Regulating AI: Differences Between the U.S. and the EU

*Raquel Xalabarder**

The topic for this lecture is very much about the future. Basically, we are talking about things we are only starting to see. So, it may still be premature to talk about regulating Artificial Intelligence ("AI"). Not too long ago, several experts and tycoons in the AI community issued an open letter last year saying, in effect, "let's pause it a bit. Let's see how to regulate it."[1]

Since then, we have seen the rise of ever more forms of AI, notably ChatGPT, Stable Diffusion, and all these AI systems and processes of diffusing images, adding noise to them and then de-noising them, so as to subsequently generate a different image or a different audiovisual recording at the request of a prompt.[2] Creators and professionals are concerned about that. And so are we, professors. In fact, AI seems to be a game changer in the world of copyright. We are not Luddites but we are starting to sympathize with them (if I may put it that way). This is how much AI is shaking our lives.

Let me start by disclosing that I am going to pose more questions than answers about these AI technologies that have come to disrupt our lives. AI is not new, per se. We have been talking about AI for decades. But we are now seeing an explosion of AI systems, resulting from a specific combination of enhanced computing power and

* Raquel Xalabarder, Chair of Intellectual Property, Universitat Oberta de Catalunya. This Article was first delivered as the 35th Annual Horace S. Manges Lecture, given at Columbia Law School on April 17, 2023.

1. *Pause Giant AI Experiments: An Open Letter*, FUTURE OF LIFE INST. (Mar. 22, 2023), https://futureoflife.org/wp-content/uploads/2023/05/FLI_Pause-Giant-AI-Experiments_An-Open-Letter.pdf [https://perma.cc/MV7P-AWUV] [https://web.archive.org/save/https://futureoflife.org/wp-content/uploads/2023/05/FLI_Pause-Giant-AI-Experiments_An-Open-Letter.pdf].

2. To generate an output from an input image, Stable Diffusion "adds noise" to the input, based on a "seed" (a number to be used by its noise generator algorithm). The amount of noise added (Denoising Strength) determines how close the output image will be to the input one. Outputs will depend on the seed number and Denoising Strength, as well as on the prompt and parameters applied to obtain them. For more explanations and examples, see *Guide: What Is a Stable Diffusion Seed and How To Use It*, ONCE UPON AN ALGORITHM (Mar. 14, 2023), https://onceuponanalgorithm.org/guide-what-is-a-stable-diffusion-seed-and-how-to-use-it/ [https://perma.cc/N4HN-VRY3] [https://web.archive.org/web/20250404215110/https://onceuponanalgorithm.org/guide-what-is-a-stable-diffusion-seed-and-how-to-use-it/].

connectivity, on the one hand, and an enormous increase of data (big data) available in digital and accessible formats, on the other. This combination has fostered the development of algorithms capable of learning fast (trained with data) and making further connections (through neuronal networks).

AI being the all-encompassing term, we use the term "machine learning" to refer to the process of training algorithms with further data and, specifically, to "deep learning" when these machine-learning algorithms involve multilayered neuronal networks capable of reading vast amounts of data. The more data we feed them, the more training data and of better quality, the better and more precise the outputs from these AI systems, the more reliable or valuable their outputs would be. And let us not forget that this data is not simply data. It is very often copyrighted works, performances, recordings.

But AI systems may be of very different kinds. For the purposes of today's lecture, I would like to distinguish among two kinds: *functional artificial intelligence systems* and *generative artificial intelligence systems*. The former, to refer to algorithms which are meant to fulfil a task, a function: translate, locate, identify patterns, obtain information, etc. For instance, internet search engines or dedicated search engines like the Google Books Project; another example: programs that transcribe handwritten texts. These programs are meant to achieve a task, to do something. By contrast, generative AI is meant to produce new outputs, new contents, new software (apparently, ChatGPT is very good at writing code, better than people who write code). Generative AI can produce "performances," "recordings," news, legal texts, such as case briefs (though the tendency of ChatGPT to invent case citations shows how wrong some of these outputs can go[3]), as well as new "works" (if you let me use the term without vesting a copyright significance).

Generative AI can also produce novels and essays. ChatGPT may not (not yet?) be very good at being correct and accurate, but it is very good at inventing and generating very well-written texts, often better than students' writings—I can confirm that. What about works of art, images, comic books, and graphic novels? New comic books and graphic novels have already been generated by Stable Diffusion and Midjourney.[4]

From a copyright perspective, this whole process poses at least three main questions:

(1) The use of copyrighted contents as inputs for the training of the AI systems: machine learning or text and data mining ("TDM"), as we call it in the European Union. Is this an infringement? Does it need a license? Can authors oppose it?

---

3. Benjamin Weiser, *ChatGPT Lawyers Are Ordered To Consider Seeking Forgiveness*, N.Y. TIMES (June 22, 2023), https://www.nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html?searchResultPosition=2 [https://web.archive.org/save/https://www.nytimes.com/2023/06/22/nyregion/lawyers-chatgpt-schwartz-loduca.html?searchResultPosition=2] (reporting on a S.D.N.Y. judge fining lawyers who filed a brief with spurious citations composed by ChatGPT).

4. Letter re: *Zarya of the Dawn* (Registration #VAu001480196), U.S. COPYRIGHT OFF. (Feb. 21, 2023), https://www.copyright.gov/docs/zarya-of-the-dawn.pdf [https://perma.cc/GV2T-EGWW] [https://web.archive.org/save/https://www.copyright.gov/docs/zarya-of-the-dawn.pdf] (canceling the copyright registration for *Zarya of the Dawn* after determining that its images were generated using Midjourney and issuing a new registration for the expressive elements original to the author).

