output. Getty Images is alleging, in its copyright and trademark infringement lawsuit, that Stability AI can replicate its watermark in synthetic images. The claim includes several assertions ranging from trademark infringement to unfair competition. The suit also comprises allegations of trademark dilution under Delaware and federal law, as well as unauthorised use of trademarks in AI-generated images violating Delaware’s Uniform Deceptive Trade Practices Act.<sup>249</sup> The case remains ongoing, so it is to be seen whether Getty Images will be successful in its claim.

### 3.2.6 US Case Law

The US has seen a flurry of lawsuits since the emergence of generative AI in 2022. Approximately forty cases have been initiated in the US, which is where most of the major AI

---

<sup>244</sup>Lanham Act of 1946, 15 USC 1051 *et seq.* Trademark Act

<sup>245</sup>Cornell Law School, ‘Trademark’ (Legal Information Institute, April 2025)

<<https://www.law.cornell.edu/wex/trademark>> accessed 26 July 2025, see also Trademark Act, s1127

<sup>246</sup>Trademark Act, s1114

<sup>247</sup>Trademark Act, s1125(a)

<sup>248</sup>Cornell Law School (n245)

<sup>249</sup>Diana Milton and Harrison Enright, ‘Getty Images v Stability AI’ (Baker Hostetler)

<<https://www.bakerlaw.com/getty-images-v-stability-ai/>> accessed 26 July 2025

developers are based.<sup>250</sup> The claims are primarily focused on copyright and highlight the uncertainties in the application of existing legal frameworks. The cases vary from copyright infringement claims based on the use of protected content in training AI models to claims concerning the generated output. As previously mentioned, fair use is applied on a case-by-case basis; thus, many of the lawsuits also consider the effectiveness of this defence. The vast majority of cases are still in the early stages of litigation. However, a few initial rulings have begun to emerge and address some of the issues surrounding generative AI.<sup>251</sup> This section will examine several of these cases, highlighting the differing arguments presented and the court's initial decisions.

### ***Getty Images (US) Inc & Ors v Stability AI Ltd***

In February 2023, Getty Images filed a complaint against Stability AI. They claimed Stability AI copied more than 12 million of their photographs, as well as associated captions and metadata. Their images were allegedly used to train the defendant's image generation AI without a license and as part of their endeavour to create a competing business. Getty Images also asserts that its copyright management information was altered or removed, that false copyright management information was provided, and its famous trademark was infringed.<sup>252</sup>

Getty Images' copyright infringement claim includes several breaches under the Copyright Act.<sup>253</sup> According to Section 106,<sup>254</sup> Getty Images is the owner or exclusive licensee of their content and is therefore entitled to exclusive rights under the Act. It is alleged that Stability AI infringed on this through accessing images, their associated titles and captions, as well as other metadata in the database to make derivatives. This occurred in direct violation of the terms of use on the website, and for Stability AI's commercial benefit.<sup>255</sup>

An infringement of false copyright management information ("CMI") is also purported under Section 1202(a).<sup>256</sup> Getty Images' watermark is considered copyright management

---

<sup>250</sup>Mark Sellman, 'Getty Images sues AI firm in landmark case for copyright law' (The Times, 09 June 2025) <<https://www.thetimes.com/uk/technology-uk/article/getty-images-copyright-trial-sr8s6fk9b>> accessed 01 August 2025

<sup>251</sup>Lisa Oratz, Tyler Robbins and Sean West, 'Initial Rulings in AI Copyright Litigation Provide Insight but Leave Questions Unanswered.' (2024) 53(3) American Bar Association 26

<sup>252</sup>*Getty Images (US), Inc. v Stability AI, Inc.*, 1:23-cv-00135-JLH p1

<sup>253</sup>Copyright Act of 1976, 17 USC s101 *et seq*

<sup>254</sup>Copyright Act of 1976, 17 USC s106

<sup>255</sup>*Getty Images (US), Inc. v Stability AI, Inc* p23

<sup>256</sup>Copyright Act of 1976, 17 USC s1202(a)

information, and Stability AI's use of a modified version of the watermark infringed this for knowingly and intentionally concealing, inducing and enabling copyright infringement.<sup>257</sup> The allegation also concerns copyright management information under Section 1202(b)<sup>258</sup> concerning its removal or alteration. It is claimed that Stability AI intentionally removed or altered the watermark and metadata associated with the images to infringe Getty Images' copyright.<sup>259</sup> Proving CMI can prove difficult. Getty Images has been considered a rare case for successfully evidencing intent to remove or alter CMI. This was only possible because certain generated images contained modified versions of Getty Images' trademark, which appeared bizarrely synthetic. Without a recognisable watermark, their claim under DMCA 1202a, 1202b, and controllable evidence of direct infringement would prove much more difficult.<sup>260</sup>



261

In terms of trademark law, Getty Images has submitted a number of claims, including trademark infringement, violating Section 32 of the Lanham Act<sup>262</sup> by using its distinctive mark, which allegedly has caused and will continue to cause the public confusion. Unfair competition and false designation of origin for potentially causing and continuing to cause the public confusion are submitted under Section 43(a) of the Lanham Act,<sup>263</sup> in addition to

<sup>257</sup> *Getty Images (US), Inc. v Stability AI, Inc* pp24-25

<sup>258</sup> Copyright Act of 1976, 17 USC s1202

<sup>259</sup> *Getty Images (US), Inc. v Stability AI, Inc* p26

<sup>260</sup> Yunah Kwon, 'Copyright in the Age of Generative AI: Legal Gaps and Artist Vulnerability' (Columbia Undergraduate Law Review, 17 June 2025) <<https://www.culawreview.org/current-events-2/copyright-in-the-age-of-generative-ai-legal-gaps-and-artist-vulnerability>> accessed 03 August 2025

<sup>261</sup> *Getty Images (US), Inc. v Stability AI, Inc*, Complaint of 03 February 2023 p18, illustration from Getty Images' lawsuit, showing their original photograph and a similar image generated by Stable Diffusion (including Getty Images' watermark)

<sup>262</sup> s32 Lanham Act, 15 U.S.C § 1114(1)

<sup>263</sup> s34(a) Lanham Act, 15 U.S.C §1125(a)

trademark dilution under Section 43(c) of the Lanham Act<sup>264</sup> and Section 3313 of the Delaware Trademark Act.<sup>265</sup>

Stability AI has claimed that their use of copyrighted materials was permitted under fair use. The court, therefore, must consider whether Stability AI's model demonstrates a transformative purpose, the commerciality of their system, the nature of Getty Images' protected content, the portion of works used and the potential market effect. Getty Images appears to have a strong case against fair use, given the substantial number of images they claim have been infringed. Similarly, an extensive portion of discernible images has been evidenced in Stability AI's outputs, which could help lead to a decision in Getty Images' favour.<sup>266</sup> The ruling should provide clarity in the modern copyright law landscape to copyright holders and AI developers alike. However, due to an extensive court docket, the case remains ongoing.<sup>267</sup>

The case illustrates the extensive scope of potential IP infringement claims. Whether all of these infringements will be substantiated is to be seen. It is also important to note that this dispute follows a separate but similar case, *Getty Images v Stability AI*,<sup>268</sup> in the UK, as well as a class action against Stability AI in California.<sup>269</sup> The case in the UK was submitted on similar grounds, which exemplifies the cross-border nature of generative AI. It is not limited to one jurisdiction but is subject to varying rules in each. As the legal landscape in this area continues to develop, litigation such as this may serve to highlight the need to adopt a harmonised international framework to facilitate easier compliance.

### ***The New York Times v Microsoft and OpenAI***

In December 2023, The New York Times filed a lawsuit against OpenAI and its financial backer, Microsoft.<sup>270</sup> This is a consolidated case with three combined plaintiffs: The New York

---

<sup>264</sup>s43(c) Lanham Act, 15 U.S.C1125(c)

<sup>265</sup>Delaware Trademark Act of 2024, 6 DE Code §3301 *et seq*

<sup>266</sup>Matthew Coulter, 'Aiming for Fairness: an Exploration into Getty Images v Stability AI and its Importance in the Landscape of Modern Copyright Law' (2024) 34 DePaul Journal of Art, Technology and Intellectual Property Law pp14-18

<sup>267</sup>James Vincent, 'Getty Images sues AI Art generator Stable Diffusion in the US for copyright infringement' (The Verge, 06 February 2023) <<https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion>> accessed 02 August 2025

<sup>268</sup>*Getty Images and Other v Stability AI* [2025] EWHC 38 (Ch)

<sup>269</sup>Blake Brittain, 'Getty Images Lawsuit says Stability AI misused photos to train AI' (Reuters, 06 February 2023) <<https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/>> accessed 01 August 2025

<sup>270</sup>*The New York Times Co v Microsoft et al* 1:23-cv-11195 (S.D.N.Y.)

Times, The New York Daily News and the Centre for Investigative Reporting. The NYT's allegations comprise a number of claims, including copying millions of articles to train OpenAI's ChatGPT without their consent. This has allegedly resulted in outputs almost identical to the NYT's originals, which they claim threatens their journalistic standards and decreases visits to their website. It is submitted that the LLM allows people to bypass the NYT's paywall, and at times, it also makes up stories which it wrongly claims were published by the NYT, harming their reputation.<sup>271</sup> Interestingly, other publishers such as News Corp. and the Associated Press have chosen to avoid litigation by instead reaching a content-sharing deal with OpenAI.<sup>272</sup>

OpenAI denies these allegations, stating articles are not directly copied by AI models in their normal use. It is also reasoned that fair use permits the use of publicly available articles to train AI models. They additionally claim that their focus on research and development makes their LLM a non-commercial purpose under the fair use criteria, and that any similarities between the outputs and NYT articles are accidental and rare. Furthermore, OpenAI alleges that the NYT manipulated outputs by using specific prompts to make the issue appear worse.<sup>273</sup>

The case is still in the early stages; however, in April of this year, Judge Sidney Stein issued an opinion granting and dismissing equally in part OpenAI's motion to dismiss. OpenAI requested a dismissal on the grounds that the statute of limitations had already passed. This was not upheld as they failed to establish that the plaintiffs should have known about the alleged infringements three years before the claim was filed. A motion to dismiss contributory infringement claims was also dismissed, as third-party end-user infringement had been plausibly alleged by the plaintiffs. The NYT and Daily News have provided substantial, widely publicised examples, which the Judge stated, "give rise to a plausible inference of copyright

---

<sup>271</sup>Nicole Chen, 'Stolen Stories or Fair Use? The New York Times v. OpenAI and the Limits of Machine Learning' (Columbia Undergraduate Law Review, 09 June 2025) <<https://www.culawreview.org/ddc-x-culr-1/nyt-v-openai-and-microsoft>> accessed 02 August 2025

<sup>272</sup>Bobby Allyn, 'The New York Times takes OpenAI to Court. ChatGPT's future could be on the line' (NPR, 14 January 2025) <<https://www.npr.org/2025/01/14/nx-s1-5258952/new-york-times-openai-microsoft>> accessed 02 August 2025

<sup>273</sup>Chen (n271)

infringement by third parties.”<sup>274</sup> A claim to dismiss trademark dilution allegations by the Daily News was also submitted and subsequently denied by the court.<sup>275</sup>

OpenAI’s motions to dismiss did not prove entirely unsuccessful. A number of their claims were upheld. Reproduction allegations under the Digital Millennium Copyright Act<sup>276</sup> were not permitted as the excerpts were not sufficiently identical to the originals, nor did they amount to copies.<sup>277</sup> OpenAI is expected to rely on the defence of fair use for the claims they did not seek to dismiss. In the opinion of the NYT, if the data had not been scraped by OpenAI, a license would have been required; therefore, compensating the NYT for their articles and avoiding market harm. There are a lot of nuances in determining market harm. Market displacement will have to be substantiated to rebut fair use’s fourth criterion.<sup>278</sup>

The lawsuit has been widely considered the first big test for AI in the copyright law space. An interesting aspect of this case is the seemingly more extensive evidence of direct memorisation and verbatim reproduction compared to other lawsuits. It is also one of the few that has survived in part, motions to dismiss. This indicates that the court believes there is a plausible legal basis there. This in no way constitutes a verdict. Decisions in this area of law are very much still in the formative stages.<sup>279</sup>

### ***Andersen v Stability AI***

In 2023, Andersen,<sup>280</sup> along with two other visual artists, filed a class action lawsuit against various generative AI developers. This was the first joint lawsuit challenging generative AI companies on the grounds of copyright infringement. This case is likely to set important groundwork for handling IP disputes going forward and has been followed by several similar class action lawsuit filings.<sup>281</sup>

---

<sup>274</sup>Maura O’Malley, ‘US Judge refuses OpenAI’s motion to dismiss New York Times copyright infringement claims’ (The Global Legal Post, 08 April 2025) <<https://www.globallegalpost.com/news/us-judge-refuses-openais-motion-to-dismiss-new-york-times-copyright-infringement-claims-887263879>> accessed 03 August 2025

<sup>275</sup>O’Malley (n274)

<sup>276</sup>Digital Millennium Copyright Act of 1998, HR 2281, 105<sup>th</sup> Congress

<sup>277</sup>O’Malley (n274)

<sup>278</sup>Rachel Reed, ‘ChatNYT’ (Harvard Law Today, 22 March 2024) <<https://hls.harvard.edu/today/does-chatgpt-violate-new-york-times-copyrights/>> accessed 02 August 2025

<sup>279</sup>Reed (n278)

<sup>280</sup>*Andersen v Stability AI* 3:23-cv-00201 (N.D. CAL)

<sup>281</sup>Yunah Kwon, ‘Copyright in the Age of Generative AI: Legal Gaps and Artist Vulnerability’ (Columbia Undergraduate Law Review, 17 June 2025) <<https://www.culawreview.org/current-events-2/copyright-in-the-age-of-generative-ai-legal-gaps-and-artist-vulnerability>> accessed 03 August 2025

The plaintiffs initially filed a number of claims, including copyright infringement, DMCA violations, and false endorsement, arising from the use of their works in Stability AI and DeviantArt's training data sets. Towards the end of 2023, the lawsuits were largely dismissed by the court with leave to amend, except for a direct claim of copyright infringement by Andersen against Stability AI. Initiating an infringement action requires works to be registered with the US Copyright Office. Andersen's direct infringement allegation survived, as some of her works had been registered. Furthermore, excluding Andersen, the plaintiffs were unable to provide sufficient evidence substantiating a claim of removed CMI in violation of DMCA 1202b. The burden of proof was on the artists to demonstrate intent to remove CMI from their works or that the generative AI models were built with the intent to facilitate copyright. The defendants were instead able to exhibit that their models were created to generate distinct new images. The court in this instance ruled in their favour.<sup>282</sup>

This highlights the notorious difficulty in proving intent to infringe. In *Getty Images*<sup>283</sup>, demonstrating intent to remove or alter CMI was only possible because of their distinct watermark. This would prove much more difficult to exhibit for independent concept artists who typically publish without a recognisable watermark. Without heavily branding their work, artists may find it difficult to prove intent to infringe. Furthermore, the progression of the class action of only Andersen's registered work shows the barriers independent artists may face in bringing infringement proceedings. This also shows the high burden of proof artists face to progress their lawsuits.<sup>284</sup>

### ***Bartz v Anthropic PBC***

*Bartz v Anthropic PBC*<sup>285</sup> is arguably the most significant ruling to date on whether copyright-protected content can lawfully be used by AI developers to train their models. A summary judgment was recently granted to Anthropic on this issue. The decision has drawn a definitive line between transformative use in training, permissible under fair use, and prohibited illicit data collection.<sup>286</sup> The judgment has been received with varied responses, viewing this

---

<sup>282</sup>Kwon (n281)

<sup>283</sup>*Getty Images and Other v Stability AI* [2025] EWHC 38 (Ch)

<sup>284</sup>Kwon (n281)

<sup>285</sup>*Bartz v Anthropic* *Bartz v Anthropic PBC* 3:24-cv-05417 (N.D. Cal)

<sup>286</sup>Barry Scannell and Leo Moore, 'Artificial Intelligence and Copyright: Significant U.S. Judgment' (William Fry, 01 July 2025) <<https://www.williamfry.com/knowledge/artificial-intelligence-and-copyright-significant-u-s-judgment/>> accessed 03 August 2025

landmark ruling as a split decision, which could define how AI developers will source, store and use training data.<sup>287</sup>

The focus of this case was the centralised library of digitised books that Anthropic PBC curated to train the LLM underlying Claude, their generative AI. The library consisted of both books purchased, which were then converted into digital form, and pirated copies downloaded from numerous websites. This resulted in various authors alleging copyright infringement for Anthropic's use of their books in its digital library and to train its LLMs. A putative class action lawsuit was filed where Anthropic sought a summary judgment. After separately evaluating the fair use criteria under Section 107 of the Copyright Act,<sup>288</sup> Judge Alsup was able to provide a summary ruling.<sup>289</sup>

After assessing each of the four fair use factors, the court concluded that the use of books to train generative AI amounted to fair use. When considering the first factor, purpose and character of use, the court favoured AI developers. This was justified on the basis that works are used to train LLMs that generate new outputs which are “quintessentially transformative”.<sup>290</sup> A key factor in this finding is Anthropic's software that ensures the output does not contain infringing content. It was acknowledged by that court that the analysis may differ if the outputs were infringing.<sup>291</sup>

On the other hand, the second factor, nature of the copyrighted work, weighed against fair use, as books contain expressive characteristics, and Anthropic chose these works for their expressive elements. Despite Anthropic copying the Author's entire works, the third factor, amount and substantiality of the portion used, was decided in favour of fair use. This was justified given the substantial data needed to train LLMs and the absence of any allegations or evidence that the Author's works appeared in the outputs. The fourth factor, the potential market effect, also favoured fair use. It was determined that the demand for the Author's works was not displaced by the generative AI model. Competition arising from generative AI works

---

<sup>287</sup>Chris Hanslik, 'Bartz v. Anthropic: A Split Decision That Reshapes AI Copyright Law' (Lexology, 17 July 2025) <<https://www.lexology.com/library/detail.aspx?g=b3d3aa98-6c02-4d1e-b194-f092e5d6a500>> accessed 03 August 2025

<sup>288</sup>Copyright Act of 1976, 17 USC s107

<sup>289</sup>Practical Law, 'Anthropic's Use of Copyrighted Materials to Train GenAI Is Fair Use: N.D. Cal' (Thomas Reuters Practical Law, 25 June 2025) <[https://uk.practicallaw.thomsonreuters.com/w-047-4069?transitionType=Default&contextData=\(sc.Default\)&firstPage=true](https://uk.practicallaw.thomsonreuters.com/w-047-4069?transitionType=Default&contextData=(sc.Default)&firstPage=true)> accessed 04 August 2025

<sup>290</sup>*Bartz v Anthropic* *Bartz v Anthropic* PBC, Order on Fair Use Document 231, p13

<sup>291</sup>Practical Law (n289)

was also not considered to be the kind of creative substitute envisioned in the Copyright Act. The court went so far as to say that even if they were to consider a hypothetical emerging market for licensing protected works to AI developers, the impact of such a market would not be protected by the Copyright Act.<sup>292</sup>

*Bartz v Anthropic* did, however, provide authors with a partial success in its fair use consideration of Anthropic's central digital library. The court considered the pirated copies taken without compensation or authorisation, in contrast with the lawfully acquired print format, which was digitised, the originals destroyed, and the digitals were not redistributed. Fair use was upheld for the lawfully purchased copies, in a similar analysis and finding to the decision on content used to train Anthropic's LLMs. The pirated copies were determined to have no fair use justification and were denied summary judgment. Illegally copying the content as a substitute for purchasing displaced demand for purchased copies, affecting the publishing market. The use also failed to meet the transformative requirement.<sup>293</sup>

*Bartz*, along with the related ruling in *Kadrey v. Meta Platforms Inc.*<sup>294</sup>, represents somewhat of a setback for authors challenging fair use. However, the court has left open the possibility of challenging fair use in future cases, highlighting how context-dependent such decisions are and underscoring the importance of stronger arguments by authors to prevail. The legal landscape in the US remains uncertain regarding the use of copyrighted works in training generative AI. As litigation progresses, clearer guidance is likely to emerge for copyright owners and AI developers. At this stage, it is clear that to successfully rebut fair use, claimants require substantial evidence and a persuasive argument.

### 3.3 International Guidelines on Artificial Intelligence

To tackle the fragmented approach to regulating AI, a number of international organisations have sought to provide consolidated guidance to encourage compliance and ethical practices. Although, as aforementioned, the publications are primarily guidance and not legally binding. However, the policies will be briefly discussed, as while they are generally not mandatory, they

---

<sup>292</sup>Practical Law (n289)

<sup>293</sup>Practical Law (n289)

<sup>294</sup>*Kadrey v. Meta Platforms Inc* 3:23-cv-03417 (N.D. Cal)

provide recommendations in the area, which may influence the development of further regulations and judgments of ongoing litigation.