(2) The outputs resulting from these AI systems, which are sometimes very similar to the inputs fed to the algorithm: Are they infringing contents? Should they be licensed? Should perhaps the answer be different depending on whether they were created by functional or generative AI?

(3) What about generative AI outputs: Can they (should they?) be protected at all? If so, who owns them? As interesting as it is, we will not have time to examine this last topic today.

And here goes another disclaimer: It is impossible to draw general conclusions since probably a case-by-case analysis would be required at all times.

So let us start with TDM to train the AI system. We take for granted that machine reading, TDM, implies a reproduction of the works, performances, recordings; anything that is protected under copyright and is ingested by the system is being, at least, reproduced. Of course, machine reading also implies reproduction of raw data. But I am not going to deal with raw data. Data as such is not protected by copyright.[5]

Machine learning also involves disaggregating the work or recording; for instance, in order to identify grammatical components such as nouns, adjectives, or verbs or in order to allow the process of "noising and denoising" used by generative AI. Furthermore, the resulting data is then stored, communicated to the public (or at least, made available) in order for the AI system to work. Such actions happen as part of the machine learning process and, from a copyright perspective, they go beyond simple act of reproduction itself. So, as long as machine learning deals with copyrighted works, does it need a license? If there is a limitation or an exception in our laws that permits such training (TDM, in Europe), should we pay attention to the scope of this exception and run the specific machine learning activities through its filter? Is there a statutory license or an authorization granted by law in some specific cases, such as for scientific research purposes? And, if there is no license (voluntary or statutory), is machine learning based on copyrighted works an infringement? In the United States, the relevant question should be: Does the fair use doctrine provide a defense against all scenarios of machine learning? These are the questions we can discuss today. But before that, and just for the sake of argument, let us pause for a moment and consider that machine learning may not be an act of exploitation subject to the exclusive rights of the author or copyright owner.

Some professors, especially in Europe, defend this position. Professor Alain Strowel proposes that we should only be talking about exploitation when the work is being used as a work.[6] A bit like what happens with descriptive uses of trademarks: One can use a

5. The European Union protects data through a *sui generis* right for databases, which grants database producers the authority to control (prohibit or authorize) the extraction and reuse of substantial or entire parts of a database. *See* Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the Legal Protection of Databases, 1996 O.J. (L 77) 20. Works, performances, recordings, i.e., everything that is protected under copyright, may be reproduced as part of algorithmic training in machine learning.

6. *See* Rossana Ducato & Alain M. Strowel, *Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out*, 43 EUR. INTELL. PROP. REV. 322, 334 (2021) (suggesting "an infringement test requiring that for the reproduction right to be infringed the *work* should be used *as a work* and perceived as a work by a public").

trademark to identify the trademark owner or its goods, but not to identify the source of one's own goods.[7] A competitor can refer to Coca-Cola in an advertisement asserting that the competitor's beverage tastes better, but the competitor cannot call its beverage Coca-Cola. The argument asserts that we should recognize an analogous exception to copyright because a machine reading a work does not involve an act of exploitation of this work when the AI program is merely recording the work as data, rather than using the work as it should be used, i.e., as an expressive work of authorship.

There is an exception or an exemption in Japanese law that is consistent with this approach. Under the Japanese TDM exception, acts of exploitation of works that are not for purposes of enjoying the work as a work will not be subject to the author's exclusive rights.[8] We will get back to this later on. For now, I just want to point out the fact that, on both sides of the Atlantic, we have all taken for granted that machine learning, TDM, involves at least an act of exploitation, of reproduction. From that starting point, different jurisdictions have dealt differently with the consequences of the initial reproduction into training data.

Let us start with Europe. The European Union would like to be a worldwide hub for AI. In 2018, the European Union started issuing several communications about AI.[9] Every few months, there has been a new communication by the European Commission, including Artificial Intelligence for Europe, [10] Coordinated Plan on Artificial Intelligence ("AI made in Europe"),[11] and Building Trust in Human-Centric Artificial Intelligence.[12] The High-Level Expert Group on AI also issued Ethics Guidelines for Trustworthy Artificial Intelligence.[13] The Commission released a regulatory proposal, known as the Artificial Intelligence Act ("AI Act").[14] The AI Act identifies different risk categories for the development of an AI system. AI systems that present an *unacceptable*

7. In U.S. trademark law, this is called "nominative fair use." *See* New Kids on the Block v. News Am. Publ'g, 971 F.2d 302, 308 (9th Cir. 1992). *See also* 15 U.S.C. § 1115(b)(4) (providing that "use, otherwise than as a mark . . . which is descriptive and used fairly and in good faith only to describe the goods or services of [the user]" constitutes a defense to trademark infringement).

8. *See* Chosakukenhō [Copyright Act], Law No. 48 of 1970 (as amended up to Jan. 1, 2022), art. 30-4 (Japan) [hereinafter, Japanese Copyright Act], *translated in* WIPO Lex, https://www.wipo.int/wipolex/en/text/584874 [https://perma.cc/S25W-ZBY2] [https://web.archive.org/save/https://www.wipo.int/wipolex/en/text/584874].

9. *See, e.g.,* Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe, COM (2018) 237 final (Apr. 25, 2018).

10. *Id.*

11. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Coordinated Plan on Artificial Intelligence, COM (2018) 795 final (July 12, 2018).

12. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: Building Trust in Human-Centric Artificial Intelligence, COM (2019) 168 final (Apr. 8, 2019).

13. HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE, ETHICS GUIDELINES FOR TRUSTWORTHY AI (Apr. 8, 2019).

14. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 final (Apr. 21, 2021).

risk will not be allowed, at all. For instance, the Chinese government's use of AI systems to control and score people's activity is an unacceptable risk. *High risk* includes, for instance, facial recognition and the AI Act also identifies liability conditions for the development of these "high risk" AI systems. The other permitted categories are *limited risk* applications and *minimal risk* applications. These risk categories follow the European Union's approach to personal data protection set out in the General Data Protection Regulation ("GDPR"). The GDPR identified different kinds of personal data, and depending on the risk involved in using that personal data, implemented different requirements and conditions to use that data.[15]

We are not talking about AI and copyright only; several other regulations may also have an impact on AI. In addition to the GDPR, we're talking about the Digital Services Act,[16] the Digital Markets Act,[17] the Data Governance Act,[18] and the Open Data Directive,[19] which in the old times used to be a Public Sector Information Directive. And now, there is a proposal on a new Data Act. How to navigate among all these regulations and succeed in implementing an AI project will certainly keep lawyers and consulting companies busy for a while.

But let us focus on TDM. The DSM Directive specifically concerns copyright and contains two exceptions and/or limitations that specifically refer to TDM.[20] The exception in Article 3 specifically allows TDM for scientific research purposes.[21] Under Article 4, member states may set exceptions or limitations allowing TDM for any other purposes, beyond scientific research.[22] Under Article 3, member states may not provide any compensation scheme to the copyright owners of the source works.[23] Article 3 is therefore a true exception, rather than a limitation subject to compensation.[24] In

15. Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1.

16. Regulation 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L 277) 1.

17. Regulation 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), 2022 O.J. (L 265) 1.

18. Regulation 2022/868 of the European Parliament and of the Council of 30 May 2022 on European Data Governance and Amending Regulation (EU) 2018/1724 (Data Governance Act), 2022 O.J. (L 152) 1.

19. Directive 2019/1024 of the European Parliament and of the Council of 20 June 2019 on Open Data and the Re-use of Public Sector Information (recast), 2019 O.J. (L 172) 56.

20. Directive 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, 2019 O.J. (L 130) 92 [hereinafter, DSM Directive].

21. *See id.* art. 3.

22. *See id.* art. 4.

23. *See id.* recital (17) ("In view of the nature and scope of the exception, which is limited to entities carrying out scientific research, any potential harm created to rightholders through this exception would be minimal. Member states should, therefore, not provide for compensation for rightholders as regards uses under the text and data mining exceptions introduced by this Directive.").

24. In Europe, we have lots of limitations (compensated exceptions) such as for private copying. *See, e.g.*, Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the

Article 4, by contrast, the European Union leaves member states the possibility to decide which TDM purposes should be authorized by law and whether or not to subject it to a compensation.[25]

Turning to the text of the DSM Directive's provisions on TDM. First, the definition of TDM is very broad: "'[T]ext and data mining' means any automated analytical technique aimed at analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations."[26] I am stopping at this definition because perhaps we could start questioning whether the 2019 DSM Directive is really prepared to exempt generative AI. Recital 8 adds that the text and data to be analyzed could be "text, sounds, images, or data," i.e., anything goes as to the subject matter covered by TDM.[27] The goal is to gain knowledge, to generate information, patterns, trends, correlations, to enable the "processing of large amounts of information with a view to gaining new knowledge and discovering new trends possible."[28] In other words, neither Article 3 nor its accompanying Recitals indicate anything close to generate new products, rather to "only" obtain information and knowledge. *De lege lata*, it makes sense, since the DSM Directive far predates the raise of generative AI. *De lege ferenda*, should we apply its provisions to deal with TDM for generative AI?

Under Article 3, the TDM exception for scientific research purposes authorizes reproduction of works and sound recordings (which in Europe, the latter are the subject of a related right), and the extraction and re-utilization of databases not covered by copyright but by the *sui generis* right.[29] Article 3 also authorizes the reproduction and making available of works covered by a new ancillary right given to press publishers, for two years, to control "online use by information society service providers" of their press publications. Research organizations and cultural heritage institutions are the beneficiaries of this uniform exception.

The term "research organizations" seems to exclude individual researchers. "Research organizations" are defined as "a university, including its libraries, a research institute, or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research . . . on a not-for-profit basis. ." and can include public-private partnerships.[30] Accordingly, a TDM project funded by private investment could also benefit from this exception, as long as the research organization has decisive influence over how the

Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society, art. 5, 2001 O.J. (L 167).

25. *See* DSM Directive, *supra* note 20, art. 4.

26. *Id.* art. 2(2).

27. *Id.* recital (8) ("New technologies enable the automated computational analysis of information in digital form, such as text, sounds, images or data, generally known as text and data mining. Text and data mining makes the processing of large amounts of information with a view to gaining new knowledge and discovering new trends possible.").

28. *Id.*

29. *See* Directive 96/9/EC, *supra* note 5, for an explanation of the sui generis right.

30. DSM Directive, *supra* note 20, art. 2(1).

results of the TDM activity are going to be used.[31] This is, on the one hand, very broad but, on the other, quite restrictive in that it is limited to nonprofit research purposes. That said, the public-private partnerships allowed by the Directive often have blurred lines. One may be developing TDM projects on a research organization funded by private money, but it is very difficult to say that the results of this TDM are not going to end up in the private hands of those who invested in it. Additionally, cultural heritage institutions "means a publicly accessible library or museum, an archive or a film or audio heritage institution."[32] The DSM Directive does not explicitly say so, but it is plausible to think that the same public-private partnership could also apply to these cultural heritage institutions.

This exemption to conduct TDM applies only to materials to which these institutions (research organizations, universities, research centers, libraries, museums, archives) have lawful access.[33] What is lawful access? This includes licensed contents or subscription access to a database, as well as open access, be it under a license, such as a Creative Commons license, or simply posted on the internet, and declared freely available to the public. Recital 14 notes that lawful access includes works freely available online.[34] In other words, as long as access is unrestricted or licensed, that content will be capable of being subject to TDM purposes for research purposes by research organizations or museums, etc.