### 3.3.1 The Council of Europe

In May 2024, the Council of Europe adopted the first internationally legally binding instrument on AI. The Council of Europe Framework Convention<sup>295</sup> “establishes a legal framework focused on safeguarding human rights, democracy, and the rule of law in AI development and usage.”<sup>296</sup> The Convention aims to ensure that developments throughout the lifecycle of AI systems are compliant with human rights, democracy and the rule of law while being conducive to technological progress and innovation.<sup>297</sup>

The framework intends to complement existing International standards while addressing any legal gaps that may result from rapid technological advances. It is essentially technology-neutral and does not regulate technology.<sup>298</sup> The convention’s preamble recognises the benefit of fostering cooperation while being conscious of both the unprecedented opportunities AI systems may offer, while acknowledging the potential impact the technology can have on human rights and dignity, democracy and the rule of law.<sup>299</sup>

The EU and the US are both signatories to the Convention. In the EU, the framework will be adopted through the AI Act<sup>300</sup>. The convention is consistent with the EU AI Act and includes some key concepts, such as a focus on human-centric AI, a risk-based approach, key principles for trustworthy AI, etc.<sup>301</sup> In terms of generative AI, it specifically encourages transparency for AI-generated content as well as in interactions with AI models.

---

<sup>295</sup>Convention on Artificial Intelligence and Human Rights, Democracy and Rule of Law (2024) CETS 225 5.IX.2024

<sup>296</sup>Irish Computer Society, ‘Comparison of Council of European Framework Convention on Artificial Intelligence with the European Union Artificial Intelligence Act’ (ICS, 29 October 2024) <<https://ics.ie/2024/10/29/comparison-of-council-of-europe-framework-convention-on-artificial-intelligence-with-the-european-union-artificial-intelligence-act/>> accessed 07 July 2025

<sup>297</sup> Council of Europe, ‘The Framework Convention on Artificial Intelligence’ (Council of Europe, 2025) <<https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence>> accessed 07 July 2025

<sup>298</sup>Council of Europe (n297)

<sup>299</sup>Convention on AI (n295)

<sup>300</sup>AI Act

<sup>301</sup>European Commission, ‘Commission Signed the Council of Europe Framework Convention on Artificial Intelligence and human rights, democracy and rule of law’ (European Commission, 05 September 2024) <<https://digital-strategy.ec.europa.eu/en/news/commission-signed-council-europe-framework-convention-artificial-intelligence-and-human-rights>> accessed 07 July 2025

Article 8 outlines that “Each Party shall adopt or maintain measures to ensure that adequate transparency and oversight requirements tailored to the specific contexts and risks are in place in respect of activities within the lifecycle of artificial intelligence systems, including with regard to the identification of content generated by artificial intelligence systems.”<sup>302</sup> Transparency and oversight are important for AI governance as they provide appropriate safeguards for activities within the lifecycle of the AI. The principle of transparency encourages openness and clarity throughout the governance of AI activities. This is an important consideration for intellectual property issues arising from generative AI training data.<sup>303</sup>

The explanatory report set out by the Council of Europe states that transparency could require communicating certain information. It continues to say that this may involve informing persons concerned and the public, where appropriate, about the data used to create, train and operate the model.<sup>304</sup> Furthermore, “Providing transparency could in addition facilitate the possibility for parties with legitimate interests, including copyright holders, to exercise and enforce their intellectual property rights.”<sup>305</sup> The Convention establishes a baseline standard in international law for AI practices by governments that respect human rights, reaffirming the continued applicability of existing human rights obligations to AI-related activities.<sup>306</sup>

### 3.3.2 Additional Guidelines: OECD, UN & G7

The OECD has published a number of recommendations on Artificial Intelligence, encouraging Members to uphold principles of trustworthy AI. The first intergovernmental standard on AI was published by the OECD and derived from the G20’s AI Principles. These Principles and the Recommendation of the Council on Artificial Intelligence, which was adopted later in 2021,

---

<sup>302</sup>Convention on AI (n295)

<sup>303</sup>Council of Europe, ‘Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and Rule of Law’ (2024) COE, CETS 225 5.IX.2024, art 8

<sup>304</sup>COE Explanatory Report (n303), art 8

<sup>305</sup>COE Explanatory Report (n303), art 8

<sup>306</sup>The Bureau of Democracy, Human Rights and Labor, ‘The Council of Europe’s Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law’ (U.S. Department of State) <<https://www.state.gov/bureau-of-democracy-human-rights-and-labor/the-council-of-europes-framework-convention-on-artificial-intelligence-and-human-rights-democracy-and-the-rule-of-law#:~:text=The%20Convention%20defines%20a%20baseline,rights%20obligations%20to%20AI%20activities>> accessed 13 July 2025

represent a commitment to the standard and an expectation that the Member States will adhere to and implement.<sup>307</sup>

The OECD recognises potential risks arising from AI, such as copyright infringement and has thus introduced principles to mitigate this. Furthermore, they note, “AI knows no borders. In this context, the principles offer a foundation for international cooperation and interoperability with guidance designed to stand the test of time in the fast-paced world of AI.”<sup>308</sup> The OECD endeavours to guide policymakers on regulations going forward to maximise benefits and mitigate risks. Similar to the Council of Europe, the OECD encourages transparency to facilitate responsible disclosure regarding AI systems. While the principle is not binding, it is certainly a consideration when discussing data scraping and IP infringements as it seeks to “provide information that enables those adversely affected by an AI system to challenge its output.”<sup>309</sup>

The OECD has further recognised the IP issues arising from generative AI trained on scraped data, a topic discussed at length in their February 2025 Artificial Intelligence paper. The report identifies many of the concerns outlined or to be discussed in this thesis. It recognises the fragmented approach to the issue and provides several considerations going forward, including raising awareness about IP issues, implementing a voluntary code of conduct to address data scraping issues, utilising technical tools to protect IP rightsholders and provide greater access to their information, and including standard contract terms to create a more responsible path to web scraping.<sup>310</sup> The paper highlights the legislative gaps when discussing generative AI and provides policy considerations in light of current litigation.

Furthermore, the United Nations (“UN”) and G7 have published resolutions to encourage countries to monitor for risks arising out of AI, protect personal data and safeguard human rights. The UN adopted a draft resolution on 21 March 2024 entitled “Seizing the opportunities of safe, secure and trustworthy AI systems for sustainable development”. More than 120 Member States backed the resolution, which encouraged protecting personal data, safeguards for human rights and monitoring AI for risks. This highlights the common underlying concerns, which states and intergovernmental

---

<sup>307</sup>White & Case (n46)

<sup>308</sup>OECD, ‘AI Principles’ (OECD) <<https://www.oecd.org/en/topics/sub-issues/ai-principles.html>> accessed 08 July 2025

<sup>309</sup>Recommendation of the Council on Artificial Intelligence (2019) OECD/Legal/0449, sec 1(3)(iv)

<sup>310</sup>OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping’ (2025) OECD Artificial Intelligence Papers, No. 33 chap 4, pp28-35

organisations alike have sought to provide guidance on. The G7 framework, while not as widely applicable, also seeks to promote safe, secure and trustworthy AI. The Hiroshima AI Process Comprehensive Policy Framework consist of four pillars: the guiding principles for organisations developing advanced AI systems, the code of conduct designed to supplement these principles, an analysis of challenges, opportunities and priorities of generative AI and a project-based cooperation with the aim to support the development of responsible AI. Neither the Guiding Principles nor the Code of Conduct are legally binding. Despite this, they have the potential to exert strong political influence internationally.<sup>311</sup>

The participating countries are not bound by the UN and G7 recommendations. Instead, they aim to reach a consensus on secure, trustworthy and safe AI systems. In light of such, Members are encouraged to develop and implement their governance and regulatory frameworks and approaches.<sup>312</sup> While there is much guidance surrounding AI governance at an international level, it is clear that it lacks a binding consolidated approach. Binding regulations are fragmented, applying only at the state level.

### **3.3.3 International Copyright Law**

Intellectual property law obligations, including copyright, are outlined in a series of international law treaties. The World Trade Organisation (“WTO”) has an agreement on Trade-Related Aspects of Intellectual Property Rights (“TRIPS”), while the WIPO has also issued some treaties with international IP responsibilities. These treaties are enforceable on their Member States, which include the EU and the US. The TRIPS Agreement,<sup>313</sup> the WIPO Copyright Agreement<sup>314</sup> and the WIPO Performances and Phonograms Treaty<sup>315</sup> contain provisions relevant to both copyright and data scraping. These agreements supplement existing treaties, the Berne and Rome Conventions,<sup>316</sup> on copyright and related rights.<sup>317</sup>

---

<sup>311</sup>White & Case (n46)

<sup>312</sup>White & Case (n46)

<sup>313</sup>Agreement on Trade-Related Aspects of Intellectual Property Rights, Annex 1C of the Marrakesh Agreement Establishing the World Trade Organization (15 April 1994)

<sup>314</sup>World Intellectual Property Organisation Copyright Treaty of 1996

<sup>315</sup>WIPO Performances and Phonograms Treaty of 1996

<sup>316</sup>Berne Convention for the Protection of Literary and Artistic Works (signed 9 September 1886, amended at Paris 24 July 1971, entered into force 10 October 1974) 828 UNTS 221, and International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organisations (opened for signature 26 October 1961, entered into force 18 May 1964), ‘Rome Convention’

<sup>317</sup>OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping’ (2025) OECD Artificial Intelligence Papers No. 33 p23

The provisions adopt several obligations, such as a prohibition against formalities. This means copyright protections are automatic and do not require registration. A three-step test was also introduced to determine when limits and exceptions to copyright may apply. The threshold is confined to special cases, does not interfere with the regular use of work, and does not unfairly harm the legitimate interests of the rightsholder. Protections are also provided for Rights Management Information (“RMI”), which includes embedded metadata as well as preventing the bypassing of technological measures (“TMs”) and protective measures for compilations of data or material.<sup>318</sup> State-level application of these protections can be seen in the aforementioned case law. For instance, Getty Images’ claim of false copyright management information derives from the WIPO Copyright Treaty provision on RMI, which was transposed into US law as CMI in the DMCA.

#### **4. Policy Approaches**

The case law and legal framework implemented in the EU and the US illustrate the varied approaches taken by jurisdictions. The EU’s reaction has been proactive, focusing on progressing its regulatory power and placing itself at the forefront of legislative development with the AI Act, whereas the US’s approach centres on fostering AI development and establishing a more reactive context-based approach. This section will assess legislative measures implemented in other jurisdictions, comparatively examining their legal framework or lack thereof, against the EU and the US. This will be followed by recommendations on policy approaches based on the legal gaps discussed throughout this thesis.

##### **4.1 Comparative analysis of jurisdictions with and without a legal framework**

This section will provide insights into the development of the legal framework of other jurisdictions, with a primary focus on China and the UK. It will assess what regulations have been implemented by each of the countries, contrasting these approaches with the EU and the US. Ongoing litigation and its potential impact on the respective legal systems will also be examined. The UK discussion will also consider other common law jurisdictions and their differing approach from civil law countries, as well as why they have proven more litigious.