This is a uniform mandatory exception, with a clear mandatory scope. All the national laws in the European Union must say the same. Why? Because the potential harm of this exception to do TDM for research purposes, by research institutions, is minimal. Hence the EU legislature's determination that the exception will be uniform and uncompensated across all national laws. Research organizations can do this TDM for free, and member states cannot provide compensation.

Furthermore, the exception cannot be overruled by contracts. If content has been licensed by subscription, the licensor cannot bar the licensee from conducting TDM; such a contractual clause would be unenforceable. In other words, anything that has been lawfully accessed or licensed can be used for TDM (machine-learning) purposes. However, the DSM Directive doesn't say anything about the use of technological protection measures ("TPM"), which remains one of the very difficult issues in the European Union. It is easy to foresee that technological protection measures could de facto prevent research organizations and cultural institutions from benefiting from the exception.[35] That result is problematic in light of the mandatory character of the

31. See *id.*

32. *Id.* art. 2(3) ("'[C]ultural heritage institution' means a publicly accessible library or museum, an archive or a film or audio heritage institution[.]").

33. *See id.*, art. 3(1); *see also id.*, recital (33). Lawful access should be understood as covering access to content based on an open access policy or through contractual arrangements between rightsholders and research organizations or cultural heritage institutions, such as subscriptions, or through other lawful means. For instance, in the case of subscriptions taken by research organizations or cultural heritage institutions, the persons attached thereto and covered by those subscriptions should be deemed to have lawful access. Lawful access should also cover access to content that is freely available online.

34. *See id.* recital (14).

35. *See id.* recital (7).

exception. If the exception cannot be contracted out of, then TPMs should not neutralize the exception's operation, either. The Directive, however, does not explicitly prohibit the use of TPMs. The Court of Justice of the European Union ("CJEU") will likely be presented with cases to clarify many of these issues.

Additionally, storage is allowed.[36] Thus, all the results from the TDM may be stored for later verification of the research results. Specific protection measures should be applied to ensure that these TDM results will be kept safe and will not be spread beyond the research organization.[37] Furthermore, member states should encourage commonly agreed best practices to keep these results safe and how to conduct TDM.[38]

And one last comment regarding Article 3 (and, as we shall see, Article 4): These TDM articles only provide an exception or limitation to the reproduction right for works and sound recordings, but not to the right of communication to the public. Paragraph 1 references "Article 2 of Directive 2001/29/EC [the Infosoc Directive]" which concerns the right of reproduction, but not Article 3 of that Directive, which covers the right of communication to the public. This brings us back to the initial comment regarding the whole purpose of TDM as envisioned by the DSM Directive (to obtain information and knowledge), according to the TDM technology available at the time of its drafting and approval. We will consider the significance of the omission of the right of communication to the public when we discuss Article 4.

Turning now to Article 4, it covers TDM for any other purposes beyond scientific research. Article 4 involves startups, journalists, individual researchers (remember they were not in Article 3), search engines, translation tools, and marketing. Search engines fall within Article 4 because, as far as I am aware, there is no national law in the European Union that allows or that has any exception or limitation for search engines. So, while in the United States, the copying in which search engines engage may be fair use,[39] the functioning of search engines according to European copyright laws remains in a very gray zone.

In Spain, we had a beautiful case that dealt with the Google search engine. The plaintiffs tried to shut down the Google search engine, claiming that Google was showing their contents and web page, and that they had not authorized Google to link, search, and show snippets of its contents. The Spanish Supreme Court agreed with the plaintiff that there is no exception or limitation in the Spanish copyright law saying

36. *See id.* art. 3(2) ("Copies of works or other subject matter made in compliance with paragraph 1 shall be stored with an appropriate level of security and may be retained for the purposes of scientific research, including for the verification of research results.").

37. *See id.*, art. 3(3) ("Rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted. Such measures shall not go beyond what is necessary to achieve that objective.").

38. *See id.* art. 3(4) ("Member States shall encourage rightholders, research organisations and cultural heritage institutions to define commonly agreed best practices concerning the application of the obligation and of the measures referred to in paragraphs 2 and 3 respectively.").

39. *Compare* Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007) (holding that copying images found on the internet and creating thumbnails was fair use), *with* VHT, Inc. v. Zillow Grp., Inc., 918 F.3d 723, 742 (9th Cir. 2019) (holding that the copying of plaintiff's photographs were not fair use because "the label 'search engine' is not a talismanic term that serves as an on-off switch as to fair use").

that search engines can search and can identify or show a bit of your work.[40] However, the Spanish Supreme Court found that Google's conduct was not infringing, and that the plaintiffs had gone too far and abused their copyright.[41] By the same token, I wonder whether a court might conclude that it would be an abuse of copyright to prevent showing results from a TDM algorithm or a TDM project.

While Article 4 is a mandatory provision for member states, they may choose to impose compensation for the scope of TDM chosen to be permitted by law. As far as I know, there is no national law that does this. Seeing that the TDM coverage under Article 4 is quite broad, encompassing TDM done by anyone for any purposes, could one expect that Article 4 may be more favorable towards generative AI? In fact, Recital 18 notes that TDM is used for "government services, complex business decisions, and the development of new applications or technologies."[42] Could this include generative AI? Even then, we would be facing the same problem as before: The TDM exception might allow for the creation of new content through generative AI, but the exception would not authorize the communication to the public of any outputs that incorporate copied content because Article 4, like Article 3, only exempts and limits the right of reproduction.