---

<sup>318</sup>OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping’ (2025) OECD Artificial Intelligence Papers No. 33 p23

#### 4.1.1 China

China has long been at the forefront of innovation, and when it comes to AI, this remains true. According to Stanford’s Global AI Vibrancy Ranking, China ranks second behind the US.<sup>319</sup> China has positioned itself as a key competitor in the AI market, having developed 13% of the world’s foundational models. In comparison, the US holds 73% of foundational models, with the EU lagging in this respect. China’s regulatory procedures have been shaped by a mix of restrictive legislation in response to the EU’s approach and facilitative measures due to US pressure. This has led to an AI-centred geopolitical competition that consists of a technological and regulatory race to see which jurisdiction will assert the most influence between the ‘Brussels Effect’, the ‘California Effect’ or the ‘Beijing Effect’.<sup>320</sup>

In 2023, China’s first administrative regulation was introduced on the management of generative AI services. The Cyberspace Administration of China (“CAC”), in coordination with six other governmental regulators, established the Interim Measures for Management of Generative Artificial Intelligence Services<sup>321</sup> (“AI Measures”). The State has also adopted Labelling Rules to help identify generative AI outputs and a Market Regulation and Standardisation framework to enhance generative AI security and governance. One of the outlined standards will enforce requirements and evaluation criteria to enhance safeguards around data sets used in generative AI pre-training and fine-tuning development. Both the Labelling Rules and Market Regulation and Standardisation will come into effect later this year.<sup>322</sup>

In the context of this discussion, China’s AI Measures are the most relevant framework to consider. Similar to the EU AI Act, the objective of the AI Measures is to promote healthy AI development and regulate its application.<sup>323</sup> The scope of application applies only to generative AI that provides services to the public.<sup>324</sup> Under the AI Measures, generative AI providers are required to guarantee the source and legality of the data used in their models. Furthermore,

---

<sup>319</sup>Human-Centered Artificial Intelligence, ‘Which countries are leading in AI?’ (Stanford University) <<https://hai.stanford.edu/ai-index/global-vibrancy-tool>> accessed 05 August 2025

<sup>320</sup>Aifang Ma, ‘Regulation in pursuit of artificial intelligence (AI) sovereignty: China’s mix of restrictive and facilitative modalities’ (2024) 34 AJIC 2

<sup>321</sup>Interim Measures for Management of Generative Artificial Intelligence Services of July 2023, ‘AI Measures’

<sup>322</sup>White & Case, ‘AI Watch: Global Regulatory tracker – China’ (White & Case, 29 May 2025) <<https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-china>> accessed 05 August 2025

<sup>323</sup>AI Measures, article 1

<sup>324</sup>AI Measures, article 2

providers must comply with Article 7 throughout all phases of training, including pre-training, fine-tuning, etc. The AI Measures ensure that training data only consists of lawful sources and that providers do not infringe upon intellectual property rights.<sup>325</sup>

This differentiates China's policy from both the US and the EU. Their approach to IP protection is seen as innovative, contrasting with other jurisdictions grappling with ethical and legal issues around using copyrighted content in training data and seeking viable regulatory solutions.<sup>326</sup> While the EU AI Act offers guidance on the use of copyrighted works in generative AI training, its provisions remain somewhat vague and lack detailed clarity, as highlighted in *Like Company v. Google*.

In 2024, these AI measures were explored in China's first case on generative AI copyright infringement. This case is known as the Ultraman or Guangzhou Internet Court case.<sup>327</sup> In its findings, the court determined the defendant was liable for copyright infringement. It concluded that a reasonable duty of care was not exercised by the provider in its generative outputs, therefore violating the AI Measures. In 2023, the plaintiff discovered that prompts, input into the defendant's generative AI, containing the word 'Ultraman' or related phrases, resulted in outputs identical or substantially similar to their character Ultraman. The court focused on two key questions to determine whether the use constituted an IP infringement. It was first assessed whether the plaintiff's rights, such as reproduction, derivative works and dissemination, were infringed. If this amounted to a copyright violation, it had to be determined what civil liabilities ensued.<sup>328</sup>

It was decided that the generation of substantially similar images, which retained the original expressions, satisfied the first criterion. In considering the second question, it was found that the defendant failed to demonstrate a reasonable duty of care, breaching articles 4 (requirement to comply with applicable policies), 12 (content must be labelled) and 15 (complaint and reporting mechanism obligation) of the AI Measures.<sup>329</sup> The defendant in their argument

---

<sup>325</sup>Sara Migliorini, 'China's Interim Measures on generative AI: Origin, content and significance' (2024) 53 Computer Law & Security Review 4, see also AI Measures, article 7

<sup>326</sup>Migliorini (n325) p5

<sup>327</sup>*Guangzhou Internet Court* (2024) Yue 0192 Min Chu 113. (2024粤0192初113号)

<sup>328</sup>Seagull Song, 'China's First Case on AIGC Output Infringement – Ultraman' (King & Wood Mallesons, 28 February 2024) <<https://www.kwm.com/cn/en/insights/latest-thinking/china-s-first-case-on-aigc-output-infringement-ultraman.html>> accessed 04 August 2025

<sup>329</sup>AI Measures, articles 4, 12 and 15

claimed that a third party developed the model, expunging their obligations under the AI Measures. The court rejected this claim and the defendant was liable to pay compensation. In its remarks, the court emphasised that providers should not be overburdened by the measures. However, they should be proactive to fulfil their reasonable and affordable duty of care. This exemplifies China's aim to foster a balance between AI development and IP protection.<sup>330</sup> To date, the Ultraman decision is the most favourable towards copyright holders. However, China's firm legislative stance is rather unsurprising given its sweeping crackdown on tech companies between 2020 and 2022.<sup>331</sup>

#### 4.1.2 United Kingdom

There is no comprehensive legal framework for AI or generative AI in the United Kingdom ("UK"). While consultations were conducted seeking recommendations on what measures should be introduced, nothing has yet been implemented. The current policy has prioritised a flexible framework with sector-specific laws governing AI in the UK.<sup>332</sup> Their legislative landscape has been somewhat influenced by its goal to maintain the UK's position as third in global AI innovation. An AI Bill, requiring companies to turn over their large models for testing by the AI Security Institute in the UK, was to be published in December 2024. However, this has been delayed and subsequently reconsidered to keep pace with the US's approach to foster AI innovation. Members of Parliament ("MPs") are hesitant to implement measures that may affect the UK's appeal to AI companies.<sup>333</sup>

The UK, similar to many jurisdictions at this time, have seen a rise in concerns around the use of copyrighted works to train AI models.<sup>334</sup> Thus far, generative AI training data issues have been dealt with under the Copyright, Designs and Patents Act.<sup>335</sup> As with the EU, the UK

---

<sup>330</sup>Song (n328)

<sup>331</sup>Angela Zhang, 'The Promise and Perils of China's Regulation of Artificial Intelligence' (2024) 63 Columbia Journal of Transnational Law 104

<sup>332</sup>White & Case, 'AI Watch: Global Regulatory tracker – United Kingdom' (White & Case, 25 March 2025) <<https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-kingdom>> accessed 06 August 2025

<sup>333</sup>Secretary of State for Science, Innovation and Technology, 'Copyright and Artificial Intelligence (2024) Copyright and AI: Consultation CP 1205 p8, see also Eleni Courea, 'UK delays plans to regulate AI as ministers seek to align with Trump administration' (The Guardian, 24 February 2025) <<https://www.theguardian.com/technology/2025/feb/24/uk-delays-plans-to-regulate-ai-as-ministers-seek-to-align-with-trump-administration>> accessed 06 August 2025

<sup>334</sup>Secretary of State for Science, Innovation and Technology, 'Copyright and Artificial Intelligence (2024) Copyright and AI: Consultation CP 1205 p10

<sup>335</sup>Copyright, Designs and Patents Act of 1988

Copyright Act has an exemption for TDM. However, its scope is more limited as the CDSM Directive was not implemented in the UK due to Brexit. It is only applicable for research and non-commercial purposes.<sup>336</sup> As part of a Consultation conducted by the Secretary of State for Science, Innovation and Technology, several policy options were proposed. These included introducing a broader data mining exception, akin to the US fair use exemption. Other recommendations included strengthening licensing requirements, as well as a proposal for data mining, with opt-outs supported by transparency measures similar to the EU, which was initially the preferred recommendation.<sup>337</sup>

The government received over 11,000 responses to its consultation. The replies are still being considered, and it will therefore require more time before a decision is made. More recently, copyright holders have voiced their apprehension regarding the introduction of a TDM exception, similar to the EU. Their concerns centre around two main issues: firstly, their preference is to license their works to ensure adequate compensation is provided. Secondly, a system similar to the EU may prove challenging to monitor and implement, as well as the fact that no adequate technical solution to uphold opt-outs has been provided yet in the EU. More recent discussions in the UK indicate that the leaning may now be towards an approach involving licensing and transparency, but a decision has not been finalised.<sup>338</sup>

Legal disputes have also begun to develop in the UK. As previously mentioned, Getty Images<sup>339</sup> has also brought a claim in the UK High Court against Stability AI, a UK-based company. Getty Images alleges in its argument that Stability AI is liable for secondary copyright infringement, passing off and trademark infringement. The plaintiff's main claim of infringement under UK copyright law was dropped, alluding to the difficulty in proving the jurisdiction in which the infringing acts took place. To succeed in a secondary infringement claim, Getty Images will have to meet three criteria: the importation of Stable Diffusion into the UK or that it was possessed, sold, let, hired or offered or exposed for sale or hire by Stability AI. It must also be proven that Stable Diffusion is both an infringing copy and an article, as

---

<sup>336</sup>Rebecca Peknam-Walsh, 'AI and copyright – where are we now?' (Fieldfisher, 10 June 2025) <<https://www.fieldfisher.com/en/services/intellectual-property/intellectual-property-blog/ai-and-copyright-where-are-we-now>> accessed 06 August 2025

<sup>337</sup>Secretary of State for Science, Innovation and Technology, 'Copyright and Artificial Intelligence (2024) Copyright and AI: Consultation CP 1205 pp12-14

<sup>338</sup>Rebecca Peknam-Walsh, 'AI and copyright – where are we now?' (Fieldfisher, 10 June 2025) <<https://www.fieldfisher.com/en/services/intellectual-property/intellectual-property-blog/ai-and-copyright-where-are-we-now>> accessed 06 August 2025

<sup>339</sup>*Getty Images and Other v Stability AI* [2025] EWHC 38 (Ch)

well as proving that they knew or had reason to believe this. The decision, in this case, is highly anticipated as it could either provide rightsholders with a precedent to assert their copyright claims or allow AI developers almost unfettered access to copyrighted works.<sup>340</sup>

Considering the case law in the US and UK, litigation surrounding IP infringements related to AI training data appears to be more prominent in common law countries. Disputes have also developed in other jurisdictions, including India and Canada. In India, a prominent news agency, ANI Media,<sup>341</sup> is currently alleging that OpenAI unlawfully scraped free and paywalled content from its website. Similarly, Canada has a dispute before the Supreme Court of British Columbia, claiming Caseway AI illegally used CanLII's database without the necessary permissions or licenses to train its AI model.<sup>342</sup> The ruling in both cases will largely depend on copyright laws and the defendant's fair dealing argument. Fair dealing is similar to the fair use doctrine in the US, as it also requires a court ruling on a case-by-case basis. However, its scope is considered to be more limited than the US exception.<sup>343</sup>

This case law trend in common law jurisdictions can be attributed to a number of factors. Firstly, common law jurisdictions tend to be driven by litigation. Unlike civil law systems, where the laws are codified, common law regulations are more open to interpretation, leading to test cases and legal boundaries being defined by court decisions. Doctrines such as fair use and fair dealing can also seem ambiguous, leading to legal challenges that seek clarity rather than waiting for legislative reform. Particularly in the US, companies with lobbying power and vast resources, legal and financial, can affect the legal landscape. Furthermore, the frameworks in these jurisdictions tend to be less stringent and are influenced more significantly by judicial precedent. Consequently, issues surrounding AI training data and intellectual property infringements have become more prominent in common law countries.