Like Article 3, Article 4 is limited to lawfully accessible content. But Article 4 presents a big difference in comparison with Article 3, in addition to the breadth of its beneficiaries. Article 4 "shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their right holders in an appropriate manner, such as machine-readable means in the case of content made publicly available online."[43] In other words, right holders may "opt out" of the application of the exception or limitation. Instead, since Article 3 cannot be opted out, Article 4 explicitly states that it "shall not affect the application of Article 3 of this Directive."[44] Research organizations will always be able to do TDM using any content to which they have lawful access. (In addition, the temporary acts of reproduction authorized by Article 5(1) of the Infosoc Directive that do not have an independent economic significance remain permissible under Article 4.[45])

How can this opt-out reservation be administered in an appropriate manner? Recital 18 references machine-readable means, such as robot exclusion protocols, robot texts, metadata, terms and conditions on a website or a service, contractual agreements, or,

40. S.T.S., Apr. 3, 2012 (T.S., No. 3942/2012, p. 1, 4) (Spain); *see also* Raquel Xalabarder, *Spanish Supreme Court Rules in Favour of Google Search Engine. . . and a Flexible Reading of Copyright Statutes?*, 3 J. INTELL. PROP., INFO. TECH., & ELEC. COM. L. 162, 163 (2012).

41. Xalabarder, *supra* note 40, at 164–65. The Court also found that the conduct at issue fell within the doctrine of *ius usus inocui*, or the "right to make a harmless use of someone else's property."

42. DSM Directive, *supra* note 20, recital (18).

43. *Id.*, art. 4.

44. *Id.*

45. *See* id. recital (9) ("There can also be instances of text and data mining that do not involve acts of reproduction or where the reproductions made fall under the mandatory exception for temporary acts of reproduction provided for in Article 5(1) of Directive 2001/29/EC, which should continue to apply to text and data mining techniques that do not involve the making of copies beyond the scope of that exception.").

basically, a unilateral declaration on a website.[46] Additionally, it is not clear whether the opt-out reservation in Article 4 should also apply to content posted before the 2019 Directive promulgation (or its 2021 deadline for implementation)? That would seem fundamentally unfair since authors who posted content before 2019 did not know that they had to reserve their copyrights so that their works would not be used for TDM purposes. I anticipate the CJEU may also have to resolve this retroactivity issue at some point.

In summary, Article 3 is very straightforward: It is a mandatory exception, and member states must adopt it as is. By contrast, Article 4 leaves a lot of leeway for member states, yet most of them have failed to make use of its full potential as it can be seen now that the Directive has been implemented in almost all the member states. Let me start with its awkward implementation in Spain.

Spain has implemented Articles 3 and 4 in a single exception without distinguishing between research uses or other uses for TDM.[47] Thus, TDM of copyrighted content is allowed for all TDM purposes in Spain as long as it has not been reserved by the copyright owner. Additionally, no special treatment exists for scientific research TDM: A copyright owner in Spain could prohibit the use of any TDM including for scientific research purposes (thus contravening Article 3 of the DSM Directive). I anticipate that, unless revised, Spain's implementation may be challenged at some point in front of the CJEU.

By comparison, the United Kingdom, which is no longer in the European Union and has no obligation to implement the DSM Directive (because that text post-dates Brexit),[48] permits TDM subject to a requirement of lawful access, and only for non-commercial research purposes. The United Kingdom was tempted to expand this exception to other users but decided not to.[49]

46. *See* id. recital (18) ("In the case of content that has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine-readable means, including metadata and terms and conditions of a website or a service.").

47. C.E., B.O.E. N. 263, Nov. 3, 2021 (Spain).

48. The United Kingdom did not implement the DSM Directive because the implementation deadline by EU member states expired after both the United Kingdom's exit from the European Union and the subsequent implementation period, when EU law by and large still applied in the United Kingdom. *See* Chris Skidmore, Minister of State for Business, Energy and Industrial Strategy, Written Answer to Question 4371 (Jan. 21, 2020), https://questions-statements.parliament.uk/written-questions/detail/2020-01-16/4371 (stating that "[a]ny future changes to the UK copyright framework will be considered as part of the usual domestic policy process").

49. *See* Intell. Prop. Off., Artificial Intelligence and Intellectual Property: Copyright and Patents: Government Response to Consultation (June 28, 2022) (U.K.) (announcing plans to broaden the TDM exception for any purpose, with copyright protection limited to lawful access); *see also* HC Deb (Feb. 1, 2023) (727) cols. 162–63 (George Freeman, MP, Minister for Science, Research and Innovation, noting that that the proposal would not be going forward after receiving a "huge response" requiring more time to "get the balance right").

In Germany, TDM of lawfully accessed contents is also permitted.[50] In Switzerland, which is not an EU country, TDM is permitted for scientific research only, with or without commercial purposes, which is wider than the DSM Directive.[51]

By comparison, in the United States TDM does not need a specific statutory exception or a limitation because of the fair use exception. TDM and machine learning may be permitted as a fair use under U.S. law. There are rulings dealing with fair use, including cases which consider considering copying the whole image contents of the internet in order to index for search engines, and show thumbnails to identify the search results,[52] or scanning whole books to index, search, and show snippets.[53] If the fair use defense did not apply, all of these actions could be infringing because they are reproducing and communicating to the public without an authorization by the copyright owner.