---

<sup>340</sup>Cerys Wyn Davis and Gill Dennis, 'Getty Images v Stability AI: why the remaining copyright claims are of more than secondary significance' (Pinsent Mason, 15 July 2025) <<https://www.pinsentmasons.com/out-law/analysis/getty-images-v-stability-ai-copyright-claims-significance>> accessed 06 August 2025

<sup>341</sup>*ANI Media Pvt Ltd v OpenAI Inc & Anr*, CS(COMM) 1028/2024

<sup>342</sup>*Canadian Legal Information Institute v. 1345750 BC Ltd. et al* (2024) SCBC Action No. S

<sup>343</sup>The University of Edinburgh, 'Copyright Exceptions and Fair Dealing' (The University of Edinburgh, 21 August 2024) <<https://library.ed.ac.uk/library-help/copyright/copyright-exceptions-and-fair-dealing>> accessed 08 August 2025

## 4.2 Policy recommendations based on legal gaps

Despite the cross-border scope of generative AI, policy approaches have proven to be rather fragmented, with legal frameworks primarily implemented at the state level. Getty Images<sup>344</sup> concurrent cases exemplify the need for cross-jurisdictional guidelines to simplify compliance for AI developers and provide IP rightsholders with greater clarity. Therefore, a voluntary code of conduct addressing legislative gaps could help address concerns posed by data scraping. This code of conduct could build upon existing codes, including the EU GPAI Code of Practice and the G7 Code of Conduct. A code of conduct could be beneficial to policymakers with its swift adoption and degree of flexibility in its implementation. Compliance with the code could be voluntary for companies. States could then create a centralised database to monitor compliance and treat non-compliance in line with their established procedures.<sup>345</sup>

Irrespective of whether states elect to implement a voluntary code of conduct, a number of provisions should be considered to address concerns about data scraping. AI developers could adopt systems to prevent safeguards and controls from being circumvented, therefore avoiding prompts or generated content likely to infringe.<sup>346</sup> For example, in the NYT,<sup>347</sup> it is claimed that prompts requesting specific paragraphs of their articles reproduce significant portions of their content, allowing users to bypass their paywall. The NYT alleges this infringes on their copyright as it affects their revenue. Introducing a provision preventing users from making such requests could alleviate this concern.

Another consideration is transparency. Fostering transparency around data scraping practices promotes better compliance and accountability. This concept has been discussed in many jurisdictions; however, its practice is currently limited. Under the EU AI Act, GPAI models are required to provide a summary of the content used in their training data set.<sup>348</sup> Outside of the EU, transparency is not as strictly enforced. In the US, no such requirement has been introduced at the federal level, whereas states such as California and New York have pending transparency policies, enforceable only at the state level. The UK is considering transparency provisions,

---

<sup>344</sup>*Getty Images (US), Inc. v Stability AI, Inc.*, 1:23-cv-00135-JLH, see also *Getty Images and Other v Stability AI* [2025] EWHC 38 (Ch)

<sup>345</sup>OECD, 'Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping' (2025) OECD Artificial Intelligence Papers No. 33 pp29-30

<sup>346</sup>OECD, 'Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping' (2025) OECD Artificial Intelligence Papers No. 33 p30

<sup>347</sup>*The New York Times Co v Microsoft et al* 1:23-cv-11195 (S.D.N.Y.)

<sup>348</sup>AI Act, art 53(1)(d)

but nothing is currently in place. A code of conduct with a transparency provision could help facilitate a coordinated approach, simplifying compliance for AI developers and making rights protection easier for rightsholders by aligning data management best practices.<sup>349</sup>

Transparency obligations could also outline specific requirements for users at various stages of the process. For example, data aggregators could adopt transparency tools disclosing information about their data sets, or users could be prevented from altering rights management information, etc. A formal reporting process could be established, providing a designated point for lodging violation complaints and reducing the burden on the courts.<sup>350</sup> In addition, alternatives to litigation, such as alternative dispute resolution, could be encouraged as a first step toward resolving disputes.

The code of conduct could also be used to clarify each jurisdiction's applicable laws. For example, the states' pertinent intellectual property laws could be embedded in the code. Clearly outlining the relevant laws would facilitate easier compliance for AI companies. It would also give rightsholders clearer and more accessible information about their legal protections, enabling them to better understand and exercise their rights. This provision could be included as in the EU AI Act under the GPAI obligations. Article 53(1)(c)<sup>351</sup> of the Act specifies that providers must establish a policy in compliance with the EU copyright and related rights law. This gives AI companies and rightsholders clarity on enforceable laws.

AI developers could also be encouraged to implement technical tools to identify reserved websites and content that should not be scraped, although this may prove difficult in practice.<sup>352</sup> TDM in the EU only permits web scraping on a commercial basis where rights have not been reserved. There has been some uncertainty surrounding what constitutes 'machine-readable' rights reservation under EU law. In *LAION*,<sup>353</sup> the court commented that natural language would likely suffice in this instance. However, this statement was *obiter dictum*, meaning it is not binding. Despite this, encouraging the voluntary use of these tools could alleviate concerns

---

<sup>349</sup>OECD, 'Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping' (2025) OECD Artificial Intelligence Papers No. 33 pp30-31

<sup>350</sup>OECD, 'Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping' (2025) OECD Artificial Intelligence Papers No. 33 pp31-32

<sup>351</sup>AI Act, art 53(1)(c)

<sup>352</sup>OECD, 'Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping' (2025) OECD Artificial Intelligence Papers No. 33 p33

<sup>353</sup>*LAION v Robert Kneschke* Hamburg District Court, 310 O.22723

and reduce infringement claims while also providing rightsholders with an easier way to protect their content by opting out.

There's no easy way to resolve the challenges posed by web scraping of protected works. States are eager to strike a balance between the innovation of AI and the protection of IP rights. A viable solution going forward could be to encourage AI developers to license protected content. This has been a topic of discussion among copyright holders and AI developers. In the EU, those who elect to opt out from commercial TDM can seek to enter into a licensing agreement.<sup>354</sup> There is a willingness to enter into licensing agreements in the US, with publicised deals between OpenAI and companies such as News Corp.<sup>355</sup> The UK has also begun to consider licensing and transparency as a potential resolution going forward.

With licensing, there are, however, a number of factors to consider. Firstly, larger AI developers like OpenAI would be in a better position to negotiate agreements given their substantial resources compared to smaller startups. This will likely place smaller enterprises at an innovative disadvantage as they would not have access to the same vast substantive content. Similarly, some creative industries are better equipped to negotiate individual or collective licensing agreements, placing some copyright holders at a disadvantage.

A final aspect to consider is that whether or not states choose to implement these proposed policies may not determine AI companies' compliance obligation. Depending on their market presence, AI developers will still be required to comply with the EU AI Act. Similar to the GDPR, the AI Act has extraterritorial scope, which is a significant feature granting it a broad and far-reaching effect. Article 2 of the AI Act<sup>356</sup> states that GPAI models on the market in the EU, irrespective of established location, are subject to the regulation. Furthermore, this applies to providers in a third country where the produced output is used in the EU.<sup>357</sup> There has been speculation as to whether the AI Act, like the GDPR, will have a global reach. The GDPR impacted companies worldwide, requiring them to comply with its rules to continue operating

---

<sup>354</sup>European Innovation Council and SMEs Executive Agency, 'Artificial Intelligence and copyright: use of generative AI tools to develop new content' (European Commission, 16 July 2024) < [https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/artificial-intelligence-and-copyright-use-generative-ai-tools-develop-new-content-2024-07-16-0\\_en](https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/artificial-intelligence-and-copyright-use-generative-ai-tools-develop-new-content-2024-07-16-0_en) > accessed 09 August 2025

<sup>355</sup>Reuters, 'Sam Altman's OpenAI signs content agreement with News Corp' (Reuters, 22 May 2024) < <https://www.reuters.com/technology/sam-altmans-openai-signs-content-agreement-with-news-corp-2024-05-22/> > accessed 09 August 2024

<sup>356</sup>AI Act, art 2(1)(a)

<sup>357</sup>AI Act, art 2(1)(c)

and processing data in the EU market. There has been conjecture as to whether the AI Act will have a similar Brussels effect. A key determining factor will be the size of the EU market. Given its substantial population and wealth, it is likely to be considered a large and attractive market for AI systems, drawing interest from online platforms and AI product providers. There are also significant opportunities to market AI systems in both the business and public sectors. It is therefore unlikely that providers will give up access to the EU. Thus, companies will either have to conform or risk being pushed out.<sup>358</sup>

## 5. Conclusion

The issue of intellectual property law infringements arising from generative AI training data is a pressing challenge in the context of technological development. The landscape has proven complex due to the rapidly evolving nature of AI and the arduous task of implementing a regulatory framework in line with these advancements. As demand for vast amounts of data to train AI models increases, concerns will continue to be raised about the data collection methods used and their subsequent impact on the protection of intellectual property rights. The difficulty for states to strike a balance between permitting data usage and protecting IP works has not proven easy to resolve. In this context, it has been complicated by a variety of factors, including ambiguous legal frameworks and strategic interests.

Although initially, any mention of generative AI was absent from the EU AI Act's draft, growing attention following the launch of ChatGPT prompted a concerted effort to introduce specific guidelines. Subsequently, providers of general-purpose AI will have obligations, such as complying with Copyright and Related Rights and providing detailed summaries of the content used in their model to promote transparency. Significantly, it has also introduced the GPAI Code of Practice, which includes a Model Documentation Form simplifying transparency disclosures and practical suggestions to comply with copyright law. Despite these policies creating a degree of legal certainty, IP rightsholders' concerns have persisted, particularly in regard to the allowance of TDM through the AI Act and CDSM. Under the current regulations, TDM is permitted without the need for permission for research or non-commercial entities and subject to opt-outs for commercial companies. The permissibility of TDM is currently being queried by *Like Company* at the Court of Justice of the European

---

<sup>358</sup>Marco Almada and Anca Radu, 'The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy' (2024) 25 German Law Journal pp654-655

Union. This case will significantly influence the permissible scope of web scraping in the EU and the ongoing validity of TDM in such contexts.

Policies in the US have proven much more fragmented in comparison to the comprehensive and unified approach adopted in the EU. Regulatory oversight of generative AI is currently driven largely at the state level, with states such as California and New York introducing transparency laws to help fill legislative gaps. The current legal framework, or lack thereof, has been influenced by the US's aim to maintain its long-standing reputation at the forefront of AI. The goal of achieving global AI dominance, regardless of its impact on intellectual property rights, is not shared by all. The proposed COPIED Act intends to promote transparency and bolster content owners' right to take legal action. However, this act remains pending.

In the meantime, intellectual property law infringements are being managed under copyright and trademark laws. In the US, concerns about using protected content in training data are mainly addressed through the fair use doctrine. While case law in this area remains in its early stages, summary judgments in *Kadrey* and *Anthropic* indicate an inclination toward permitting fair use for generative AI. However, the courts have left open the possibility of challenging fair use once again in future cases. Decisions were noted as being context-dependent, with strong evidence of needing a substantive argument by rightsholders to prevail. The legal landscape in the US continues to remain uncertain, with rightsholders claiming use of their content is detrimentally affecting their industry. In contrast, AI developers persist in asserting that without fair use, generative AI would not be possible, as well as claiming that the data is only used to develop language patterns, not to reproduce protectable expressions.

A need for governance at an international level has been recognised by various international organisations. The Council of Europe has established a legal framework with human-centric safeguards. However, outside of transparency requirements, it does not provide considerable resolutions to IP infringement allegations. Recommendations on generative AI have also been established by the OECD, UN and G7. Proposals for a voluntary code of conduct addressing data scraping and suggestions for technical tools to protect IP are mentioned; however, these are all non-binding. Despite this, a need for cross-border measures is clear given Getty Images' concurrent cases in the US and UK. Web scraping is not confined to one jurisdiction and its wide-reaching extent should be reflected in governing policies.

Given the ongoing litigation worldwide, most notably in the US, but also in the EU, China, the UK, etc., appropriate measures should be adopted to address the current legal uncertainty. It is unlikely that this issue will resolve itself without proactive measures to balance intellectual property rights with the use of web scraping for generative AI datasets. What is clearly needed is a mixture of legal approaches, some of which have already been established and technical solutions. Several recommendations could be adopted by states to simplify compliance for AI developers and provide IP rightsholders with greater clarity. This might involve a code of conduct designed with enough flexibility to enable jurisdictions to adapt it to their existing legal frameworks. Technical systems could be adopted to prevent safeguards and controls from being circumvented. This could include implementing software that avoids prompts or generated content likely to infringe. An emphasis on transparency can be seen in several jurisdictions, which could promote improved compliance and accountability. Licensing, in particular, could help strike a balance between data use and the protection of copyright holders, ensuring that rightsholders continue to profit from their works while enabling AI developers to access vast, reliable datasets.

Despite these recommendations, there is no simple solution to the challenges posed by web scraping of protected works. As generative AI continues to evolve, the legal framework will remain in flux, shaped by ongoing legislative debate and judicial decisions, particularly given ongoing litigation that will significantly influence how IP infringements from data scraping are addressed.

## Bibliography

### Case law:

#### Canada

*Canadian Legal Information Institute v. 1345750 BC Ltd. et al* (2024) SCBC Action No. S

#### China

*Guangzhou Internet Court* (2024) Yue 0192 Min Chu 113. (2024粤0192初113号)

#### EU

*C-250/25 Like Company v Google Ireland* [2025]

#### France

Decision 20-MC-01 of 9 April 2020 on requests for interim measures by the Syndicat des éditeurs de la presse magazine, the Alliance de la presse d'information générale and others and Agence France-Presse

Decision 21-D-17 of 12 July 2021 on compliance with the injunctions issued against Google in Decision 20-MC-01 of 9 April 2020

Cases No 19/0074F, 19/0078F and 19/0080F – Related Rights Commitments Proposal of Alphabet Inc, Google LLC, Google Ireland Limited and Google France

Decision 22-D-13 of 21 June 2022 regarding practices implemented in the press sector

Decision 24-D-03 of 15 March 2024 regarding compliance with the commitments in

Decision 22-D-13 of 21 June 2022 of the Autorité de la concurrence regarding practices implemented by Google in the press sector

#### Germany

*LAION v Robert Kneschke* Hamburg District Court, 310 O.22723

#### India

*ANI Media Pvt Ltd v OpenAI Inc & Anr*, CS(COMM) 1028/2024

#### UK

*Getty Images and Other v Stability AI* [2025] EWHC 38 (Ch)

#### US

*Authors Guild, Inc. v. Google Inc.*, No. 13-4829-cv (2d Cir. Oct. 16, 2015)

*Google LLC v. Oracle Am., Inc.*, 593 U.S. 1, 19 (2021)

*Andy Warhol Foundation v. Goldsmith* 598 U.S. 508 (2023)

*The New York Times Co v Microsoft et al* 1:23-cv-11195 (S.D.N.Y.)

*Andersen v Stability AI* 3:23-cv-00201 (N.D. Cal)

*Kadrey v Meta Platforms* No. 3:23-cv-03417 (N.D. Cal)

*Getty Images (US), Inc. v Stability AI, Inc.*, 1:23-cv-00135-JLH

*Getty Images (US), Inc. v Stability AI, Inc.*, 1:23-cv-00135-JLH, Complaint of 03 February 2023

*Andersen et al v Stability AI* 3:23-cv-00201 Document 166 (N.D. Cal 2024)

*Bartz v Anthropic Bartz v Anthropic PBC* 3:24-cv-05417 (N.D. Cal)

*Bartz v Anthropic Bartz v Anthropic PBC*, Order on Fair Use Document 231

*Kadrey et al v Meta Platforms, Inc.*, 3:2023cv03417 Document 598 (N.D. Cal 2025)

## **Legislation:**

### **China**

Interim Measures for Management of Generative Artificial Intelligence Services of July 2023

### **Council of Europe**

Convention on Artificial Intelligence and Human Rights, Democracy and Rule of Law (2024)  
CETS 225 5.IX.2024

### **EU**

Code of Practice for General-Purpose AI Models (2025) Copyright Chapter

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 (Directive on Copyright in the Digital Single Market) *OJ L 130/1* (“CDSM Directive”)

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 (Information Society Directive) [2001] *OJ L 167/10* (“InfoSoc Directive”)

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) [2024] *OJ L*, 2024/1689 (“AI Act”)

### **France**

Law No 2019-775 of 24 July 2019, on the Creation of Neighbouring Rights for the Benefit of Press Agencies and Publishers

### **Germany**

Gesetz über Urheberrecht und verwandte Schutzrechte (“Urheberrechtsgesetz”), Act on Copyright and Related Rights version applicable since 7 June 2021

## **OECD**

Recommendation of the Council on Artificial Intelligence (2019) OECD/Legal/0449

## **UK**

Copyright, Designs and Patents Act of 1988

## **US**

AI Training Data Transparency Act of 2025, S 6955

An Act regarding the ownership of model training and content generated by a generative AI tool (2025) HB1876

Content Origin Protection and Integrity from Edited and Deepfaked Media Act of 2025 S 1396

Copyright Act of 1976, 17 USC ss101-810

Delaware Trademark Act of 2024, 6 DE Code §3301 *et seq*

Digital Millennium Copyright Act of 1998, HR 2281, 105<sup>th</sup> Congress

Executive Office of the President “Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People” (October 2022)

Executive Order 14110 of October 30, 2023. Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

Executive Order 14179 of January 23, 2024. Removing Barriers to American Leadership in Artificial Intelligence

Generative artificial intelligence: training data transparency (2024) AB2013

Lanham Act of 1946, 15 USC 1051 *et seq.* Trademark Act

National Artificial Intelligence Initiative of 2020, 15 USC 9401

One Big Beautiful Bill Act of 2025, HR 492, 119<sup>th</sup> Congress

Transparency in AI (2025) H 1168

## **WIPO**

Berne Convention for the Protection of Literary and Artistic Works (signed 9 September 1886, amended at Paris 24 July 1971, entered into force 10 October 1974) 828 UNTS 221

International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organisations (opened for signature 26 October 1961, entered into force 18 May 1964), ‘Rome Convention’

WIPO Performances and Phonograms Treaty of 1996

World Intellectual Property Organisation Copyright Treaty of 1996

## **WTO**

Agreement on Trade-Related Aspects of Intellectual Property Rights, Annex 1C of the Marrakesh Agreement Establishing the World Trade Organization (15 April 1994)

## **Books:**

Ryan W. White and Chirag Shah, *Information Access in the Earlier Era of Generative AI* (1<sup>st</sup> edn, Springer 2024)

Tom Taulli, *Generative AI: How ChatGPT and Other AI Tools Will Revolutionise Business* (1<sup>st</sup> edn, Apress L. P. 2023)

## **Reports:**

Council of Europe, ‘Explanatory Report to the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and Rule of Law’ (2024) COE, CETS 225 5.IX.2024

European Commission, ‘Artificial Intelligence - Questions and Answers’ (2024) European Commission QANDA/21/1683

IP and Frontier Technologies Division, ‘Generative AI Navigating Intellectual Property’ (2024) WIPO Conversation, 8<sup>th</sup> Session

OECD, ‘G7 Hiroshima Process on Generative Artificial Intelligence (AI)’ (2023) Japanese G7 Presidency and the G7 Digital and Tech Working Group

OECD, ‘Intellectual Property Issues in Artificial Intelligence Trained on Data Scraping’ (2025) OECD Artificial Intelligence Papers No. 33

Secretary of State for Science, Innovation and Technology, ‘Copyright and Artificial Intelligence (2024) Copyright and AI: Consultation CP 1205

United States Copyright Office, ‘Copyright and Artificial Intelligence Part 3: Generative AI Training’ (2025) A Report of the Register of Copyrights Pre-Publication Version

US Congress, ‘Highlights of the 2023 Executive Order on Artificial Intelligence for Congress’ (2024) US Congress Reports R47843

## **Journal Articles:**

Aifang Ma, ‘Regulation in pursuit of artificial intelligence (AI) sovereignty: China’s mix of restrictive and facilitative modalities’ (2024) 34 AJIC 1

Anett Pogácsás, 'New Ways of Plagiarism? At the Meeting Point of the Specific Form of Co-Authorship and the Transparency Criteria of the EU AI Regulation' (2025) IIC <  
<https://link.springer.com/article/10.1007/s40319-025-01608-2>> accessed 05 July 2025

Angela Zhang, 'The Promise and Perils of China's Regulation of Artificial Intelligence' (2024) 63 Columbia Journal of Transnational Law 102