As the Supreme Court stated in *Campbell v. Acuff-Rose*, "some opportunity for fair use of copyrighted materials has been thought necessary to fulfill copyright's very purpose, 'to promote the Progress of Science and useful Arts.'"[54]

Let us consider the application of the previously-mentioned Google Books case in that light. Google's scanning and storing in a searchable database of millions of in-copyright books and delivering "snippets" responsive to search terms enabled users to identify books that contained the content they sought. Judge Pierre Leval affirmed the lower court's ruling that providing users with the ability to search large texts and be shown snippets of books that meet the search criteria changed the text's purpose and character, in delivering information *about* the relevant books.[55] The service provided by Google made a "transformative use" of the copied books because it informed, identified, and located books, without impinging on the authors' rights to make derivative works. In providing a new service, Google was not exploiting works as works. Its new service added value (and could even help identify books that were already forgotten) but did not supplant or supersede the market for the scanned works. The significant public benefits of the project were sufficient to trump the infringement claim, the unauthorized reproduction and communication to the public that was being done by Google. And it is in that sense of providing a new service, that the requirement of "transformative use" may be understood, rather than providing new derivative

50. Gesetz über Urheberrecht und verwandte Schutzrechte [Act on Copyright and Related Rights], *translated in* the Federal Office of Justice, https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html [https://perma.cc/2BGQ-5B4N] [https://web.archive.org/web/20250219045145/https://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html] (Ger.).

51. Bundesgesetz über das Urheberrecht und verwandte Schutzrechte vom 9. Oktober 1992 [Federal Act of October 9, 1992, on Copyright and Related Rights], *translated in* WIPO Lex, https://www.wipo.int/wipolex/en/text/584729 [https://perma.cc/E9JW-JMTP] [https://web.archive.org/web/20250307192807/https://www.wipo.int/wipolex/en/text/584729] (Switz.).

52. *See, e.g.*, Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003); Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146 (9th Cir. 2007).

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

54. Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 575 (1994) (quoting U.S. Const. art. I, § 8, cl. 8).

55. *Authors Guild*, 804 F.3d at 217–18.

works, as the Google Books ruling seems to be read, which, clearly, conflicts with the very scope of the exclusive rights.

So, if we apply Google Books to what I called functional (not generative) AI, it may easily qualify as a fair use. That would be my bet and I am sure courts in the United States will be called upon to resolve cases of that kind.

But let me move on to Japanese law. Basically, Japanese law permits acts of exploitation that are necessary to achieve three goals, as long as these acts are not "for the enjoyment of the works used": for testing technological developments; for (any) TDM purposes; and for playing music to animals or plants (e.g., to make them more productive).[56] All of these uses are permitted also for commercial purposes, as long as they do not "unreasonably prejudice the interests of the copyright holder."[57] It is unclear whether or not Article 30 would allow the use of copyrighted works in AI outputs because as soon as the algorithm issues these outputs, someone is enjoying the work, at least if the work is substantially reproduced in the outputs.

Furthermore, Japanese law does not allow authors to opt out of having their works harvested and there is no requirement in Japanese law that the copied materials have been lawfully accessed. Japan, therefore, may become a safe haven for machine learning and text and data mining.[58]

So, it seems, for now, that Japan has the most welcoming environment for technology companies seeking to text and data mine and scrape the web in other ways. The United States may be also leaning towards prioritizing technological development over copyright, at least with respect to "functional AI" ("generative" AI remains to be seen). And lastly, the European Union seems to be the outlier here because statutory authorization for TDM (and perhaps other machine reading processes) is solely envisaged for scientific research purposes.

Far more difficult to predict, is whether machine learning exceptions and fair use defenses can also apply to the training of generative AI**.** In the United States, courts will have to decide soon whether machine learning to train Generative AI systems is also the kind of "transformative use" exempted under fair use, which—as always—will very much depend on the specific facts and circumstances of each case. On the one hand, the fair use case for generative AI outputs may be stronger if the outputs do not substantially reproduce the inputs (copyrighted works and performances). Admittedly, this seems to be a paramount consideration after the Supreme Court's 2023 decision in *Andy Warhol Foundation v. Goldsmith.*[59] But the Court has also emphasized that the need

56. *See* GOV'T OF JAPAN, INTEGRATED INNOVATION STRATEGY 2022 (June 30, 2022), https://www8.cao.go.jp/cstp/english/strategy_2022.pdf [https://perma.cc/3B2H-6ED6] [https://web.archive.org/web/20250307194534/https://www8.cao.go.jp/cstp/english/strategy_2022.pdf] (setting forth three areas for growth, including promoting advanced and emerging technologies).

57. Japanese Copyright Act, *supra* note 8, art. 30.

58. Min Jeong Lee & Natsuko Katsuki, *Google Alums Go Big in Japan with Months-Old Startup Sakana AI*, BLOOMBERG LAW (Apr. 11, 2024), https://www.bloomberglaw.com/product/blaw/bloombergterminalnews/bloomberg-terminal-news/SAOYIOT0AFB4?criteria_id=9c7124c5fe575996e07e171c65ad1c9d (describing interest to develop "[Japan]'s untapped AI landscape").

59. Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023).

for a justification for the copying will vary with the extent to which the defendant's use does not substitute for the plaintiff's use of its work.[60] If one determines that a use that generates an output that does not substitute for the plaintiff's exploitation of its work is sufficiently "transformative," then it may not matter that the output fails to comment on, criticize, or otherwise be "about" the copied inputs. That said, the question of substitution may require a different assessment in the AI content, because an output might not compete with any particular copied work, but it might compete with the copied author's future prospects of being hired to create works that AI can now generate.[61] Whether courts in fair use cases can or should take account of that kind of substitution is a big question. One might wonder if these outputs advance or undermine the very goal of copyright, which is to advance the progress of science and useful arts, and to do that by encouraging authorship. I think there is room for doubting that generative AI will foster creativity. It is one thing to identify, locate, generate information, but creating new industries, new performances, and new works (if we want to call them that) seems very different to me.