Artha Dermawan, 'Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese "nonenjoyment purposes"' (2024) 27 The Journal of World Intellectual Property 44

Celeste Shen, 'Fair Use, Licensing, and Authors' Rights in the Age of Generative AI' (2024) 22 Northwestern Journal of Technology and Intellectual Property 157

Francesca Palmiotto, 'The AI Act Roller Coaster: The Evolution of Fundamental Rights Protection in the Legislative Process and the Future of the Regulation' (2025) 16 European Journal of Risk Regulation 770

Javier Torre de Silva and López de Letona, 'The Right to Scrape Data on the Internet: From the US Case *hiQLabs, Inc. v. LinkedIn Corp.* to the ChatGPT Scraping Cases: Differences Between US and EU Law' (2024) 5 Global Privacy Law Review 5

Kaigeng Li, Hong Wu and Yupeng Dong, 'Copyright Protection during the training stage of generative AI: Industry-oriented U.S. law, rights-oriented EU law, and fair remuneration rights for generative AI training under the UN's international governance regime for AI' (2024) 55 Computer Law and Security Review 1

Katharina de la Durantaye, 'Control and Compensation. A Comparative Analysis of Copyright Exceptions for Training Generative AI' (2025) 56(4) IIC 737

Laurie Harris, 'Generative Artificial Intelligence: Overview, Issues and Considerations for Congress' (2025) IF12426 Library of Congress 1

Lisa Oratz, Tyler Robbins and Sean West, 'Initial Rulings in AI Copyright Litigation Provide Insight but Leave Questions Unanswered.' (2024) 53(3) American Bar Association 26

Marco Almada and Anca Radu, 'The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy' (2024) 25 German Law Journal 646

Marco Almada and Nicolas Petit, 'The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights' (2025) 62 Common Market Law Review 85

Matthew Coulter, 'Aiming for Fairness: an Exploration into *Getty Images v Stability AI* and its Importance in the Landscape of Modern Copyright Law' (2024) 34 DePaul Journal of Art, Technology and Intellectual Property Law 1

Nicoletta Rangone and Luca Megale, 'Risks Without Rights? The EU AI Act's Approach to AI in Law and Rulemaking' (2025) *European Journal of Risk Regulation* 1

Nidia Andrade Moreira, Pedro Miguel Freitas, and Paulo Novais, 'The AI Meets General Purpose AI: The Good, The Bad and The Uncertain' (2023) 14116 *Lecture Notes in Computer Science* 157

Nir Kshetri, 'Navigating EU Regulations: Challenges for US Technology Firms and the Rise of Europe's Generative AI Ecosystem' (2024) 57(10) *IEEE Computer Society* 112

Oskar Gstrein, Noman Haleem and Andrej Zwitter, 'General Purpose AI regulation and the European Union Act' (2024) 13(3) *Internet Policy Review* 1

Qian Li and Konrad Kollnig, 'Data Scraping for the Training of Generative AI: Lessons from Chinese Case Law and Regulation' (2024) 25(2) *Computer Law Review International* 33

Saliltorn Thongmeensuk, 'Rethinking copyright exceptions in the era of generative AI: Balancing innovation and intellectual property protection' (2024) 47(2) *The Journal of World Intellectual Property* 278

Sara Migliorini, 'China's Interim Measures on generative AI: Origin, content and significance' (2024) 53 *Computer Law & Security Review* 1

Stanley Greenstein and Mauro Zamboni, 'Navigating the legislative dilemma: evaluating the EU AI Act's approach to regulating emerging technologies' (2025) *The Theory and Practice of Legislation* 1

Stephanie Houde, Vera Liao, Jacquelyn Martino, Michael Muller, David Piorkowski, John Richards, Justin Weisz and Yunfeng Zhang, 'Business (mis)Use Cases of Generative AI' (2020) 2848 *HAI-GEN* 1

The International Journal of Science, 'AI must play fair when they use academic data in training' (2024) 632 *Nature* 953

### **Websites:**

Audrey Pope, 'NYT v. OpenAI: The Time's About -Face' (*Harvard Law Review*, 10 April 2024) <<https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/>> accessed 26 May 2025

Autorité de la concurrence, 'Related Rights: the Autorité fines Google €250 million for non-compliance with some of its commitments made in June 2022 (Autorité de la concurrence, 20 March 2024) <<https://www.autoritedelaconcurrence.fr/en/press-release/related-rights-autorite-fines-google-eu250-million-non-compliance-some-its>> accessed 28 July 2025

Barry Scannell and Leo Moore, ‘Artificial Intelligence and Copyright: Significant U.S. Judgment’ (William Fry, 01 July 2025) <<https://www.williamfry.com/knowledge/artificial-intelligence-and-copyright-significant-u-s-judgment/>> accessed 03 August 2025

Barry Scannell and Leo Moore, ‘CJEU to Rule on AI Chatbots and Copyright in Landmark Case Against Google’ (William Fry, May 2025) <<https://www.williamfry.com/knowledge/cjeu-to-rule-on-ai-chatbots-and-copyright-in-landmark-case-against-google/>> accessed 27 July 2025

Ben Gould, ‘California’s New AI Laws Focus on Training Data, Content Transparency’ (Cooley, 16 October 2024) <<https://www.cooley.com/news/insight/2024/2024-10-16-californias-new-ai-laws-focus-on-training-data-content-transparency>> accessed 13 July 2024

Bird & Bird, ‘AI Regulatory Horizon Tracker’ (Bird & Bird, 30 April 2025) <<https://www.twobirds.com/en/capabilities/artificial-intelligence/ai-legal-services/ai-regulatory-horizon-tracker>> accessed 06 July 2025

Blake Brittain, ‘Getty Images Lawsuit says Stability AI misused photos to train AI’ (Reuters, 06 February 2023) <<https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/>> accessed 01 August 2025

Blake Brittain, ‘Judge explains order for New York Times in OpenAI copyright case’ (Reuters, 4 April 2025) <<https://www.reuters.com/legal/litigation/judge-explains-order-new-york-times-openai-copyright-case-2025-04-04/>> accessed 3 July 2025

Bobby Allyn, ‘The New York Times takes OpenAI to Court. ChatGPT’s future could be on the line’ (NPR, 14 January 2025) <<https://www.npr.org/2025/01/14/nx-s1-5258952/new-york-times-openai-microsoft>> accessed 02 August 2025

Cade Metz, Cecelia Kang, Sheera Frenkel, Stuart Thompson, and Nico Grant, ‘How Tech Giants Cut Corners to Harvest Data for A.I.’ (New York Times, 8 April 2024) <<https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html>> accessed 2 July 2025

Caitlin Andrews, ‘European Commission receives final version of General-Purpose AI Code of Practice’ (iapp, 10 July 2025) <<https://iapp.org/news/a/european-commission-receives-final-version-of-general-purpose-ai-code-of-practice>> accessed 11 July 2025

Cerys Wyn Davis and Gill Dennis, ‘Getty Images v Stability AI: why the remaining copyright claims are of more than secondary significance’ (Pinsent Mason, 15 July 2025) <<https://www.pinsentmasons.com/out-law/analysis/getty-images-v-stability-ai-copyright-claims-significance>> accessed 06 August 2025

Chris Hanslik, 'Bartz v. Anthropic: A Split Decision That Reshapes AI Copyright Law' (Lexology, 17 July 2025) <<https://www.lexology.com/library/detail.aspx?g=b3d3aa98-6c02-4d1e-b194-f092e5d6a500>> accessed 03 August 2025

Cornell Law School, 'Trademark' (Legal Information Institute, April 2025) <<https://www.law.cornell.edu/wex/trademark>> accessed 26 July 2025

Council of Europe, 'The Framework Convention on Artificial Intelligence' (Council of Europe, 2025) <<https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence>> accessed 07 July 2025

David Morgan and David Shepardson, 'US Senate Strikes AI regulation ban from Trump megabill' (Reuters, 01 July 2025) <<https://www.reuters.com/legal/government/us-senate-strikes-ai-regulation-ban-trump-megabill-2025-07-01/>> accessed 13 July 2025

Desara Dushi, 'How is ChatGPT regulated by the EU AI Act: Reflections on Higher Education' (Global Campus of Human Rights, 05 September 2024) <<https://www.gchumanrights.org/preparedness/how-is-chatgpt-regulated-by-the-eu-ai-act-reflections-on-higher-education/>> accessed 10 July 2025

Diana Milton and Harrison Enright, 'Getty Images v Stability AI' (Baker Hostetler) <<https://www.bakerlaw.com/getty-images-v-stability-ai/>> accessed 26 July 2025

Eleni Courea, 'UK delays plans to regulate AI as ministers seek to align with Trump administration' (The Guardian, 24 February 2025) <<https://www.theguardian.com/technology/2025/feb/24/uk-delays-plans-to-regulate-ai-as-ministers-seek-to-align-with-trump-administration>> accessed 06 August 2025

EUIPO, 'Germany - Hamburg District Court, 310 O.22723, LAION v Robert Kneschke, [27 September 24] (European Union Intellectual Property Office) <<https://www.euipo.europa.eu/en/law/recent-case-law/germany-hamburg-district-court-310-o-22723-laion-v-robert-kneschke>> 31 July 2025

European Commission, 'AI Act' (European Commission, 03 June 2025) <<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>> accessed 10 July 2025

European Commission, 'Commission Signed the Council of Europe Framework Convention on Artificial Intelligence and human rights, democracy and rule of law' (European Commission, 05 September 2024) <<https://digital-strategy.ec.europa.eu/en/news/commission-signed-council-europe-framework-convention-artificial-intelligence-and-human-rights>> accessed 07 July 2025

European Commission, ‘General Purpose AI Code of Practice now Available’ (European Commission, 10 July 2025)

<[https://ec.europa.eu/commission/presscorner/detail/en/ip\\_25\\_1787](https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1787)> accessed 11 July 2025

European Commission, ‘General Purpose AI Models in the AI Act – Questions and Answers’ (European Commission, 12 June 2025) <<https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers>> accessed 10 July 2025

European Innovation Council and SMEs Executive Agency, ‘Artificial Intelligence and copyright: use of generative AI tools to develop new content’ (European Commission, 16 July 2024) <[https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/artificial-intelligence-and-copyright-use-generative-ai-tools-develop-new-content-2024-07-16-0\\_en](https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/artificial-intelligence-and-copyright-use-generative-ai-tools-develop-new-content-2024-07-16-0_en)> accessed 09 August 2025

European Parliament, ‘EU AI Act: first regulation on Artificial Intelligence’ (European Parliament, 08 June 2023)

<<https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence#:~:text=In%20April%202021%2C%20the%20European,risk%20they%20pose%20to%20users.>> accessed 09 July 2025

European Union, ‘Copyright and related rights in the information society’ (European Union, 23 September 2021) <<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=LEGISSUM:l26053>> accessed 17 July 2025

European Union, ‘EU copyright’ (European Union) <[https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legisum:eu\\_copyright](https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legisum:eu_copyright)> accessed 17 July 2025

Grant Gross, ‘What IT leaders want from US AI regulation’ (Cio, 04 September 2024) <<https://www.cio.com/article/3498566/what-it-leaders-want-from-us-ai-regulation.html>> accessed 13 July 2024

Human-Centered Artificial Intelligence, ‘Which countries are leading in AI?’ (Stanford University) <<https://hai.stanford.edu/ai-index/global-vibrancy-tool>> accessed 05 August 2025  
iapp25, ‘Global AI and Policy Tracker’ (iapp 25, May 2025)

<<https://iapp.org/resources/article/global-ai-legislation-tracker/>> accessed 06 July 2025

Irish Computer Society, ‘Comparison of Council of European Framework Convention on Artificial Intelligence with the European Union Artificial Intelligence Act’ (ICS, 29 October 2024) <<https://ics.ie/2024/10/29/comparison-of-council-of-europe-framework-convention-on-artificial-intelligence-with-the-european-union-artificial-intelligence-act/>> accessed 07 July 2025

James Vincent, 'Getty Images sues AI Art generator Stable Diffusion in the US for copyright infringement' (The Verge, 06 February 2023)

<<https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion>> accessed 02 August 2025

Jennifer Rankin, 'EU accused of leaving 'devastating' copyright loophole in AI Act' (The Guardian, 19 February 2025) <<https://www.theguardian.com/technology/2025/feb/19/eu-accused-of-leaving-devastating-copyright-loophole-in-ai-act>> accessed 10 August 2025

Joe El Khoury, 'How Have Pre-Training Datasets for Large Language Models Evolved?' (Medium, 13 July 2024) <<https://medium.com/@jelkhoury880/how-have-pre-training-data-sets-for-large-language-models-evolved-13d74c01f8e8>> accessed 14 July 2024

Katherine Grillaert, Matt Kennedy and Chinasa T. Okolo, 'Risks of State-Led AI Governance in a Federal Policy Vacuum' (Tech Policy Press, 06 February 2025)

<<https://www.techpolicy.press/risks-of-state-led-ai-governance-in-a-federal-policy-vacuum/>> accessed 13 July 2025

Keith Foote, 'A Brief History of Generative AI' (Dataversity, 5 March 2024)

<<https://www.dataversity.net/a-brief-history-of-generative-ai/>> accessed 30 June 2025

Mark Sellman, 'Getty Images sues AI firm in landmark case for copyright law' (The Times, 09 June 2025) <<https://www.thetimes.com/uk/technology-uk/article/getty-images-copyright-trial-sr8s6fk9b>> accessed 01 August 2025

Martin Braun, Anne Valley, Itsiq Benizri and Nancy Stephen, 'Navigating Generative AI Under the European Union's Artificial Intelligence Act' (WilmerHale, 02 October 2024)

<<https://www.wilmerhale.com/en/insights/blogs/wilmerhale-privacy-and-cybersecurity-law/20241002-navigating-generative-ai-under-the-european-unions-artificial-intelligence-act>> accessed 10 July 2025

Maura O'Malley, 'US Judge refuses OpenAI's motion to dismiss New York Times copyright infringement claims' (The Global Legal Post, 08 April 2025)

<<https://www.globallegalpost.com/news/us-judge-refuses-openais-motion-to-dismiss-new-york-times-copyright-infringement-claims-887263879>> accessed 03 August 2025

Min Jun Jung, 'AI Essentials: What are model weights?' (Medium, 10 October 2024)

<<https://engineadvocacyfoundation.medium.com/ai-essentials-what-are-model-weights-2e5b47ec77a1>> accessed 24 July 2025

Nathalie Moreno and Amanda McAllister, 'Key insights into AI regulations in the EU and the US: navigating the evolving landscape' (Kennedys, 21 January 2025)

<<https://kennedyslaw.com/en/thought-leadership/article/2025/key-insights-into-ai-regulations-in-the-eu-and-the-us-navigating-the-evolving-landscape/>> accessed 12 July 2025

Nicole Chen, ‘Stolen Stories or Fair Use? The New York Times v. OpenAI and the Limits of Machine Learning’ (Columbia Undergraduate Law Review, 09 June 2025)

<<https://www.culawreview.org/ddc-x-culr-1/nyt-v-openai-and-microsoft/>> accessed 02 August 2025

OECD, ‘AI Principles’ (OECD) <<https://www.oecd.org/en/topics/sub-issues/ai-principles.html>> accessed 08 July 2025

Practical Law, ‘Anthropic’s Use of Copyrighted Materials to Train GenAI Is Fair Use: N.D. Cal’ (Thomas Reuters Practical Law, 25 June 2025)

<[https://uk.practicallaw.thomsonreuters.com/w-047-4069?transitionType=Default&contextData=\(sc.Default\)&firstPage=true](https://uk.practicallaw.thomsonreuters.com/w-047-4069?transitionType=Default&contextData=(sc.Default)&firstPage=true)> accessed 04 August 2025

Rachel Reed, ‘ChatNYT ’ (Harvard Law Today, 22 March 2024)

<<https://hls.harvard.edu/today/does-chatgpt-violate-new-york-times-copyrights/>> accessed 02 August 2025

Rebecca Peknam-Walsh, ‘AI and copyright – where are we now?’ (Fieldfisher, 10 June 2025) <<https://www.fieldfisher.com/en/services/intellectual-property/intellectual-property-blog/ai-and-copyright-where-are-we-now>> accessed 06 August 2025

Reuters, ‘Sam Altman’s OpenAI signs content agreement with News Corp’ (Reuters, 22 May 2024) <<https://www.reuters.com/technology/sam-altmans-openai-signs-content-agreement-with-news-corp-2024-05-22/>> accessed 09 August 2024

Seagull Song, ‘China’s First Case on AIGC Output Infringement – Ultraman’ (King & Wood Mallesons, 28 February 2024) <<https://www.kwm.com/cn/en/insights/latest-thinking/china-s-first-case-on-aigc-output-infringement-ultraman.html>> accessed 04 August 2025

Sonya Huang, Pat Grady and GPT-3, ‘Generative AI: A Creative New World’ (Sequoia, 19 September 2022) <<https://www.sequoiacap.com/article/generative-ai-a-creative-new-world/>> accessed 02 May 2025

Stanford HAI staff, ‘Global AI Power Rankings’ (Stanford University Human-Centred Artificial Intelligence, 21 November 2024) <<https://hai.stanford.edu/news/global-ai-power-rankings-stanford-hai-tool-ranks-36-countries-in-ai>> accessed 27 July 2025

The Bureau of Democracy, Human Rights and Labor, ‘The Council of Europe’s Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law’ (U.S. Department of State) <<https://www.state.gov/bureau-of-democracy-human-rights-and->

[labor/the-council-of-europes-framework-convention-on-artificial-intelligence-and-human-rights-democracy-and-the-rule-of-law#:~:text=The%20Convention%20defines%20a%20baseline,rights%20obligations%20to%20AI%20activities.>](#) accessed 13 July 2025

The University of Edinburgh, ‘Copyright Exceptions and Fair Dealing’ (The University of Edinburgh, 21 August 2024) <<https://library.ed.ac.uk/library-help/copyright/copyright-exceptions-and-fair-dealing>> accessed 08 August 2025

Thomas Reuters Foundation, ‘AI governance in the United States’ (Thomas Reuters Foundation, 07 May 2025) <[https://www.trust.org/toolkit/ai-governance-for-africa-part-1/7-ai-governance-in-united-states/#\\_ftn1](https://www.trust.org/toolkit/ai-governance-for-africa-part-1/7-ai-governance-in-united-states/#_ftn1)> accessed 12 July 2025

Tom Krantz, ‘What is the AI Bill of Rights’ (IBM, 27 September 2024) <<https://www.ibm.com/think/topics/ai-bill-of-rights>> accessed 12 July 2024

US Senate Committee on Commerce, Science and Transportation, ‘Cantwell, Blackburn, Heinrich Introduce Legislation to Increase Transparency, Combat AI Deepfakes & Put Journalists, Artists & Songwriters Back in Control of Their Content’ (US Senate Committee on Commerce, Science and Transportation, 11 July 2024) <<https://www.commerce.senate.gov/2024/7/cantwell-blackburn-heinrich-introduce-legislation-to-combat-ai-deepfakes-put-journalists-artists-songwriters-back-in-control-of-their-content>> accessed 13 July 2025

US Senate Committee on Commerce, Science and Transportation, ‘Cantwell, Blackburn, Heinrich Reintroduce Bipartisan Bill to Increase Transparency, Combat AI Deepfakes & Put Journalists, Artists & Songwriters Back in Control of Their Content’ (US Senate Committee on Commerce, Science and Transportation, 10 April 2025) <<https://www.commerce.senate.gov/2025/4/cantwell-blackburn-heinrich-reintroduce-bipartisan-bill-to-increase-transparency-combat-ai-deepfakes-put-journalists-artists-songwriters-back-in-control-of-their-content>> accessed 13 July 2025

White & Case, ‘AI Watch: Global Regulatory tracker – China’ (White & Case, 29 May 2025) <<https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-china>> accessed 05 August 2025

White & Case, ‘AI Watch: Global Regulatory tracker – United Kingdom’ (White & Case, 25 March 2025) <<https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-kingdom>> accessed 06 August 2025

White & Case, ‘AI Watch: Global Regulatory Tracker’ (White & Case, 2025)

<<https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker#home>>

accessed 06 July 2025

WIPO, ‘Germany DE048-j 2024 WIPO IP Judges Forum Informal Case Summary –

Hamburg Regional Court, Germany [2024]: Robert Kneschke v LAION e.V. Case No. 310

227/23’ (WIPO) <<https://www.wipo.int/wipolex/en/text/592042>> accessed 31 July 2025

Yunah Kwon, ‘Copyright in the Age of Generative AI: Legal Gaps and Artist Vulnerability’

(Columbia Undergraduate Law Review, 17 June 2025)

<<https://www.culawreview.org/current-events-2/copyright-in-the-age-of-generative-ai-legal-gaps-and-artist-vulnerability>> accessed 03 August 2025

Zack Schor, ‘Andersen v Stability AI: The Landmark Case Unpacking the Copyright Risks of

AI Image Generators’ (NYU Journal of Intellectual Property & Environmental Law, 2

December 2024) <[https://jipel.law.nyu.edu/andersen-v-stability-ai-the-landmark-case-](https://jipel.law.nyu.edu/andersen-v-stability-ai-the-landmark-case-unpacking-the-copyright-risks-of-ai-image-generators/)

[unpacking-the-copyright-risks-of-ai-image-generators/](https://jipel.law.nyu.edu/andersen-v-stability-ai-the-landmark-case-unpacking-the-copyright-risks-of-ai-image-generators/)> accessed 05 July 2025