Moving to the European Union, as mentioned, the EU legislature was not thinking about generative AI when drafting Articles 3 and 4 of the 2019 DSM Directive. The definition of TDM in this Directive was clearly aimed at gathering data for analytic uses (to extract information, patterns, and knowledge), but not (at least, not formally so) at "generating" new material.[62] Furthermore, generative AI not only reproduces but transforms the works in the data set, or at least makes derivative works or protected subject matter (i.e., a database), and the AI program communicates the outputs to the public. Whether the CJEU will allow the TDM exception in Article 4 of the DSM Directive to apply to machine learning for use in generative AI is yet to be seen. Additionally, as we mentioned before, Articles 3 and 4 of the DSM Directive do not concern the right of communication to the public, let alone the right of transformation (which is not harmonized in the European Union). Regardless of all these caveats, public debate in Europe seems to take for granted that the TDM provision in Article 4 of the DSM Directive is applicable to permit the training of generative AI systems

60. *Id.* at 533–36.

61. *See* David Cremer et al., *How Generative AI Could Disrupt Creative Work*, HARV. BUS. REV. (Apr. 13, 2023), https://hbr.org/2023/04/how-generative-ai-could-disrupt-creative-work [https://perma.cc/JD3S-BUCL] [https://web.archive.org/web/20250307211631/https://hbr.org/2023/04/how-generative-ai-could-disrupt-creative-work]; Winston Cho, *The Hollywood Jobs Most at Risk from AI*, HOLLYWOOD REP. (Jan. 30, 2024), https://www.hollywoodreporter.com/business/business-news/ai-hollywood-workers-job-cuts-1235811009/ [https://perma.cc/2VXR-5S9S] [https://web.archive.org/web/20250307211715/https://www.hollywoodreporter.com/business/business-news/ai-hollywood-workers-job-cuts-1235811009/]; Alexander Cuntz et al., *Artificial Intelligence and Intellectual Property: An Economic Perspective*, World Intell. Prop. Org. (WIPO), Economic Research Working Paper No. 77 (2024).

62. *See* DSM Directive, *supra* note 20, arts. 3–4. Note that the term used in the DSM Directive is not revised in the AI Regulation adopted in March of 2024. *See* 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 O.J. (L 1689) 1.

unless the copyright owners have "opted out" of it. Whether or not this is the correct interpretation will be determined by courts, but seeing the prevailing reading of this provision among stakeholders, there seems to be a good argument that Article 4 of the DSM Directive, notwithstanding the caveats mentioned above, will end up governing not only TDM for purposes of "functional" AI, but also for "generative" AI.

Turning again to Japanese law, it is unclear whether or not Article 30 would allow the use of copyrighted works in AI outputs because the minute the algorithm issues these outputs, someone is enjoying the work, at least if the work is substantially reproduced in the outputs.

Another open question: Does it make a difference if the AI output consists almost entirely of newly generated content? You can have a creative project that is almost entirely generated by an AI algorithm. AI has the ability to generate deepfakes. Programs can copy the voice of an actor or any other person whose voice it can access a recording of, even the voice of the president, and manipulate the voice to say things that he or she has not said. Certainly, there are positive instances where this could be used as well as negative ones. An actor's performance in a movie can be translated and performed in other languages in his or her own voice, for example. But this would be a case where the use was authorized. Not all such uses are done with permission, and the consequences of unauthorized uses of another's voice can be serious and even dangerous.

But deepfakes are not the only material that may be generated "new" by AI and we have to wonder how current laws would apply to these situations. Will Article 4 of the DSM Directive, which may be implemented and interpreted by each member state differently, have any bearing on AI's generation of "new" works?

Before we wrap up, let me quickly look at the outputs generated by these algorithms. I can think of two issues worth looking into: Whether or not these outputs may qualify as direct infringement, and whether these AI systems may qualify as an indirect, contributory infringement—specially, through the *Grokster* case and the inducement theory. As a rule of thumb, it seems that if these outputs are infringing, then these systems are clearly inducing or contributing to this infringement.

As far as direct infringement, it might be argued that there is no infringement (notwithstanding the copying of copyrighted inputs) if the outputs do not reproduce recognizable portions of the source material. In Europe, the CJEU decided a case involving a two-second extract of a 1977 song, "Metall auf Metall," from an old German group called Kraftwerk, which was continuously sampled (looped) in a hip-hop song.[63] The CJEU said that sampling violates a phonogram producer's exclusive right to reproduce and distribute their phonogram, unless the sample is "in a modified form unrecognizable to the ear."[64] One might extrapolate from that decision a threshold requirement of recognizability, but one should also be cautious because the case

63. Case C-416/17, Pelham and Others v. Hütter and Schneider-Esleben, ECLI:EU:C:2019:624 (July 29, 2019).

64. *Id.* ¶ 39.

involved sound recordings, which under EU law are treated somewhat differently than other kinds of works of authorship covered by traditional copyright law.[65]

In the United States, you have had similar issues also relying on the fair use doctrine. In *Cariou v. Prince*, the Second Circuit held that many of the images generated by Richard Prince on the basis of Cariou's photographs of Rastafarians could constitute fair use, and remanded a few to be reassessed by the lower court.[66] Although the Second Circuit explained (perhaps wrongly) that many of the thirty pictures amounted to a "transformative" fair use of Cariou's works, it is not less true that in this case the recognizability test played an important role.[67]

Then there is the other Prince. The one whose image was re-used by the Andy Warhol Foundation. In *Warhol v. Goldsmith*, the Supreme Court also used the transformative and the fair use doctrine, basically to say this was not a fair use because the series of fifteen works done by Warhol in 1984 was "substantially similar" to the original photograph taken by Goldsmith in 1981, as licensed by *Vanity Fair*.[68] In EU language, we may reach the same result by arguing that Warhol's series could not be exempted as a quotation or a parody, and that no other statutory exception would permit the making of these derivative works.

And one last recent case in the United States, the ruling by the Supreme Court in favor of fair use in *Google v. Oracle*. [69] The Court confirmed the concept of "transformativeness" for the purposes of the fair use defense, in the sense of a new product or service, not so much a derivative work. The fact that Google had copied part of the code used in Oracle's Java's APIs, so that the new code could be written for the Android system, was found to be "transformative" in the sense that Google had developed a "new product" with significant public benefits.[70] In summary, we could conclude that either under the U.S. fair use doctrine or in Europe, under the CJEU's non-recognizable test, similar conclusions could be reached: If the results generated by AI allow for the original work(s) to be identified, to be recognized, they would

65. *See* Irina Eidsvold-Tøien & Are Stenvik, *Copyright for Performers—An Obligation Under International Law*, 18 J. INTELL. PROP. L. & PRAC. 754 (2023).

66. Cariou v. Prince, 714 F.3d 694 (2d Cir. 2013). Cariou had taken pictures of Rastafarians and published them in an art book. Prince transformed these pictures (by adding eyes and hair, placing guitars in their hands, and making collages of several photos) and sold the distorted photographs for a lot of money. In some of them, the original painting could be easily identified. In others, it could not be. The Supreme Court denied Cariou's petition for a writ of certiorari, 571 U.S. 1018 (2013), and the case settled in 2014.

67. *Cariou*, 714 F.3d at 710 (finding it significant that a source photograph was "heavily obscured and altered to the point that Cariou's original is barely recognizable").

68. Andy Warhol Found. for the Visual Arts v. Goldsmith, 598 U.S. 508 (2023). An original photograph taken by Goldsmith in 1981 was licensed in 1984 to Warhol to create a reference for a silkscreen illustration of Prince to be published, by agreement with Goldsmith, only once, with her credited. Warhol created not one, but fifteen new works (the "Prince series"). Goldsmith found out in 2016, when Condé Nast published another work in the Prince series (licensed by the Warhol Foundation) on a magazine cover. The Second Circuit concluded against a fair use because the series retained the essential elements (lighting, shades, contrasts) of the original photograph without adding anything new. A classic case of "derivative work" under the scope of exclusive rights, and a classic case of contractual infringement too. *Id.*

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

70. *Id.* at 30.

constitute an infringement; instead, if the generative AI outputs do not allow for the original work(s) (trained and machine learned and put in the database) to be identified, then these outputs would not be infringing, *per se*. And notice that we are only assessing the infringing character of the AI outputs, not the eventual infringement incurred in making them possible through TDM and machine learning processes.

On the issue of secondary liability, the test may be quite different in the United States and European Union. Indirect liability has not been harmonized in the European Union, and it remains a matter of national laws, usually based on concepts such as negligence, unjust enrichment, as well as contribution to the damage caused.[71] For our purposes, the only exception is the "safe harbours" for Internet Service Providers in the e-Commerce Directive (2001),[72] and more recently confirmed by the Digital Service Act (2022),[73] and specifically, the special rule in Article 17 of the DSM Directive applicable to user-generated-content platforms.[74] A more interesting question might be whether AI systems and platforms could be considered under Article 17 of the DSM Directive. That would very much be counterintuitive because, under Article 17, these platforms require a license to reproduce and communicate works to the public, and, as already explained, the transformation right is not part of the EU *acquis*.

In the United States, generative AI systems might be examined under contributory infringement doctrine (after Sony/Betamax[75]) and the inducement test (after *Grokster*[76]). In the *Betamax* case, the court found the existence of fair use because many copies were being made for time-shifting purposes; subsequently, based on the lack of an underlying infringement the Court concluded that there was no contributory infringement.[77] Taking into account that many generative AI outputs cannot be deemed a fair use, examining generative AI under the Sony contributory infringement doctrine (and its test of "substantial non-infringing uses") may lead to the opposite result and assign liability to generative AI platforms/producers for contributory infringement. Also under the inducement doctrine (*Grokster*), generative AI systems may be seen as inherently designed and trained to precisely produce new outputs that

71. The concept of negligence usually refers to a breach of a standard "duty of care," and lack of action upon knowledge. Contribution involves an active role (a substantial participation in the sense of directly, and knowingly, contributing to the infringement), but also the incitement or fostering of the action. Unjust enrichment relies on the existence of an economic interest and actual knowledge. *See* Raquel Xalabarder, *Intermediarios Tecnológicos y Bases para la Responsabilidad. Intervención Voluntaria u Obligatoria* [*Technological Intermediaries and Bases for Liability: Voluntary or Mandatory Intervention*], *in* DIFUSIÓN Y GESTIÓN DE LAS OBRAS PROTEGIDAS POR EL DERECHO DE AUTOR EN INTERNET [DISSEMINATION AND MANAGEMENT OF COPYRIGHT-PROTECTED WORKS ON THE INTERNET] 405–21 (2016).

72. Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market ("Directive on Electronic Commerce"), arts. 12–15, 2000 O.J. (L 178).

73. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act), arts. 4–8, 2022 O.J. (L 277).

74. *See* DSM Directive, *supra* note 20, art. 17.

75. *See* Sony Corp. of Am. v. Universal City Studios, Inc., 464 U.S. 417 (1984).

76. *See* Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913 (2005).

77. *Sony*, 464 U.S. at 442.

will likely infringe copyright. But these are just speculations, and U.S. courts will most likely examine them in the near future.

Some concluding remarks. We are only starting to identify a few issues that arise related to AI. It is hard to regulate if we do not know yet what we are regulating. We have more questions than solutions. Some say we should wait and see where technology takes us. Copyright will certainly play an important role in the development of AI technology. And we better get it right, because we have all seen the *Terminator* movie, and we know how it ends. We do not wish to send someone from the future into the present to address any mistakes committed. Thank you very much for inviting me to share with you some of the many open questions.